PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

March 2015

FOREWORD

These proceedings are a written report of the eighteenth Sawtooth Software Conference, held in Orlando, Florida, March 25–27, 2015. One-hundred eighty attendees participated.

The focus of the Sawtooth Software Conference continues to be quantitative methods in marketing research. The authors were charged with delivering presentations of value to both the most sophisticated and least sophisticated attendees. Topics included choice/conjoint analysis, surveying on mobile platforms, MaxDiff, TURF, market segmentation, and product portfolio optimization.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each of the papers are photographs of authors and co-authors. We appreciate their cooperation for these photos! It lends a personal touch and makes it easier for readers to recognize these contributors at the next conference.

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

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

We look forward to the next conference!

Sawtooth Software

October, 2015

CONTENTS

SUMMARY OF FINDINGS

The eighteenth Sawtooth Software Conference was held in Orlando, Florida, March 25–27, 2015. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2015 Sawtooth Software Conference Proceedings.

Mobile Choice Modeling: A Paradigm Switch (Dirk Huisman and Jeroen Hardon, SKIM Group): Jeroen and Dirk reviewed the history of market research data collection practice: evolving from paper to CATI, then eventually to web-based and lastly mobile interviewing. Each shift in data collection methodology was met by resistance but each change largely has been validated. The most recent challenge is that the change to mobile interviewing involves dealing with smaller screen sizes and lower attention spans for respondents. Yet, a great benefit of device-based interviewing nowadays is that it allows researchers to test many elements of the modern web-based marketplace. Website commerce can be near-perfectly imitated within market research surveys, allowing researchers to test modifications to websites that can have immediate positive impact on sales. The modifications can be more than just A vs. B, using conjoint experimental designs to simultaneously test a variety of aspects of an offering to determine the best combinations of changes that can improve conversion rates.

MaxDiff on Mobile (Jing Yeh and Louise Hanlon, Millward Brown): At the same time that the use of MaxDiff is increasing as a technique for measuring the importance or preference of items, the prevalence of respondents taking surveys on mobile is also increasing. Jing and Louise studied the impact of asking MaxDiff questions on mobile devices and examined ways to make MaxDiff surveys work well irrespective of the device used to display the survey. The authors compared MaxDiff scores between interviews taken on the PC, tablet, and smartphones. After pulling demographically matching samples, the MaxDiff scores were nearly identical irrespective of which device the survey was completed on. They found that different interviewing devices tended to be used by different types of people. In addition, MaxDiff surveys on smartphone took longer to complete than on PCs and had higher dropout rates. They concluded that devices themselves did not impact substantive results, but the nuances of the demographic groups represented via the devices should not be overlooked.

A Forecaster's Guide to the Future: How to Make Better Predictions (David Bakken, Foreseeable Futures Group): In his presentation, David described why he felt predictions and forecasts often fail. Some of the reasons, he argued, were due to methods that rely entirely on historical associations, due to too many assumptions, lack of a causal model to explain how the particular future comes about, and models that are either too simple or too complex. David described how agent-based models may be used to predict emergent behavior that depends on the interactions between agents (such as consumers and sellers) as well as between agents and their environment. For example, the Bass diffusion model can be realized as an agent-based simulation that explicitly models word-of-mouth networks. The best models, said Bakken, involve disaggregate (typically individual-level) approaches that use the simplest model to capture the important behavior of the system of interest.

Wallet Economics?: Credit Card Choice Based Conjoint—Beyond Preference and Application (Dimitry Estrin, Michelle Walkey, Vision Critical; Vidya Subramani, Client Bank; Carla Wilson, VISA; Jang Tang and Rosanna Mau, Vision Critical): The authors described a conjoint analysis approach to create portfolios of credit card offerings that not only appeal to customers but are profitable to the firm. Credit cards make money mostly via transaction fees, annual fees, and interest charges. The costs to the firm include costs for acquiring customers, the rewards paid out, and redemption costs. Although traditional CBC research can identify the proportion of respondents likely to adopt a credit card, CBC alone does not indicate how customers will use the cards, which directly impacts profitability. So, the authors modified CBC surveys to include additional questions. Respondents were asked how much they would spend per month on the credit card shown on the screen. This information was bridged with information respondents provided regarding what types of expenditures they would make on a new credit card depending on the reward level for different merchant categories. The authors built an integrated simulation that predicted the likelihood of adopting the cards, expenditures, rewards that each respondent would receive, probable attrition, and also rewards not redeemed. This model validated well against actual known figures for monthly profit per card as well as other metrics. In sum, they were pleased that they had built a model that balanced the often conflicting needs of appealing to consumers while maintaining profit margins.

Conjoint for Financial Products: The Example of Annuities (Suzanne B. Shu, Robert Zeithammer, UCLA and John Payne, Duke University): Annuities seem to offer for many people in the US an opportunity to reduce their risk and insure themselves against outliving their savings. However, few people actually purchase annuities and when they do they often make poor decisions regarding different annuity offerings. The authors employed conjoint analysis to study and make recommendations regarding how insurance companies should market annuities to consumers and how regulators can help consumers make better decisions. They found that if insurance companies were required to give concrete information ("do the math") regarding the expected payoff value of different annuity packages, consumers would be able to make better decisions that benefited them. At the same time, if insurance companies would "do the math" for consumers and show the payout rates for different annuities, they could gain an edge in terms of increasing the likelihood that consumers would purchase them.

Comparing Message Bundle Optimization Methods: Should Interactions Be Addressed Directly? (Dimitri Liakhovitski, GfK; Faina Shmulyian, MetrixLab and Tatiana Koudinova, GfK): Dimitri and his co-authors examined multiple methods for finding near-optimal bundles of messages for promoting a product or service. The approaches involved MaxDiff, traditional ratings scales, and two variations of choice-based conjoint (CBC). They analyzed the MaxDiff results two ways—by simply summing the MaxDiff preference scores and by defining reach and applying TURF. They analyzed rating scale results using TURF. They analyzed CBCs with and without interaction terms. They found that the TURF-based procedures performed least well of the approaches investigated, most likely because TURF does not directly address semantic synergies among messages. Simply summing the MaxDiff preference scores worked better than the TURF-based approaches. The best method among those they tested was CBC with interaction terms. This does not mean that TURF is not an appropriate method for other optimization problems (e.g., line optimization), the authors concluded. It is just that TURF is not best suited for message bundling applications.

Using TURF Analysis to Optimize Reward Portfolios (Paul Johnson and Kyle Griffin, Survey Sampling International): Paul and Kyle described how Survey Sampling International (SSI), like other panel providers, faces the challenge of keeping their panelists happy and involved in the panel. The rewards SSI offers its panelists is the key way to improve retention and activity rates. Rewards are most likely given in the form of gift certificates that may be redeemed at a variety of retail partners. However, it costs SSI to manage such a program. Those costs include attracting new panelists, the price of buying gift certificates, the costs of managing inventory of gift certificates (some may expire if not used in time), and any volume discounts some retailers may provide to SSI. The authors used questionnaires (employing both MaxDiff+TURF and a self-explicated TURF approach) to ask respondents their preferences for gift certificates from different retailers. Their analysis pointed to streamlined portfolios of retailers that could satisfy panelists as well as reduce the cost of managing the rewards program to SSI. They were able to compare the survey results (stated preference for retailers) actual choice behavior (picking gift cards) for these same respondents and found excellent validation for the survey results. The findings will allow SSI to provide better rewards for their panelists while potentially lowering the costs for managing the panel.

Bandit Adaptive MaxDiff Designs for Huge Number of Items (Kenneth Fairchild, Bryan Orme, Sawtooth Software, Inc. and Eric Schwartz, University of Michigan): Sometimes researchers use MaxDiff to find the best few items among large lists of items. In such situations, traditional level-balanced MaxDiff designs are inefficient, spending a lot of respondent effort evaluating least desirable items. A statistical approach called Thompson Sampling has been used to solve “bandit” problems (so called because of the academic example involving maximizing the payout when playing multiple slot gambling machines, also known as “one-armed bandits”). It turns out that the same theory may be applied to MaxDiff problems involving huge numbers of items. After a few respondents have been interviewed using MaxDiff, aggregate logit may be used to estimate the means and standard errors for the items in the list. Applying Thompson Sampling based on the aggregate logit parameters, most-preferred items are then oversampled for subsequent respondents. The logit results are updated in a continuous, real time way, after new respondents have been interviewed. Using robotic respondents answering with realistic preference functions and error, the authors demonstrated that the bandit MaxDiff approach can be as much as 4x more efficient at identifying the top few items than the traditional level-balanced approach.

What Is the Right Size for My MaxDiff Study? (Stan Lipovetsky, Dimitri Liakhovitski and Mike Conklin, GfK North America): MaxDiff studies are commonly employed nowadays and the authors develop a theory for sample size planning. They conducted various simulations to validate their proposed formula for estimating needed sample size. The results give a tool for practitioners to use when planning sample sizes for MaxDiff studies.

“Performance, Motivation and Ability”—Testing a Pay-for-Performance Incentive Mechanism for Conjoint Analysis (Philip Sipos and Markus Voeth, University of Hohenheim): For a number of years now, researchers have proposed ways to try to motivate respondents to provide more truthful and higher-quality conjoint analysis data. These efforts are grouped under the term *incentive alignment*. The central idea is to give respondents rewards or other motivations so that they realize there is a consequence for their choices in a conjoint questionnaire and act in their self interest, which in turn provides better data to the researcher. Previous researchers have rewarded respondents with products that either exactly matched profiles they picked in conjoint questionnaires or were near fits to choices they made. But, Philip and Markus pointed out that such rewards are not always feasible in market research, particularly when the cost of the product or service involved is prohibitive. They propose giving respondents a higher payout (incentive) based on their performance in the conjoint interview. Performance can be measured in terms of internal consistency or hit rate. College students served as respondents to a conjoint analysis survey, where some received additional payment based on performance. The authors found a statistically significant improvement in holdout predictive validity among respondents who were given additional incentive based on performance. Moreover, Philip and Markus demonstrated that—though motivation constitutes an important factor to enhance performance—high performance is also about respondents' ability to make the cognitive effort required during a conjoint task.

Perceptual Choice Experiments: Enhancing CBC to Get from Which to Why (Bryan Orme, Sawtooth Software, Inc.): Traditional CBC simulators tell us which products are preferred, but provide no insights into *why* they are preferred. Bryan introduced perceptual choice experiments as a way to enhance traditional CBC simulators to give greater insights into the perceptions, motivations, and attitudes of respondents toward the product concepts defined in the market simulation scenarios. The approach involved adding perceptual pick-any agreement questions beneath the standard CBC questions. For each product concept, respondents click whether they agree that it is associated with given perceptual items. Bryan used aggregate logit to build models that predict the likelihood that respondents would agree that any product concept (defined using the attributes and levels of the CBC experiment) would be associated with each of many perceptual items such as fun, creates memories, educates, etc. The agreement scores may be shown as interactive heat-maps within Excel-based market simulators. The main drawbacks are that it takes about double the respondent effort to complete CBC surveys that have the included perceptual agreement questions and the sample sizes needed to stabilize the perceptual models can add to data collection costs. But, some good news may counteract the bad news: Bryan's empirical test suggested that respondents may give better CBC data when they additionally are asked the perceptual choice agreement questions.

Profile CBC: Using Conjoint Analysis for Consumer Profiles (Chris Chapman, Kate Krontiris and John S. Webb, Google): Product design teams in technology often use qualitative research to develop consumer descriptions (often known as "personas") for design inspiration and targeting. For example, a persona may read as, "Kathleen is 33 years old and is a stay-at-home mom with two children . . ." While a persona may be a good way for managers to attach a memorable description to a market segment made up of many people, few consumers will fit the *exact* description for every attribute. This makes it difficult to size the market for a target persona. Google Social Impact was interested in quantifying what weighted percent of respondents at least approximately fit into different personas they had already developed in qualitative field research regarding engagement with elections and civic life. The authors developed a conjoint analysis study with attributes derived from the qualitative personas, such as: I'm not working or in school right now; I spend as much time with my family as I can; I try to do as much civic engagement as I can, etc. Respondents saw partial-profile CBC tasks and picked within each task the concept that best represented them. The authors used latent class analysis to identify six key civic profiles, which gave composite class descriptions and market sizing. They concluded by discussing CBC design principles they suggest for such research, including response format, number of levels shown, and number of concepts.

RUM and RRM—Improving the Predictive Validity of Conjoint Results? (Jeroen Hardon and Kees van der Wagt, SKIM Group): Kees and Jeroen provided a useful overview of the differences between Random Utility Modeling (RUM—the additive compensatory model) versus RRM (Random Regret Modeling—a relatively new non-IIA bound method of modeling

CBC experiments which may be done with most any commercial logit-based utility estimation routine, including CBC/HB). RRM posits that respondents pick the concept within a task that minimizes their regret for not getting aspects that were better in the competing concepts. Kees and Jeroen compared the predictive validity of the different models with a few CBC studies, finding mixed results. They also investigated a hybrid model which incorporated both RUM and RRM characteristics. The hybrid model also produced mixed results, with some challenges of multicolinearity to overcome (which the authors handled via additional utility constraints in CBC/HB). Although RRM seems promising for certain kinds of product categories and applications, it doesn't always lead to better models than the standard RUM model specification. The hybrid approach would seem to offer some of the benefits of RRM, but could be a safer approach due to its leverage of the robust RUM model. One challenge for RRM modeling is that only ordered attributes (like speed and price) may be RRM coded.

Capturing Individual Level Behavior in DCM (Peter Kurz, TNS Infratest and Stefan Binner, bms market research + strategy): Peter and Stefan illustrated how sometimes in larger DCM study designs, respondents can be observed following certain decision rules in their choice questionnaires and yet the part-worth utilities estimated by HB can suggest otherwise. The authors pointed out that sparse designs (many attribute levels to estimate relative to few choices made per respondent) result in quite a bit of Bayesian smoothing of respondents toward the population (or covariate) means. They showed simulations that varied the number of tasks per respondent. With fewer tasks, the respondents' utilities tend to be shrunk more toward the population means. Peter and Stefan pointed out that when the number of tasks is few and the respondent's personal preferences differ from the vast majority of respondents, HB utilities for that respondent may seem to conflict with that individual's preference and more revert to the population preferences. They concluded by recommending that researchers be on the lookout for issues due to Bayesian smoothing and recognize that DCM models have great predictive accuracy in terms of total market, but can have problems for small segments or niches. They suggested that if you expect sparse data for specific and important sub-segments, then you should apply covariates and also oversample such sub-segments.

Occasion Based Conjoint—Augmenting CBC Data to Improve Model Quality (Björn Höfer and Susanne Müller, IPSOS): Björn and Susanne described how to enhance CBC questionnaires with additional questions regarding product use occasions to deliver better insights. In addition to the standard CBC questions respondents indicate which SKUs they use under different occasions (pick-any data) and how relevant each occasion is. Since the CBC data collection does not differ from standard CBC the integration of occasions is shifted to the utility estimation and/or preference share calculation. In their methodological comparison they found that the Occasion-Based Conjoint (OBC)—although it improves the estimation of substitution effects (face validity)—does not perform better than a standard volumetric CBC model in terms of internal and external validity criteria. Nevertheless the integration of occasions can be recommended to benefit from more realistic estimates of new product sales potential and substitution effects as well as from the additional insights on the motivations behind product choice that can support marketing strategy decisions.

Precise FMCG Market Modeling Using Advanced CBC (Dmitry Belyakov, Synovate Comcon): CBC studies for Fast Moving Consumer Goods (FMCG) categories can become quite complex. Often there are a dozen or more offerings (SKUs) that differ in terms of brands, package sizes, product forms, and prices. Dmitry described different strategies for designing

CBC questionnaire for complex FMCG studies and for modeling the data. Dmitry reported results for a simulation study involving consideration sets. With consideration sets designs, respondents see only the SKUs within the CBC tasks that they screen in (would consider). The coding method that compares both accepted and dropped SKUs (in a series of binary paired comparisons) to a threshold SKU parameter performed best among those he tested. The second challenge Dmitry described involved modeling price sensitivity for dozens of SKUs. If the data were sufficient, ideally the researcher would estimate a separate price slope for each SKU. But, the data are typically too sparse to do a good job with this approach. Dmitry suggested a method of grouping SKUs into just a few segments based on the slope of their aggregate logit alternative-specific price coefficients. The SKUs within the same segments can be coded with a shared price slope to economize on the total number of parameters for HB estimation while still capturing good SKU-based price information.

Defining the Employee Value Proposition (Tim Glowa, Garry Spinks, and Allyson Kuper, Bug Insights): Tim and his co-authors described how conjoint analysis may be used to help retain employees by improving rewards packages. Designing rewards packages involves not only measuring what employees think is valuable but balancing those desires against the costs of different programs. Tim suggested using best-worst conjoint, which is a conjoint-style variation on the traditional MaxDiff survey. With best-worst conjoint, respondents are shown an employment package (just as a conjoint profile, composed using one level from each of many attributes) and indicate which one level from that profile has the most positive impact on them and which one level from that profile has the least positive impact. Using logit-based analysis (e.g., logit, latent class MNL, HB), scores are estimated for each attribute level, similar to conjoint analysis. Some of the challenges of doing best-worst conjoint among employees are: 1) studies often are global, spanning multiple countries and languages, 2) large sample sizes, sometimes 20,000+, 3) the emotional sensitivity of the subject matter to the respondents, and 4) managing anxiety and expectations of the employees.

Menu-Based Choice: Probit as an Alternative to Logit? (Christian Neuerburg, GfK Marketing & Data Sciences): The commonly-used HB-logit is often the tool researchers use for modeling menu-based choice (MBC) data. However, some literature recommends multivariate probit, which is a theoretically more appealing model for modeling an array of dependent variables from a menu. Christian conducted a very extensive simulation study to examine the pros and cons of HB-logit vs. multivariate probit. He created 288 synthetic datasets (with known "true" preferences) that varied the degree of respondent heterogeneity, the menu complexity, sample size, number of tasks, and the assumptions of the behavioral choice models. He found that HB-logit models consistently outperformed multivariate probit models for nearly all of the conditions. Furthermore, HB-logit is much quicker to run and both open source and commercial software is available for HB-logit. Some of his more detailed findings were 1) relatively few tasks are needed for good individual-level models under HB, 2) individual hit rates are relatively unaffected by sample size, and 3) larger sample sizes are quite useful for reducing the error in predictions for aggregate shares of choice. The multivariate probit models are less parsimonious, are more complex to estimate, and are less scalable to larger commercial MBC studies than the HB-logit approach.

Combining Latent-Class Choice, CART, and CBC/HB to Identify Significant Covariates in Model Estimation (George Boomer, StatWizards LLC and Kiley Austin-Young, Comcast Corp.): George and his co-author Kiley explained that covariates are often important, for example, gender in the handbag market, income in the exotic car market, age in the market for geriatric medicine. They proposed an approach for identifying key covariates and incorporating them into a CBC simulation within a time frame that comports with practitioners' schedules.

Their approach makes use of three techniques applied to a common data set. First, CBC/HB is employed to produce a set of individual-level utilities. Second, a latent-class choice (LGC) estimation identifies groups of respondents who share a common set of utilities. Third, CART is used to improve upon LGC's covariate classification. Finally, the latent classes and significant covariates from modern data mining techniques are brought together in a common market simulator. The authors used both a simulated data set and a disguised, real-world example from the telecommunications industry to illustrate this approach.

Uncovering Customer Segments Based on What Matters Most to Each (Ewa Nowakowska, GfK Custom Research North America and Joseph Retzer, Market Probe): Ewa and Joseph discussed an approach to clustering data called co-clustering. Co-clustering is an emerging method that connects two data entities, e.g., rows and columns of the data. Typically factor analysis is used to find groupings of variables and cluster analysis finds groupings of cases. Co-clustering simultaneously finds groupings of variables and cases by taking into account the pairwise (dyadic) relationship between the two. Among other aspects of co-clustering, the authors demonstrated how the same respondent can simultaneously belong to multiple co-clusters as well as how a particular variable may be used to define more than one co-cluster. An illustration employing airline traveler data was reviewed. The variables included attitudinal and behavioral information about the respondent as well as customer satisfaction data regarding specific airlines. Co-clustering may be done within R using the "blockcluster" package.

Climbing the Content Ladder: How Product Platforms and Commonality Metrics Lead to Intuitive Product Strategies (Scott Ferguson, North Carolina State University): One of the challenges of using conjoint analysis in product line optimization problems is that many of the solutions may not make sense from a business standpoint. Scott began by reviewing his previous effort at the Sawtooth Software conference regarding multi-objective search, which finds solutions that concurrently satisfy multiple goals such as profit and market share. Beyond satisfying multiple objectives, Scott's new work dealt with the problem of creating product portfolios that make sense in terms of their structure. It is less expensive to provide multiple products that share a lot of characteristics, so a product portfolio that has a lot of commonality and yet reaches a variety of people would be desirable. Even though imposing increased commonality upon the solution space usually comes at the expense of other goals such as market share or profit, Scott reported that portfolios that emphasize commonality will also tend to avoid less extreme products.

A Machine-Learning Approach to Conjoint Analysis: Boosting and Blending Ensembles (Kevin Lattery, SKIM Group): Kevin's presentation explored what might happen if machine learning enthusiasts analyzed conjoint analysis results. First, he pointed out the recent successes that machine learning has had for prediction problems, notably the $1 Million Netflix prize. The winner of the grand prize, as well as all the leading methods were Ensemble approaches. Ensemble analyses blend different, diverse predictive models to improve overall predictions. Critical to their success is having a large number of quality, yet different solutions to a prediction problem. Kevin demonstrated how ensembles of latent class solutions can improve prediction for conjoint analysis problems, even surpassing the predictive rates of HB (for RLH across 3 holdout tasks in two different studies). He first generated diverse models by different random seeds, followed by pruning those models with higher correlations among their predictions. Kevin then attempted to improve upon the randomly generated ensembles by using a boosting approach. He tried several different modifications of AdaBoost. His best boosted approach used the Q-function based on standardizing the likelihood across specific tasks. However, even this method did not improve over the ensembles generated from different random seeds.

The Unreliability of Stated Preferences When Needs and Wants Don't Match (Marc R. Dotson and Greg M. Allenby, Fisher College of Business, The Ohio State University): Inviting respondents to take your conjoint analysis survey that actually have need states that lead to higher engagement in the interview yields significantly more reliable data. That's the conclusion that Marc and Greg drew after applying a statistical model that explored the mechanism through which relevance (i.e., when needs and wants match) impacts consumer choice. They reported results for an empirical study with 567 respondents and concluded that not correcting for unreliable respondents has the potential to introduce parameter bias. By screening out respondents who didn't have any of the needs met by the product category, out-of-sample hit probability improved. The authors suggested that practitioners use stricter screening criteria to ensure the choice surveys are relevant to respondents and that they really have need states that put them in the market to consider and purchase the products in question.

MOBILE CHOICE MODELING: A PARADIGM SWITCH

DIRK HUISMAN
JEROEN HARDON
SKIM GROUP

INTRODUCTION

Mobile research is rapidly growing and has been hot for a couple of years. In marketing research, every 5–10 years a new mode of data collection expands our opportunities to connect with consumers and collect data. Each new platform creates buzz and arousal in the research industry. The switch to mobile devices is similar. However, the limitations of mobile devices lead to challenges, such as the difficulty of creating relevant choice tasks on a mobile device. For us, the real excitement is not in taking such a challenge, but in the paradigm switch.

The use of mobile devices is increasingly typical in consumer behavior, and the opportunities and limitations of the mobile devices offer new marketing models and research solutions. Mobile devices play a dominant role in marketing and in the uptake of e-commerce: product reviews are available in a second; while shopping in stores, consumers compare prices and promotions with online information; the share of online and mobile shopping is increasing rapidly; and in the future, the share of online shopping may surpass 50% of many households' consumption. Because a mobile device is one of the shopping environments of e-commerce, we can mimic the online buying reality in choice modeling studies. This capability enables us to integrate research with the primary business process (real transactions).

PLATFORM SWITCHES OVER TIME

Market research is an industry and like any other industry the development can be analyzed from an industrial economic perspective. Innovation is one of the drivers of fundamental changes in industries, and only innovations which deliver real value through leaps in effectiveness and/or leaps in efficiency stick. The leaps of effectiveness often are value related. Extra value is offered because insights are delivered that could not be delivered before. The examples of value based progress are more selective than the efficiency based progress, because value is need and context specific (i.e., client industry). These examples of progress are often related to changes in marketing practice, hence changes in needs. Leaps in efficiency are often technology and operations driven and normally operations are gradually switched to new platforms.

From its early days choice modelling has witnessed four platform switches.

In the sixties of the previous century marketing became a science and many models and theories regarding the role and impact of the marketing instruments were developed combining econometrics, psychometrics, psychology, sociology etc. These were also the days that the first conjoint analysis models were developed. To collect choice or preference data, respondents had to either rank cards or scenarios (showing different specifications of the same features), or they had to fill in matrices of feature combinations, all on paper and with pencil. These were also the days that Computer Aided Telephone Interviewing was introduced and the days of big mainframe computers; IT departments managed like a kingdom and huge halls with computer interviewers, coders and data entry personnel. The switch to computer aided telephone interviewing added process control in the data collection process, more efficiency and more speed. However choice modelling was too new and complex to be impacted by the first platform switch.

Figure 1. The Four Platform Switches in Market Research "Witnessed" by Choice Modeling

From paper and pencils to CATI

From PP/CATI to CAPI

From PP/CATI/CAPI to web based

From PP/CATI/CAPI/web to Mobile

The second platform switch came in the seventies and early eighties with the introduction of the personal computer. Mainframes were replaced by PCs, data were directly entered during the interview, and the control was an integrated part of the interview software. Again the platform switch brought efficiency, speed and control. This is the period when Richard Johnson, founder of Sawtooth Software, integrated marketing science in the interview. Adaptive Conjoint Analysis was pure magic, artificial intelligence: respondents saw with their own eyes and experienced that the PC directly implemented their previous answers and created choices which were hard to make, the computer knew what they wanted. But more importantly marketing science became available to more organizations, it was socializing of marketing science, because it became feasible to many companies to apply marketing science. These early days advanced PC based systems were a breakthrough, because breakthrough methodology was enabled by simple programming. All the codes needed to create a questionnaire were summarized on one small sheet of paper. But one has to realize that the questionnaires themselves were only text based.

These early days PC-based systems as well as the researchers and marketers who used the systems were frontrunners-explorers who really did change marketing. So it was not only a platform switch but also a paradigm switch. You can imagine the vibe of these days, doing things we could not do before.

During the following decade in the nineties you see primarily process improvements, method improvements, and the exponential growth of processor capacity as well as the start of the World Wide Web. With the internet we have the next platform switch, particularly in the first 10 years of the web. In market research it was another platform to do what we did before, but now faster, more efficient and with a greater reach. And of course with the increased processor capacity we started to use multimedia, visualize, and make the interview more real and intuitive. But bottom line we did what we did before faster and more efficiently.

It is 5 to 10 years after the introduction of the World Wide Web, when it became a commonly used medium, that we see the next paradigm switch in market research. Instead of asking, we observe and analyze what people say and post. Although extremely relevant I leave it aside because for choice modelling we have not made the connection with social media yet. In combination with the internet and social media we see the rise of mobile devices, tablets as well as smartphones. These mobile devices define the latest platform switch and a paradigm switch as well.

Mobile was in the first place an extension to the PC-based and web-based platforms. For data collection the PC is partly substituted, depending on the country and the category ranging from 20% to in the near future 50%, by a tablet or a smartphone. So deliberate or coincidentally mobile devices are part of the sample. That sounds simple, but the screen and context are completely different and upfront often you do not know which device will be used to answer the questions. Can we use the same questionnaire? How to adapt to reality?

It is a matter of programming, but based on the detection of the device, during interviews it will be defined how the stimuli and the questionnaire will be shown. The design of the choice tasks and of the questionnaire will be device specific.

Comparing Mobile Study Results with PC Based Study Results

As a community of researchers, at every platform switch we are triggered to show that the results of studies on the new platform are as good as or better than the standard at that time. A walk through the Sawtooth Software conference proceedings of the past 25 years gives you quite a few nice examples of our willingness to show that new platforms are better and provide competitive advantages.

The platform switch to mobile is no exception as was shown at the 2013 Sawtooth Software conference[1]. We also ran several "validation studies" comparing results from desktops and mobile phones. This taught us a lot about the necessary conditions for running choice-based conjoint studies on mobile devices:

> *Sample:* A larger sample allows for asking fewer questions while still arriving at robust model estimates. You can use the rule of thumb of having twice the sample you would use in a desktop study when running your survey on mobile.

[1] Chris Diener et al. Making Conjoint Mobile: Adapting Conjoint to the Mobile Phenomenon. Sawtooth Software Conference Proceedings 2014
Joseph White. Choice Experiments in Mobile Web Environments

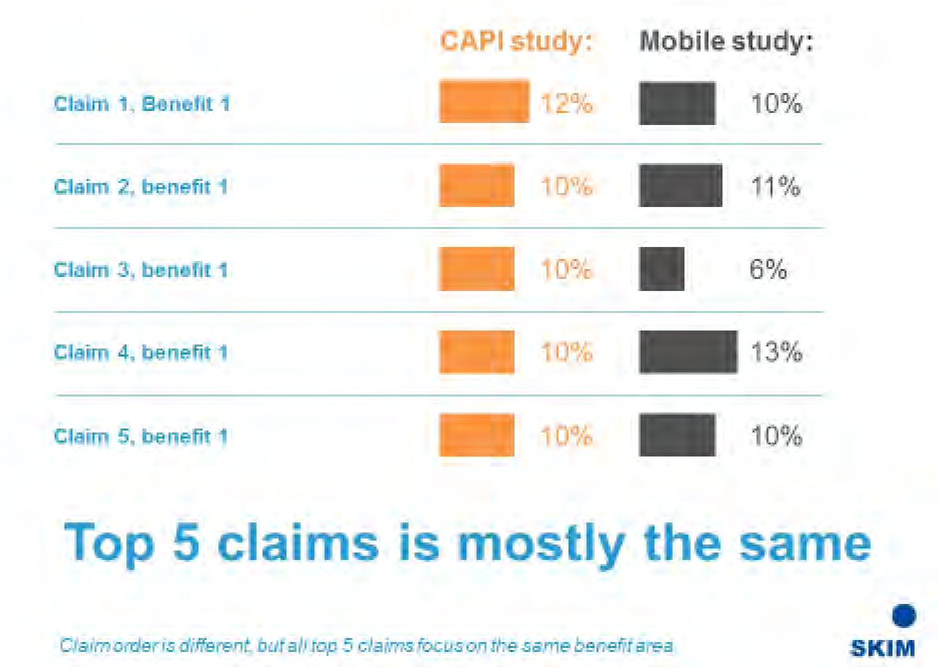
Statistical design: Reduce the complexity of your model. 3*3 mobile CBC works best with perfectly balanced designs—don't waste conjoint tasks on exclusions, prohibitions, or alternative specific attributes.

Layout: Make use of the space you have—max. 5 attributes, concise level descriptions, work with icons and brand logos where possible.

Still the key question is whether the outcomes are different and which one reflects reality best? We conducted a study in Vietnam regarding laundry detergents for hand washing to test commercial product claims aimed to drive sales. The test study was based on MaxDiff and it made sense to conduct the test in an Asian country, because mobile research in Asia is more common than in the Western world. In the PC-based questionnaire the test was based on 12 choice or best-worst tasks and a 20 minute interview. The mobile interview lasted 5 minutes and we included only 3 tasks, but we had a larger sample. So the total number of tasks was the same. The turnaround time of the CAPI interview was 3 weeks and of the mobile study 1 week and the field costs of the PC based study were twice the costs of the mobile study.

Comparing the results we learned that the top 5 claims were the same and focused on the same benefit area, but the order of the claims differed and in the mobile study the differences were slightly more outspoken.

Figure 2 Comparing the Outcome from a PC Based Sample and the Mobile Sample



Top 5 claims is mostly the same

Claim order is different, but all top 5 claims focus on the same benefit area

A potential reason for these differences may be the incomparability of the samples (sample differences). In earlier studies in Europe we only found minor sample differences, but in hindsight that might be a panel provider's initiative to create comparability between the samples. In this comparative study in Vietnam the differences were much larger and significant, particularly regarding gender, education and age.

To eliminate the sample effect we matched a subsample from the mobile sample with the CAPI-based sample. Based on the matching samples the order differences disappeared and there were no significant differences in outcomes between the two studies anymore. So it was not the platform that explains the differences but the sample.

We have to realize that samples of mobile owners are representative of the population of mobile owners, just like the sample of PC owners are representative for the population of PC owners. In rare cases this sample is completed with a subsample of non-PC owners who participate in interviews in a central location. Over time the population of mobile owners will represent the total population better than the population of PC owners, but in the intervening period it is advisable to include both subsamples (PC users and Mobile users). In case the objective of the survey requires that lower social classes are well represented it is advisable to draw a sample of feature phone users and adapt the questionnaire for that subsample[2]. It is important to be critical on the sample population, even when you stratify them based on demographics to mimic total population.

The New Mobile Reality

Extrapolating the trends it is more than likely in the future that tablets and smartphones will have replaced PCs as the primary information processing device of the consumer. For market research this means we have to adapt to this new reality, which is far more diverse than what we call "the new 5x5 reality." The 5x5 reality reflects the trends that respondents are only willing to respond seriously and spontaneously during a short period of time, say 5 minutes is the max. And the smallest screen size to read the question or observe the stimuli will be 5 inches wide. Of course there will be larger smartphone screens and people may use smart TVs or tablets etc., but when we want to reach everybody any time the small smartphones should be the standard from which we can expand the questionnaire to better-looking, more complicated alternatives.

The perk of mobiles as a platform to interact with the respondent is that you can get real time, in the moment and context specific information, but this bonus does not compensate for the reduction of the length of the interview or for the limitation of the screen size. So for complex studies we will have either the option to use other devices (tablets, smart TVs) or to decompose the questionnaire into a string of mini questionnaires, which will be integrated at the end. The advantage of the string is that we can apply in the consecutive interviews what we have learned in the early mini interviews.

How will future choice models look like on mobile devices? To answer this question we have to bear in mind that consumer behavior is changing rapidly. Depending on the country the share of online shopping may ultimately grow to over 50% of total household consumption and online means increasingly it will be on mobile devices. This means that the device used to buy online will be the same as the device to use for the choice experiment and even for the offline buying; we cannot deny the dominant role of the mobile devices. When defining the development tracks for future choice experiments we distinguish 3 tracks: A. adapting choice task regarding traditional choice experiments for the mobile device, B developing choice models that mimic online buying, C. integrating the choice model in the real buying process by evolving the current A/B testing into A/Z testing.

[2] Robin de Rooy, Mark Shoubridge: Bringing High Tech Research to Low Tech Devices, MaxDiff on feature phones and low end Androids. MRMW Asia Conference, 2015

Figure 3. Three Development Tracks for Mobile Choice Modeling

3 tracks

Adapt traditional models

Mimic online buying

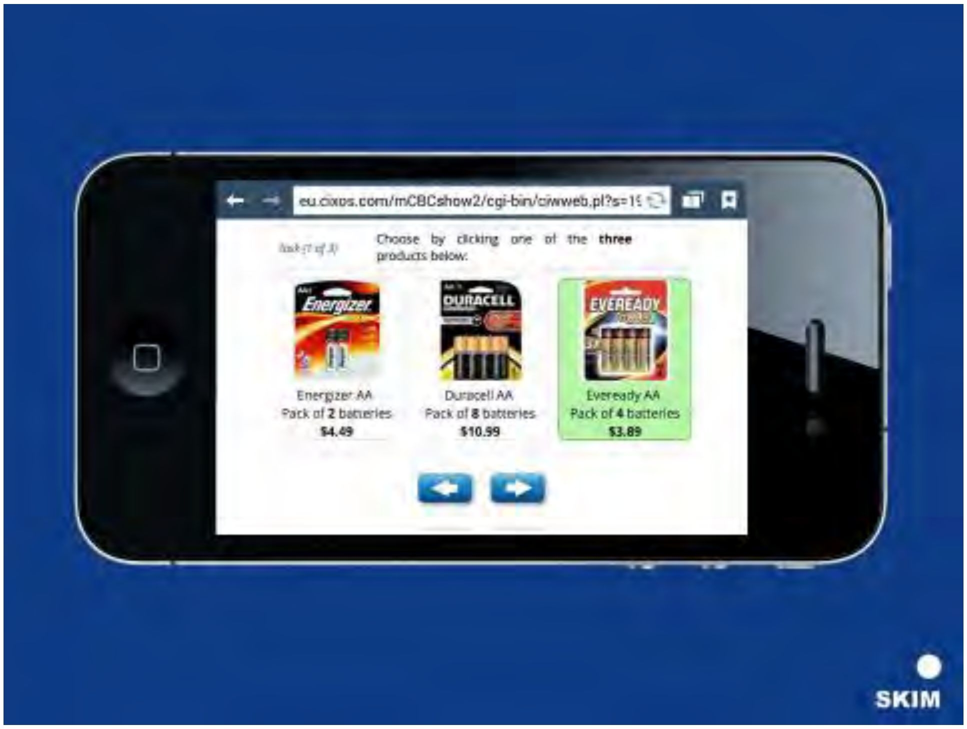
From A/B to A/Z testing

Track 1. Adapt Traditional Models

Traditional choice tasks including virtual shelf studies require wide screens to mimic reality and to mimic the complexity best. These choice experiments might move to tablets and smart TV's, but the screen size of smartphones can't be used for these experiments. The motto for mobile research clearly is: "Doing more with less." That is why we need to develop solutions with fewer conjoint tasks, fewer concepts on the screen and shorter surveys overall to get even better results, hence increasing the sample size.

This made us start with the end in mind. We forced ourselves to think through what mobile research is really all about and how we imagine the ideal mobile conjoint exercise: no more than 3 taps on a smartphone. In Figure 4 a typical mobile choice task is shown. So we can't mimic the complexity of the offline buying situation, but for many objectives it functions perfectly well. A lesson learned during the conjoint studies in this track is that it is better to use visuals designed to be seen on a small screen. Originally we used pictures of products you see on the shelves. However the small text on the package is not readable, so it is better to simplify and only show the essential elements of the product or package. Complex designs could be cut into small studies that can be linked to each other, but this needs further validation and in the meantime for these studies one should focus on devices with the required screen size and screen qualities.

Figure 4. A Typical Choice Task on a Mobile Device.



eu.cixos.com/mCBCshow2/cgi-bin/ciwweb.pl?s=1

task (1 of 3)

Choose by clicking one of the **three** products below:

Energizer AA
Pack of **2** batteries
$4.49

Duracell AA
Pack of **8** batteries
$10.99

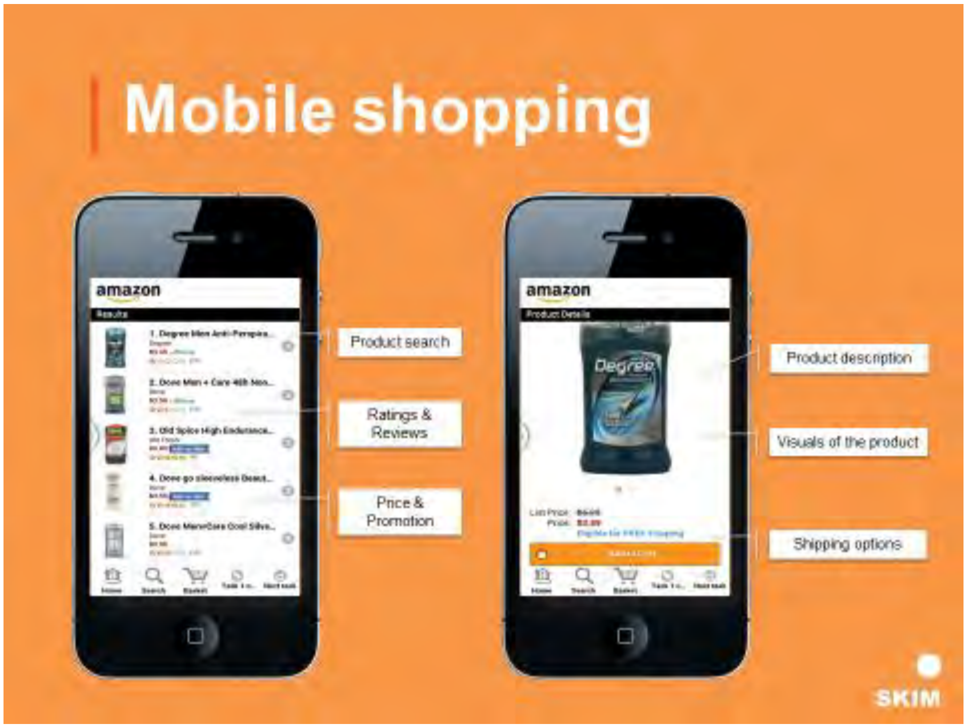
Eveready AA
Pack of **4** batteries
$3.89

Track 2. Mimic Online Buying

Online buying and e-commerce is different from offline buying because other stimuli influencing the decision play a role. These extra attributes are an integrated part of the offer and the buying decision. To mimic the buying situation these features should be included in the model. Examples are product search, rating and reviews, product descriptions, the visualization of the product and shipping options. Like with virtual shelf research, where consumers select products from the shelves, mimicking the specific outlet where they normally buy, online we should mimic the website or the online shop. An example is a choice task replicating a buying situation at Amazon.

It comes down to decomposing the online buying situation and buying processes and defining a design which brings the respondent in 2 clicks into the buying situation which is of relevance to him or her, and in this situation he or she should see only relevant choice options. However to mimic online buying the search options and providing extra information options should be integrated in the model as well.

Figure 5. Decomposing and Mimicking the Online Shopping Behavior



Mobile shopping

Product search

Ratings & Reviews

Price & Promotion

Product description

Visuals of the product

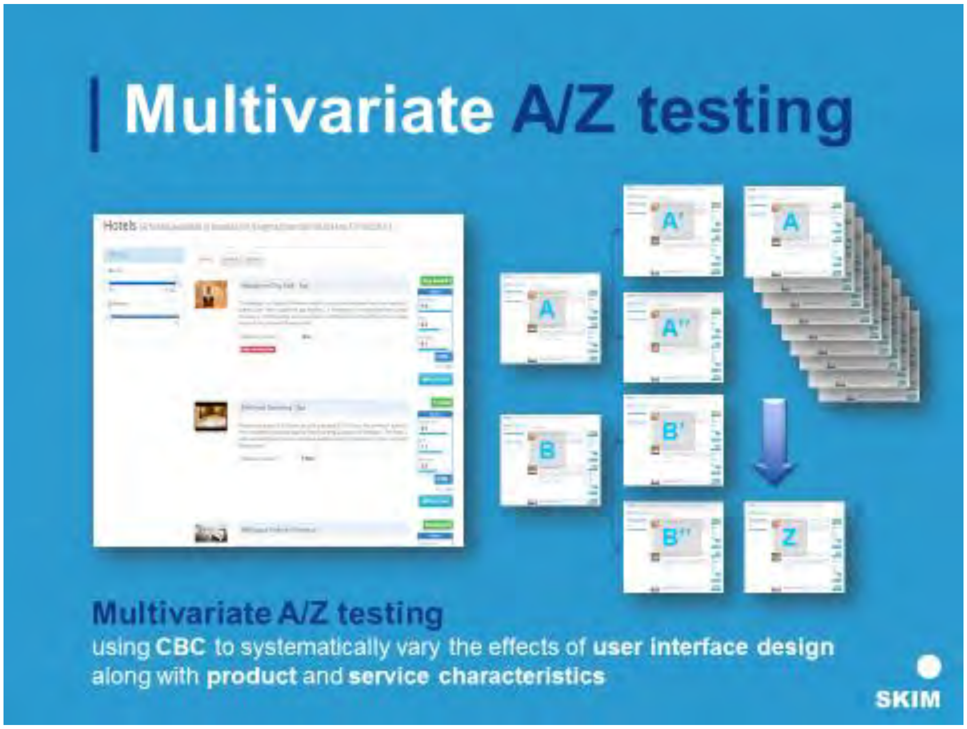
Shipping options

SKIM

Track 3. Integrating the Choice Experiment into the Real Buying Process from A/B to A/Z

When we are able to mimic the online buying situation it is only a small step to test in the real buying environment leading to real transactions. This requires the commitment and collaboration of the web shop. Testing websites or web shops is not uncommon. Normally web shops conduct A/B tests, comparing two version of a website and analyzing which one scores best. Websites are complex and include many interaction design elements, which can be varied in many ways in order to optimize the website or maximize revenue. A/B testing is too simplistic to optimize. In order to optimize and to test the impact of the design elements and the product elements (the attributes) for choice modelling you normally create a balanced design in which the various attribute combinations are shown in Z combinations. Instead of showing only two combinations (A/B testing) per respondent we show a limited selection of the A to Z combinations and across all the visitors and buyers ultimately we are able to identify the impact of each attribute on choice. This approach was successfully tested for booking hotel rooms, but unfortunately the test was not tested on the real booking infrastructure. In real life the adaptive choice models should be applied, because consumers want to select and book a.s.a.p the room they like best, so they want the website to think along and steer him/her to that offer.

Figure 6. Testing During the Real Buying Process Based on a Balanced Design of Website Elements



| Multivariate A/Z testing

Multivariate A/Z testing
using **CBC** to systematically vary the effects of **user interface design** along with **product** and **service characteristics**

SKIM

Now What?

Offline Behavior

To advise on how to best anticipate and influence offline behavior we still foresee choice modelling based on choice settings that mimic offline reality. It will be based on the same systems used to create online choice tasks, but the experiments will be conducted on tablets and smart TVs. Creating shelves, creating a balanced design, fielding and analyzing are all tasks that need to be fine-tuned to the specific situation and questions to be solved. It is always tailor made, but even then all these tasks can be and will be automated, which is essential because in the near future these studies will have to be completed in a few days max.

Mimicking offline behavior we will also measure in context and real time in front of the shelves, using the mobile device to test additional information to influence the decision. This will be a new element to add to the existing models.

For the sake of argument and because the models are different we separated the offline choice modelling from the online choice modeling. In reality the consumers will alternate, experiences and drivers from the online environment will be taken into the offline world and vice versa. Modelling this multichannel behavior is one of the key challenges for the near future.

Online Behavior

Modelling online choice behavior has become mature in a short period of time and from a model technical point of view one could say, "It is just another context with extra parameters."

From a marketing and strategic perspective it is much more challenging because consumer behavior is changing and the impact of the choice is not the same. Take for instance line optimization or pricing policy for razor blades—a typical product that is bought regularly and can be characterized as low frequency repeat buying. Although it is repeat buying, for most men it is a deliberate choice, buying the same blades over and again, but once in a while they are triggered to upgrade. Online buying reduces inconvenience so it adds value, because the blades are delivered automatically. What happens is that a repetitive choice has become a subscription choice. You don't choose repetitively and the brand loses its micro moment of reinforcement during the deliberate choice at time of purchase. Amazon, or whatever web shop, controls the subscription and erodes the brand value. In our choice models we should learn to cater to these changes and long-term effects, a great opportunity to experiment. In the meantime we should be critical in using the marketing metrics that are derived from offline behavior to optimize in a new multichannel reality.

As a researcher the real excitement is testing in real life online. Not by just observing, but by creating balanced designs to really know directly what is driving choice for whom, in tuning the offer to the individual sensitivity. It is the methodological rigor and creativity that provides researchers the opportunity and the "right" to act in real time. The need for speed forces market research to provide instant gratification. Testing in real time places the researcher in the cockpit.

In the eighties during interviews we predicted choice and respondents were amazed, but it took half a year before that product was developed and could be bought. For many products nowadays we test, learn, constantly adapt to changing individual needs and start the delivery process. That is the real paradigm switch, the excitement and the vibe of the eighties is back.

Dirk Huisman

Jeroen Hardon

MaxDiff on Mobile

Jing Yeh
Louise Hanlon
Millward Brown

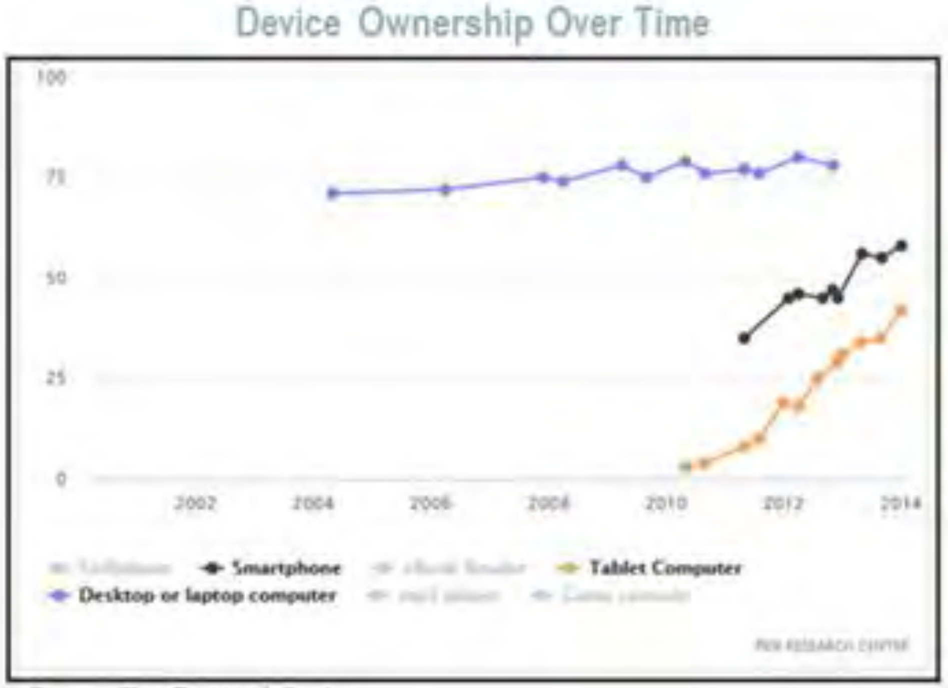
Abstract

As MaxDiff continues to be a household name in the market research industry, and as the world and human race become more technologically mobile; it is important for us researchers to understand the implications of device agnostic data collection for MaxDiff surveys. We provide learnings and recommendations from two case studies that employed MaxDiff on mobile devices.

Introduction

It is obvious that the world and human race are becoming more technologically mobile. The percentages of people who own smartphones or tablets have increased dramatically in recent years, while the percentage of people who own laptops/desktops has been relatively flat (see Figure 1).

Figure 1: Device Ownership Over Time

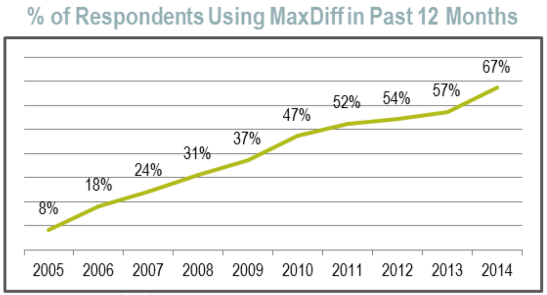


Source: Pew ResearchCenter

In order for market research to keep up with the increasingly technologically mobile human race, the industry has made advancements in adapting online surveys to mobile devices. Millward Brown has found that if market research surveys do not become mobile-enabled, our industry will increasingly run the risk of having inaccurate sample representation. Mobile surveying is becoming our only way of reaching some groups. For example, young males are a particularly hard to find group; and in the US, as much as 50% of Hispanics access surveys via mobile devices[1].

MaxDiff as a survey methodology continues to increase in popularity in the market research industry. 67% of respondents to Sawtooth Software's customer feedback survey reported using MaxDiff in 2014; up 10% from 2013 (see Figure 2). As MaxDiff continues to become an ever present tool in the research industry, it is important for researchers to identify usage guidelines for MaxDiff on mobile-enabled surveys.

Figure 2: MaxDiff Usage Over Time



Source: Sawtooth Software

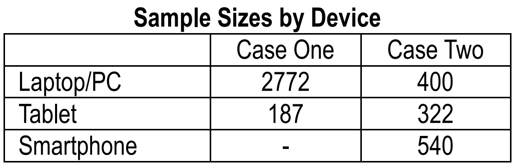
We learned from the 2013 Sawtooth Software Conference that conjoint on mobile works pretty well with, naturally, simpler respondent tasks showing more engaging the surveys (Diener, et al.; White). We also have an increasing amount of evidence that device type alone does not create large differences in response[2]. But what are the implications for employing MaxDiff on mobile devices? The research questions for this paper were:

- What is the impact of asking MaxDiff questions on mobile devices?
- How can we adapt MaxDiff designs to be more device agnostic?

METHODOLOGY

These research questions were examined using two project cases—Case One was for an online information provider and Case Two was a research and development project from Millward Brown and Kantar. Each case included a MaxDiff exercise as part of a larger project effort. Each case had groups of respondents who differed based on the type of device they used to take the survey—laptop/PC versus mobile platform. For the mobile group, Case One included tablets only, while Case Two had separate sample for examining tablet results as well as smartphone results. The sample sizes by device type for each case are shown in Figure 3.

Figure 3: Sample Sizes by Device for Each Case



Sample Sizes by Device

	Case One	Case Two
Laptop/PC	2772	400
Tablet	187	322
Smartphone	-	540

[1] Millward Brown R&D

[2] Millward Brown R&D

While device type was our main axis of comparison, we also examined the impact of demographics and estimation method (pooled versus isolated) in our analysis to ensure that any differences in our key metrics were truly driven by device.

Accounting for Demographics

Demographics were not matched by device during fieldwork, so we investigated the extent that results differ by device when demographics are matched and when they are not matched. The matching of demographics was done by randomly removing respondents from the device grouping that had a larger sample size until no differences in demographics remained.

Accounting for Utility Estimation Approach

Since the HB estimation approach borrows information from respondents when estimating utilities, results were computed based on pooled as well as isolated estimation. Pooled estimation refers to estimating utilities by pooling data across devices, then examining results by filtering the MaxDiff scores by device. Isolated estimation refers to estimating utilities separately by device.

In summary, we compared MaxDiff results by device, demographic matching, and estimation method. Within each case comparison cell, we examined the following key metrics: MaxDiff scores themselves, client recommendations that would result from the MaxDiff scores, hit rate, and percent certainty. Given the absence of a holdout task for both Case One and Case Two, the last MaxDiff question from each block was excluded from utility estimation and used to calculate hit rates.

Case One: Online Information Provider

Case One was based on a study for an online information provider in the USA who commissioned research to guide their strategy for increasing traffic to their website. The survey was 25 minutes in length and was conducted on laptop/PCs and tablets. Study qualification criteria included income and category relevancy specifications (e.g., 2002 or later vehicle owners and intenders). The MaxDiff, which was only one portion of the larger research effort, was designed to measure website features that motivated site visitation. 17 items were tested in MaxDiff using 11 screens per respondent and 4 items per screen. A summary of the research design is shown in Figure 4.

Figure 4: Case One Research Design Summary

Country of Fieldwork:	US
Category:	Online Information Provider
Length of Interview:	25 minutes
Sample Size:	Laptop/PCs (Shown as PCs throughout): n=2772 Tablets: n=187
MaxDiff:	Website features that would make them most/least likely to visit the site 17 items 11 screens per respondent 4 items per screen 10 blocks

As previously mentioned, MaxDiff results were examined by device, demographic matching, and estimation method, which resulted in the following cells shown in Figure 5:

Figure 5: Case One Sample Sizes Per Cell

MATCHED SAMPLE				NOT MATCHED SAMPLE			
Pooled N=386		Isolated		Pooled N=415		Isolated	
PC N=199	Tablet N=187	PC N=1852	Tablet N=187	PC N=228	Tablet N=187	PC N=2772	Tablet N=187

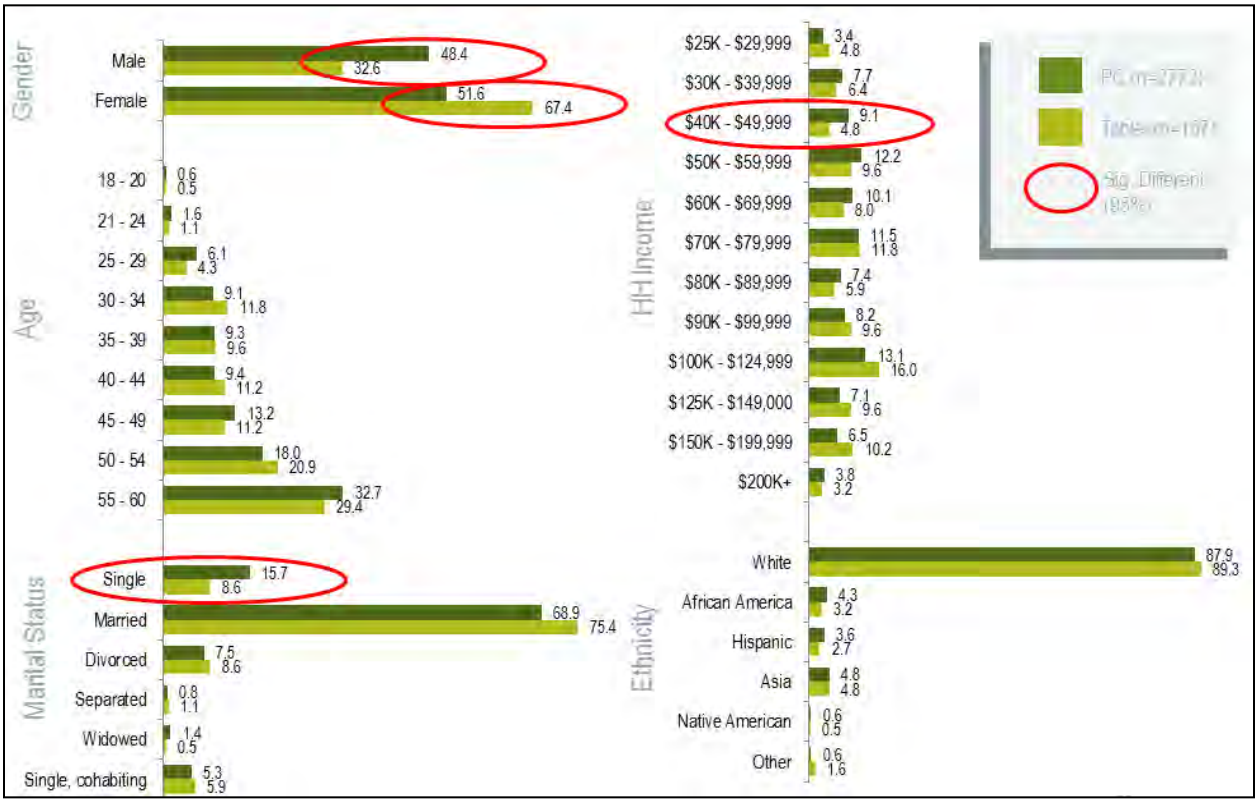
For matched sample, pooled estimation, PC sample size was dropped to N=199 to make it more comparable in size to tablet sample which was N=187. This was done to prevent PC responses from dominating the HB utility estimation.

Case One Findings

Different Types of Devices Attract Different Types of People

Basic profiling showed that there were some significant demographic differences based on the device used to complete the survey. As shown in Figure 6, PC users are more likely to be single (15.7% for PC versus 8.6% for tablet), male (48.4% for PC versus 32.6% for tablet), and earning less than tablet users (9.1% of PC earning $40,000–$49,999 versus 4.8% tablet).

Figure 6: Demographic Profiling of PC and Tablet Users

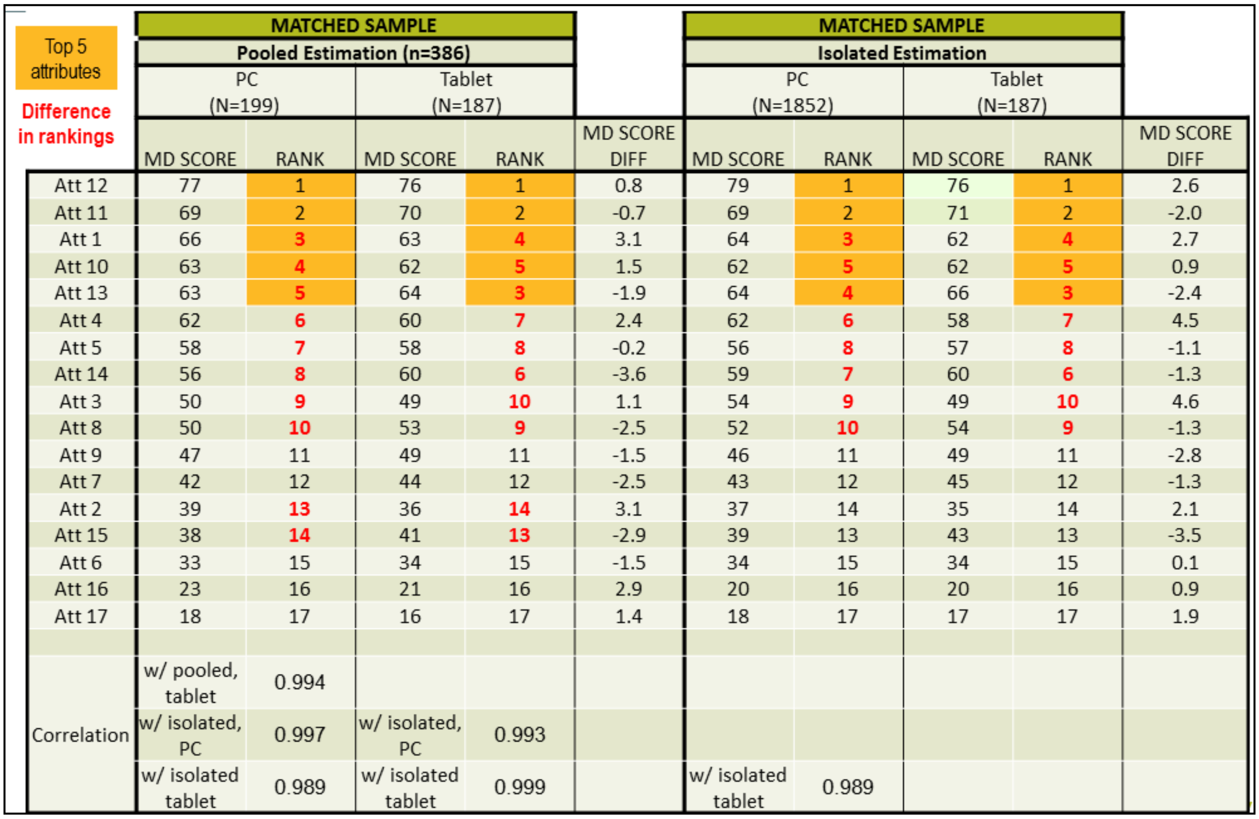


Devices Do Not Significantly Impact MaxDiff Results

We found that similar client recommendations would be reached using PC and tablet data regardless of demographic matching and utility estimation approach.

As shown in Figure 7, which displays MaxDiff scores by device and estimation approach with demographic matching; the top and bottom ranking items were the same by device. While there were differences in ranking for those items in the middle, absolute differences in scores here (and for all items) were minimal. Moreover, the gaps between items were preserved, in particular those in between the 1st and 2nd ranked items and 15th and 16th. As a result we observed extremely high correlations between PC and tablet MaxDiff scores using both estimation approaches.

Figure 7: MaxDiff Scores with Demographic Matching



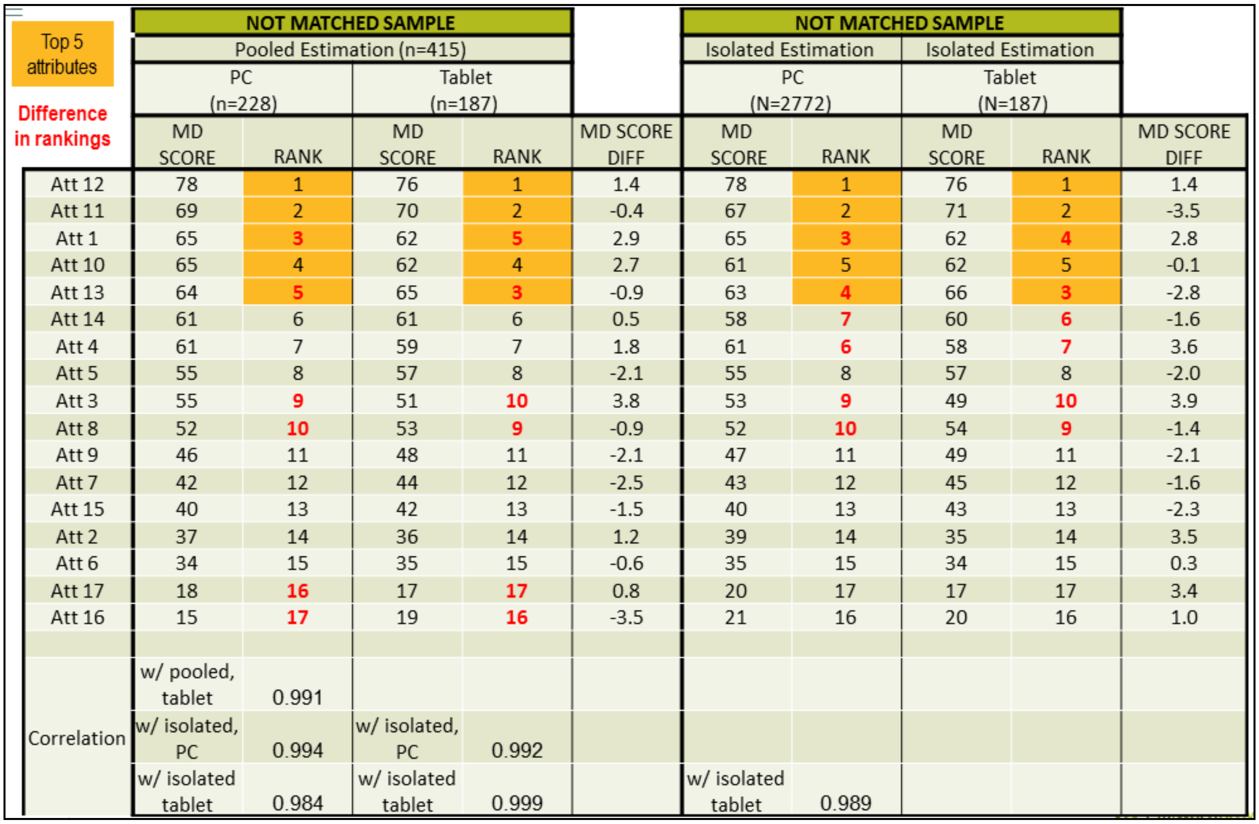
Top 5 attributes

Difference in rankings

	MATCHED SAMPLE					MATCHED SAMPLE				
	Pooled Estimation (n=386)					Isolated Estimation				
	PC (N=199)		Tablet (N=187)			PC (N=1852)		Tablet (N=187)		
	MD SCORE	RANK	MD SCORE	RANK	MD SCORE DIFF	MD SCORE	RANK	MD SCORE	RANK	MD SCORE DIFF
Att 12	77	1	76	1	0.8	79	1	76	1	2.6
Att 11	69	2	70	2	-0.7	69	2	71	2	-2.0
Att 1	66	3	63	4	3.1	64	3	62	4	2.7
Att 10	63	4	62	5	1.5	62	5	62	5	0.9
Att 13	63	5	64	3	-1.9	64	4	66	3	-2.4
Att 4	62	6	60	7	2.4	62	6	58	7	4.5
Att 5	58	7	58	8	-0.2	56	8	57	8	-1.1
Att 14	56	8	60	6	-3.6	59	7	60	6	-1.3
Att 3	50	9	49	10	1.1	54	9	49	10	4.6
Att 8	50	10	53	9	-2.5	52	10	54	9	-1.3
Att 9	47	11	49	11	-1.5	46	11	49	11	-2.8
Att 7	42	12	44	12	-2.5	43	12	45	12	-1.3
Att 2	39	13	36	14	3.1	37	14	35	14	2.1
Att 15	38	14	41	13	-2.9	39	13	43	13	-3.5
Att 6	33	15	34	15	-1.5	34	15	34	15	0.1
Att 16	23	16	21	16	2.9	20	16	20	16	0.9
Att 17	18	17	16	17	1.4	18	17	17	17	1.9
Correlation	w/ pooled, tablet	0.994								
	w/ isolated, PC	0.997	w/ isolated, PC	0.993						
	w/ isolated tablet	0.989	w/ isolated tablet	0.999		w/ isolated tablet	0.989			

Even when samples were not matched demographically, the same conclusions were reached across devices and estimation approaches as shown in Figure 8. The top 5 attributes were the same as when samples were matched. Still absolute differences were minimal and rank order was consistent, leading to the same client recommendation of what is important to consumers. Overall, device, demographic equalization and estimation technique do not impact business conclusions.

Figure 8: MaxDiff Scores with No Demographic Matching



Top 5 attributes

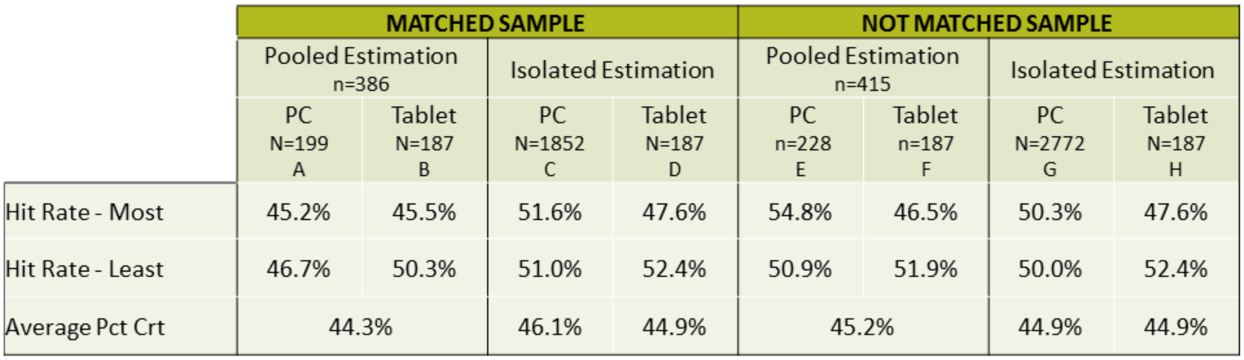
Difference in rankings

	NOT MATCHED SAMPLE					NOT MATCHED SAMPLE				
	Pooled Estimation (n=415)					Isolated Estimation		Isolated Estimation		
	PC (n=228)		Tablet (n=187)			PC (N=2772)		Tablet (N=187)		
	MD SCORE	RANK	MD SCORE	RANK	MD SCORE DIFF	MD SCORE	RANK	MD SCORE	RANK	MD SCORE DIFF
Att 12	78	1	76	1	1.4	78	1	76	1	1.4
Att 11	69	2	70	2	-0.4	67	2	71	2	-3.5
Att 1	65	3	62	5	2.9	65	3	62	4	2.8
Att 10	65	4	62	4	2.7	61	5	62	5	-0.1
Att 13	64	5	65	3	-0.9	63	4	66	3	-2.8
Att 14	61	6	61	6	0.5	58	7	60	6	-1.6
Att 4	61	7	59	7	1.8	61	6	58	7	3.6
Att 5	55	8	57	8	-2.1	55	8	57	8	-2.0
Att 3	55	9	51	10	3.8	53	9	49	10	3.9
Att 8	52	10	53	9	-0.9	52	10	54	9	-1.4
Att 9	46	11	48	11	-2.1	47	11	49	11	-2.1
Att 7	42	12	44	12	-2.5	43	12	45	12	-1.6
Att 15	40	13	42	13	-1.5	40	13	43	13	-2.3
Att 2	37	14	36	14	1.2	39	14	35	14	3.5
Att 6	34	15	35	15	-0.6	35	15	34	15	0.3
Att 17	18	16	17	17	0.8	20	17	17	17	3.4
Att 16	15	17	19	16	-3.5	21	16	20	16	1.0
Correlation	w/ pooled, tablet	0.991								
	w/ isolated, PC	0.994	w/ isolated, PC	0.992						
	w/ isolated tablet	0.984	w/ isolated tablet	0.999		w/ isolated tablet	0.989			

Devices Show the Same Predictive Power and Fit Measures

Hit rate of both the "most" and "least" items was comparable across devices, demographic samples, and estimation approaches (no significant differences at 95%). Percent certainty was also at parity across devices, demographic matching, and estimation method. (see Figure 9)

Figure 9: Hit Rates and Percent Certainty for Case One



	MATCHED SAMPLE				NOT MATCHED SAMPLE			
	Pooled Estimation n=386		Isolated Estimation		Pooled Estimation n=415		Isolated Estimation	
	PC N=199 A	Tablet N=187 B	PC N=1852 C	Tablet N=187 D	PC n=228 E	Tablet n=187 F	PC N=2772 G	Tablet N=187 H
Hit Rate - Most	45.2%	45.5%	51.6%	47.6%	54.8%	46.5%	50.3%	47.6%
Hit Rate - Least	46.7%	50.3%	51.0%	52.4%	50.9%	51.9%	50.0%	52.4%
Average Pct Crt	44.3%		46.1%	44.9%	45.2%		44.9%	44.9%

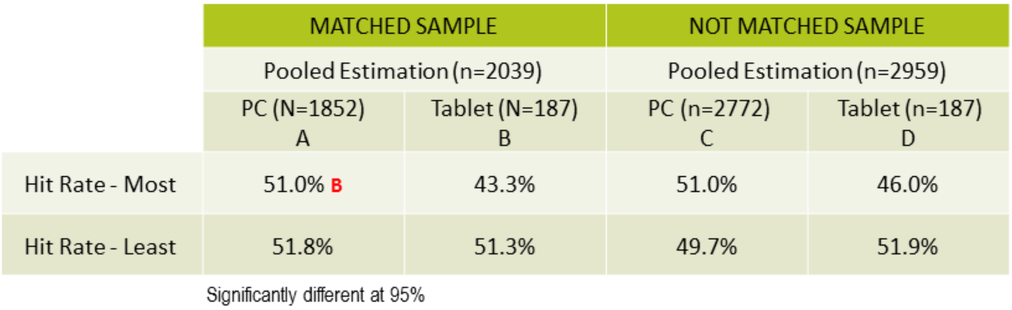
No significant differences found at 95%

Word of caution: For pooled utility estimation, ensure sample sizes per device are balanced to promote hit rate accuracy. We first calculated hit rates when sample sizes across devices were dominated by PC responses in the following ways:

- Matched sample: 91% of sample was PC, 9% tablet
- Unmatched sample: 94% of sample was PC, 6% tablet

We found that pooled estimation led to lower hit rates for tablets for "most" responses, as shown Figure 10, likely due to PC responses overshadowing tablet nuances. Once the sample composition was balanced, and tablet sample was comparable in size to the PC sample, no significant differences in hit rates occurred (as shown in Figure 9).

Figure 10: Hit Rates for Unbalanced Sample (Dominated by PC Responses)



	MATCHED SAMPLE		NOT MATCHED SAMPLE	
	Pooled Estimation (n=2039)		Pooled Estimation (n=2959)	
	PC (N=1852) A	Tablet (N=187) B	PC (n=2772) C	Tablet (n=187) D
Hit Rate - Most	51.0% B	43.3%	51.0%	46.0%
Hit Rate - Least	51.8%	51.3%	49.7%	51.9%

Significantly different at 95%

Case One Conclusions

- Different types of devices do attract different types of people; be aware when sampling.

- Devices do not significantly impact MaxDiff results. MaxDiff scores across devices result in similar conclusions, predictiveness, and model fit.

- So long as devices make up even proportions of the sample, pooled estimation of utilities provides the same results as estimation based on isolating each device.

Case Two: R&D from Millward Brown and Kantar

Case Two was based on a research and development project focused on device agnostic research. The subject of the survey was soft drinks. The survey was 20 minutes in length and fielded in the USA. To qualify for the survey, respondents must have purchased carbonated soft drinks at least once in the past year. Respondents took the survey on laptops/PCs, tablets, or smartphones. The MaxDiff, which was only one portion of the larger research effort, was designed to measure the persuasiveness of messages. 10 items were tested in MaxDiff using 5 screens per respondent and 4 items per screen. A summary of the survey design is shown in Figure 11.

Figure 11: Case Two Research Design Summary

Country of Fieldwork:	US
Category:	Soft Drinks
Length of Interview:	20 minutes
Sample Size:	PC n=400 Tablet n=322 Smartphone n=540
MaxDiff:	Dr Pepper messages measured on persuasion 10 items 5 screens per respondent 4 items per screen 20 blocks

MaxDiff results were examined by device, demographic matching, and estimation method, which resulted in the following cells shown in Figure 12:

Figure 12: Case Two Sample Sizes Per Cell

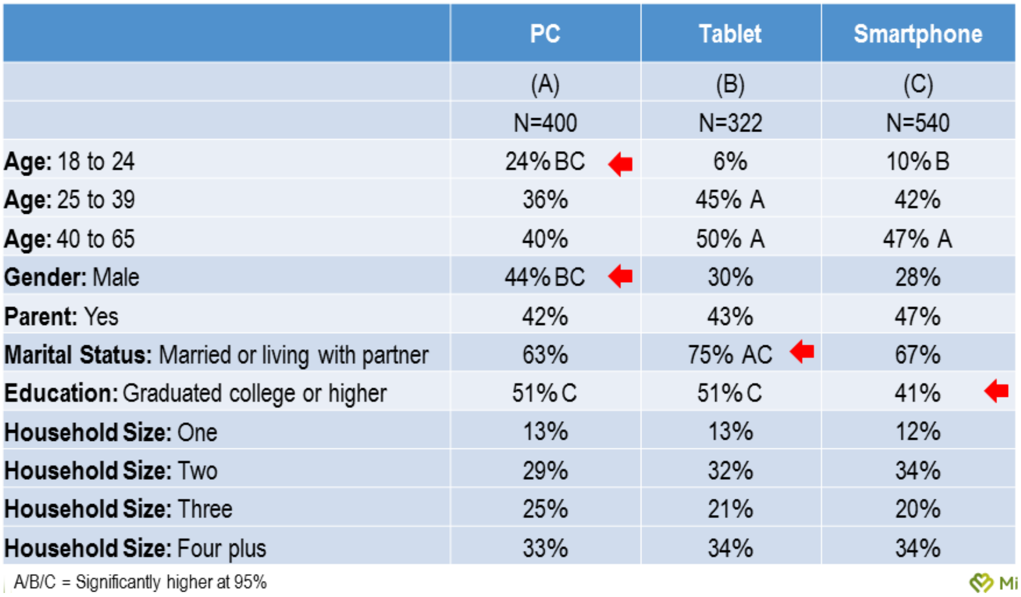
MATCHED SAMPLE						NOT MATCHED SAMPLE					
Pooled N=744			Isolated			Pooled N=1262			Isolated		
PC N=248	Tablet N=248	Smart N=248	PC N=248	Tablet N=248	Smart N=248	PC N=400	Tablet N=322	Smart N=540	PC N=400	Tablet N=322	Smart N=540

CASE TWO FINDINGS

Like in Case One, There Are Some Demographic Differences Across Devices for Case Two

PC survey-takers were more likely to be younger (24% age 18–24 for PCs versus 6% for tablet and 10% for smartphone) and male (44% for PC versus 30% for tablet and 28% for smartphone). Similar to Case One, PC users were more likely to be male and single compared to Tablet users (63% married for PC versus 75% for tablet). Tablet survey-takers were more likely to be married. Smartphone survey-takers were more likely to be less educated (41% college educated or higher for smartphone versus 51% for PC and tablet). (see Figure 13)

Figure 13: Case Two Demographic Differences by Device



	PC	Tablet	Smartphone
	(A)	(B)	(C)
	N=400	N=322	N=540
Age: 18 to 24	24% BC ⬅	6%	10% B
Age: 25 to 39	36%	45% A	42%
Age: 40 to 65	40%	50% A	47% A
Gender: Male	44% BC ⬅	30%	28%
Parent: Yes	42%	43%	47%
Marital Status: Married or living with partner	63%	75% AC ⬅	67%
Education: Graduated college or higher	51% C	51% C	41% ⬅
Household Size: One	13%	13%	12%
Household Size: Two	29%	32%	34%
Household Size: Three	25%	21%	20%
Household Size: Four plus	33%	34%	34%

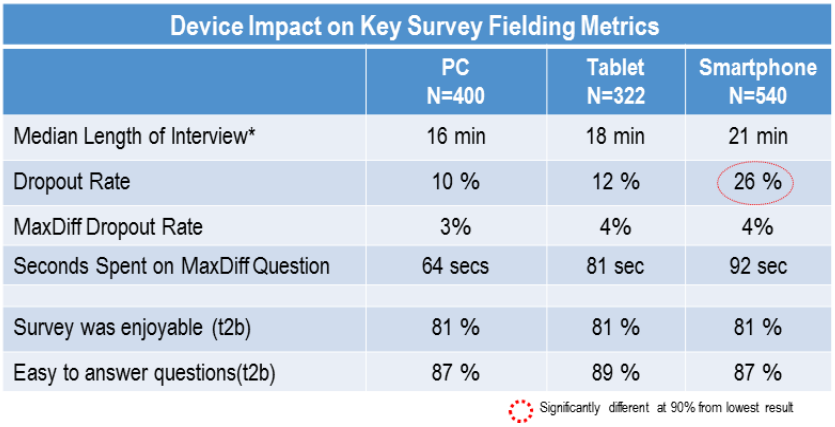
A/B/C = Significantly higher at 95%

Mi

Surveys on Smartphones Should Take Special Care to Be Concise and Easy to Complete

Key survey fielding metrics indicate that surveys on smartphones should take special care to be concise and easy to complete. Smartphone surveys took longer to complete (21 minutes for smartphones versus 16 minutes for PCs and 18 minutes for tablets) and had a higher dropout rate (26% versus 10% and 12%) as shown in Figure 14. Dropout rates for MaxDiff did not vary across devices (4% for smartphones versus 3% for PCs and 4% for tablets), likely due to MaxDiff's placement later in the questionnaire. However, MaxDiff questions on smartphones took longer to complete than other survey questions and devices (92 seconds on smartphones versus 64 seconds on PCs and 81 seconds on tablets).

Figure 14: Key Survey Fielding Metrics



Device Impact on Key Survey Fielding Metrics

	PC N=400	Tablet N=322	Smartphone N=540
Median Length of Interview*	16 min	18 min	21 min
Dropout Rate	10 %	12 %	26 %
MaxDiff Dropout Rate	3%	4%	4%
Seconds Spent on MaxDiff Question	64 secs	81 sec	92 sec
Survey was enjoyable (t2b)	81 %	81 %	81 %
Easy to answer questions(t2b)	87 %	89 %	87 %

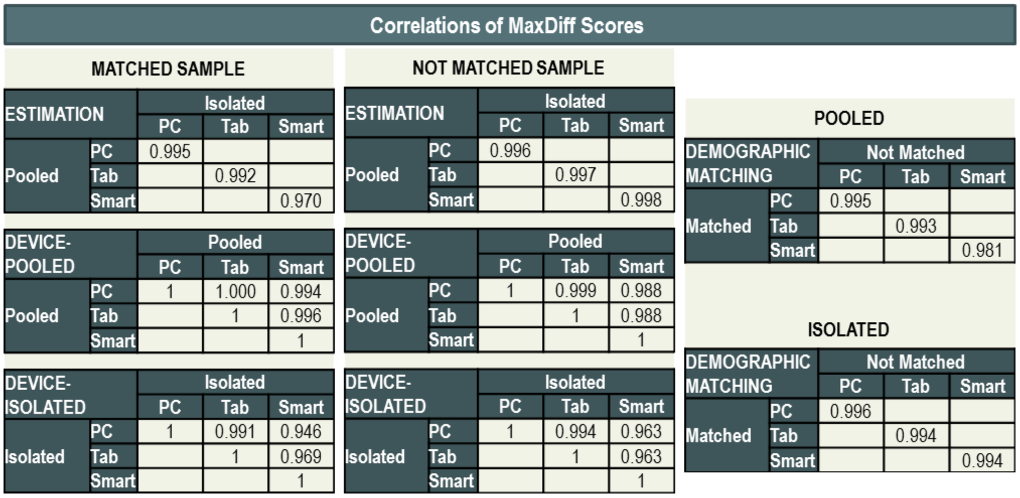
◌ Significantly different at 90% from lowest result

Despite differences in length of interview and dropout rates, survey satisfaction for smartphones was the same as PCs and tablets (81% top two box enjoyable across devices; and 87% top two box easy to answer versus 87% and 89%) as shown in Figure 14.

MaxDiff Scores Across Devices Are Very Similar

MaxDiff scores are highly correlated across estimation methods, devices, and demographic matching, as shown in Figure 15, indicating that the scores are very similar. Most correlations are about 0.99, with the lowest correlation at 0.946.

Figure 15: Correlations of MaxDiff Scores



Correlations of MaxDiff Scores

MATCHED SAMPLE

ESTIMATION		Isolated		
		PC	Tab	Smart
Pooled	PC	0.995		
	Tab		0.992	
	Smart			0.970

DEVICE-POOLED		Pooled		
		PC	Tab	Smart
Pooled	PC	1	1.000	0.994
	Tab		1	0.996
	Smart			1

DEVICE-ISOLATED		Isolated		
		PC	Tab	Smart
Isolated	PC	1	0.991	0.946
	Tab		1	0.969
	Smart			1

NOT MATCHED SAMPLE

ESTIMATION		Isolated		
		PC	Tab	Smart
Pooled	PC	0.996		
	Tab		0.997	
	Smart			0.998

DEVICE-POOLED		Pooled		
		PC	Tab	Smart
Pooled	PC	1	0.999	0.988
	Tab		1	0.988
	Smart			1

DEVICE-ISOLATED		Isolated		
		PC	Tab	Smart
Isolated	PC	1	0.994	0.963
	Tab		1	0.963
	Smart			1

POOLED

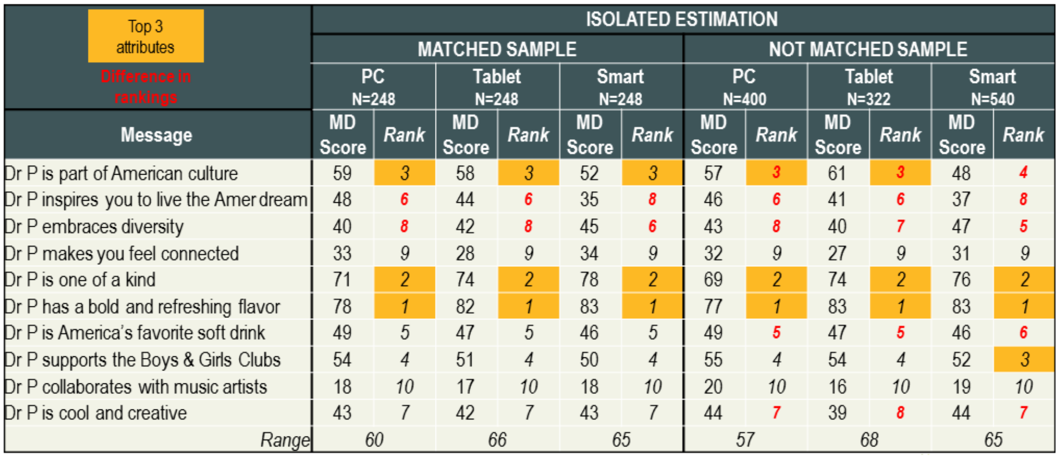
DEMOGRAPHIC MATCHING		Not Matched		
		PC	Tab	Smart
Matched	PC	0.995		
	Tab		0.993	
	Smart			0.981

ISOLATED

DEMOGRAPHIC MATCHING		Not Matched		
		PC	Tab	Smart
Matched	PC	0.996		
	Tab		0.994	
	Smart			0.994

Isolated estimation produced MaxDiff scores that led to similar client recommendations across devices. With isolated utility estimation, scores are pretty consistent across devices even without demographic matching, but the matching helps. The top 2 messages are the same with a relatively consistent gap between 1st and 2nd messages (5- to 10-point gap). The top 3 messages are pretty consistent, with a consistently sizeable drop from the 2nd to 3rd message (from 70's/69 to 50's/60's). There are some shifts in rankings in the middle and demographic matching helps. Ranges in scores become more in line with demographic matching. All in all, key takeaways from MaxDiff using isolated estimation across devices are very similar. Devices themselves essentially do not impact results, but the nuances of the demographic groups represented via the devices should not be overlooked. (see Figure 16)

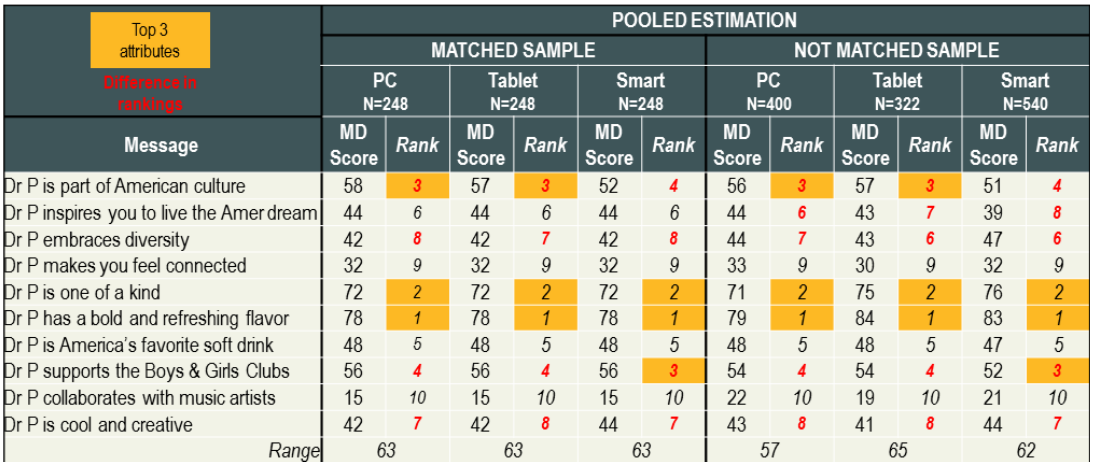
Figure 16: MaxDiff Scores for Isolated Estimation



Top 3 attributes Difference in rankings	ISOLATED ESTIMATION											
	MATCHED SAMPLE						NOT MATCHED SAMPLE					
	PC N=248		Tablet N=248		Smart N=248		PC N=400		Tablet N=322		Smart N=540	
Message	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank
Dr P is part of American culture	59	3	58	3	52	3	57	3	61	3	48	4
Dr P inspires you to live the Amer dream	48	6	44	6	35	8	46	6	41	6	37	8
Dr P embraces diversity	40	8	42	8	45	6	43	8	40	7	47	5
Dr P makes you feel connected	33	9	28	9	34	9	32	9	27	9	31	9
Dr P is one of a kind	71	2	74	2	78	2	69	2	74	2	76	2
Dr P has a bold and refreshing flavor	78	1	82	1	83	1	77	1	83	1	83	1
Dr P is America's favorite soft drink	49	5	47	5	46	5	49	5	47	5	46	6
Dr P supports the Boys & Girls Clubs	54	4	51	4	50	4	55	4	54	4	52	3
Dr P collaborates with music artists	18	10	17	10	18	10	20	10	16	10	19	10
Dr P is cool and creative	43	7	42	7	43	7	44	7	39	8	44	7
Range	60		66		65		57		68		65	

Pooled estimation also produced MaxDiff scores that led to similar client recommendations across devices. Across devices and demographic matching, the top performing messages were consistent. The top 2 messages are the same with a relatively consistent gap between 1st and 2nd messages (7- to 9-point gap). There is a consistently sizeable drop in persuasive ability from the 2nd to 3rd message (drop from 70's to 50's). There are some shifts in rankings in the middle, but scores are very close. Ranges in scores become more in line with demographic matching and sample size balance. PC and tablet scores are slightly more similar than smartphone. All in all, with pooled estimation, the key client takeaways are also the same across devices and demographic matching. (see Figure 17)

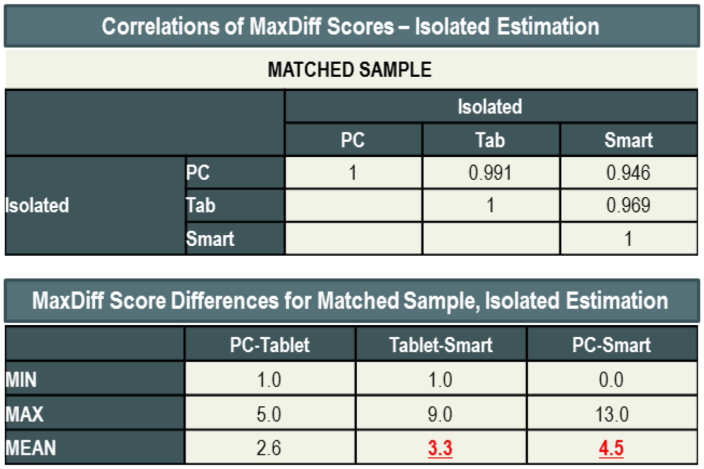
Figure 17: MaxDiff Scores for Pooled Estimation



Top 3 attributes	POOLED ESTIMATION											
	MATCHED SAMPLE						NOT MATCHED SAMPLE					
Difference in rankings	PC N=248		Tablet N=248		Smart N=248		PC N=400		Tablet N=322		Smart N=540	
Message	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank	MD Score	Rank
Dr P is part of American culture	58	3	57	3	52	4	56	3	57	3	51	4
Dr P inspires you to live the Amer dream	44	6	44	6	44	6	44	6	43	7	39	8
Dr P embraces diversity	42	8	42	7	42	8	44	7	43	6	47	6
Dr P makes you feel connected	32	9	32	9	32	9	33	9	30	9	32	9
Dr P is one of a kind	72	2	72	2	72	2	71	2	75	2	76	2
Dr P has a bold and refreshing flavor	78	1	78	1	78	1	79	1	84	1	83	1
Dr P is America's favorite soft drink	48	5	48	5	48	5	48	5	48	5	47	5
Dr P supports the Boys & Girls Clubs	56	4	56	4	56	3	54	4	54	4	52	3
Dr P collaborates with music artists	15	10	15	10	15	10	22	10	19	10	21	10
Dr P is cool and creative	42	7	42	8	44	7	43	8	41	8	44	7
Range	63		63		63		57		65		62	

Although MaxDiff results were similar across all devices, PC and tablet scores were more similar than smartphone scores. We hypothesized that PC and tablet scores would be more similar than smartphone scores. We used demographically matched sample with isolated estimation to get a pure read of scores by device. The findings supported our hypothesis, but also reiterated the large similarities in MaxDiff scores across devices in general (as previously shown). All correlations are very high, but the correlation between PC and tablet MaxDiff scores (r=0.991) is higher than both the correlation between PC and smartphone scores (r=0.946) and the correlation between tablet and smartphone scores (r=0.969). All mean differences are small, but the mean difference between PC and tablet MaxDiff scores (2.6) is smaller than both the mean difference between PC and smartphone scores (4.5) and the difference between tablet and smartphone scores (3.3). (see Figure 18)

Figure 18: Correlations and Mean Differences for Isolated Estimation, Matched Sample



Correlations of MaxDiff Scores – Isolated Estimation

MATCHED SAMPLE

		Isolated		
		PC	Tab	Smart
Isolated	PC	1	0.991	0.946
	Tab		1	0.969
	Smart			1

MaxDiff Score Differences for Matched Sample, Isolated Estimation

	PC-Tablet	Tablet-Smart	PC-Smart
MIN	1.0	1.0	0.0
MAX	5.0	9.0	13.0
MEAN	2.6	3.3	4.5

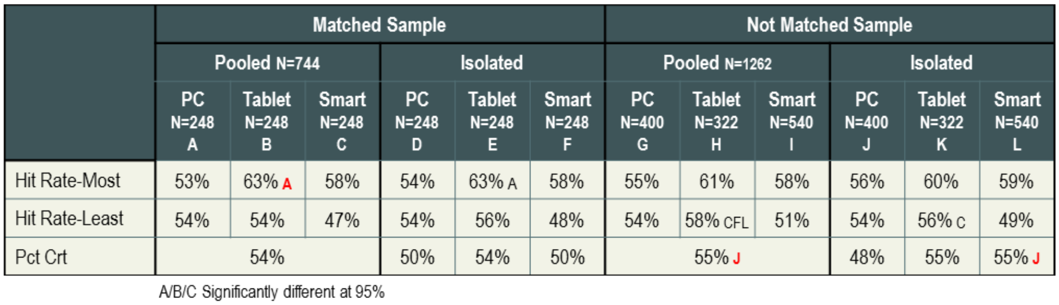
In Terms of Predictiveness and Fit, Devices Are at Parity for the Most Part

For the most part, there were no meaningfully significant differences in model accuracy across devices, estimation methods, and demographic equalization. Some notable differences, as shown in Figure 19, indicated that PC scores had somewhat lower accuracy as follows:

- Matched sample, pooled estimation: Tablets have a higher “hit rate for most” than PC (63% versus 53%).
- Not-matched sample, pooled estimation: Pooled estimation had a higher pct crt than isolated estimation for PC (55% versus 48%).
- Not-matched sample, isolated estimation: Smartphone had a higher pct crt than PC (55% versus 48%).

Although not proven here, perhaps respondent engagement on PCs is lower than respondent engagement on other devices. (see Figure 19)

Figure 19: Hit Rates and Percent Certainty



	Matched Sample						Not Matched Sample					
	Pooled N=744			Isolated			Pooled N=1262			Isolated		
	PC N=248 A	Tablet N=248 B	Smart N=248 C	PC N=248 D	Tablet N=248 E	Smart N=248 F	PC N=400 G	Tablet N=322 H	Smart N=540 I	PC N=400 J	Tablet N=322 K	Smart N=540 L
Hit Rate-Most	53%	63% A	58%	54%	63% A	58%	55%	61%	58%	56%	60%	59%
Hit Rate-Least	54%	54%	47%	54%	56%	48%	54%	58% CFL	51%	54%	56% C	49%
Pct Crt	54%			50%	54%	50%	55% J			48%	55%	55% J

A/B/C Significantly different at 95%

Conclusions for Case Two

- Take care to be concise with smartphone surveys as they take longer to complete and have higher dropout rates. In particular, MaxDiff questions take longer than other questions to complete.
- MaxDiff Scores across devices, estimation methods, and demographic matching lead to similar client recommendations.
- In terms of predictiveness and fit, devices were mostly at parity, with some weaknesses for PCs.
- Devices themselves essentially do not impact results, but the nuances of the demographic groups represented via the devices should not be overlooked.

Overall Learnings Based on Case One and Case Two and Potentials for Future Research

What Is the Impact of Asking MaxDiff Questions on Mobile Devices?

When sampling for device agnostic MaxDiff surveys, be aware that different devices will give rise to different types of consumers demographically, and these differences may vary depending on the population being studied. For both Case One and Case Two, demographic differences by device were evident. However, across Case One and Case Two, at times the differences themselves were different. For example: For Case One there were no age differences, but for Case Two PC users were younger than tablet users.

Although we found mostly similarities in the MaxDiff scores across devices, it is still important to examine results by device. Demographic differences by device are still evolving so it would be prudent for researchers to be diligent in examining utility estimations by device. In both Case One and Case Two, we saw that devices themselves essentially do not impact results, but the nuances of the demographic groups represented via the devices should not be overlooked.

Utilities can be estimated based on pooling data across devices so long as the sample contains a balance of sample sizes by device. As shown in Case One, pooled estimation using a sample that was about 90% PC respondents caused PC responses to overwhelm tablet nuances, resulting in a significantly lower hit rate for tablets. Once sample composition was balanced by device, there was no difference in hit rates across devices.

PC respondents may be less engaged than respondents on mobile platforms. Although hit rates were overall very close across devices, demographic matching, and estimation approaches; there were a few instances in Case Two where PC hit rates were lower than smartphone and tablet hit rates. Perhaps future research could further investigate this topic.

How Can We Adapt MaxDiff Designs to Be More Device Agnostic?

In terms of adapting MaxDiff designs to be device agnostic, simplicity helps. Both Case One and Case Two were relatively simple MaxDiff exercises—Case One employed 4 items per screen and 11 screens per respondent, Case Two employed 4 items per screen and 5 screens per respondent. The conjoint on mobile papers from the 2013 Sawtooth Software Conference showed that simpler conjoint tasks helped in utility estimation and respondent engagement (Diener, et al.; White). However, further examination of the impact of more complex MaxDiff designs on mobile platforms would be beneficial. Future R&D could investigate the consequences of simpler versus more complex MaxDiff designs on mobile platforms.

Surveys on smartphones should take special care to be concise and easy to complete. As we saw in Case Two, smartphone surveys take longer to complete and have higher dropout rates than surveys on PCs and tablets. Nonetheless, remember that respondent satisfaction was the same across PC's, tablets, and smartphones as shown in Case Two.

Final Thoughts

Although we as an industry have made progress on the device agnostic front of market research, overall the trends, and thus the learnings, are still evolving. We hope that the two cases discussed in the paper are helpful to this evolution and look forward to furthering our knowledge as a community.

Acknowledgements

Thank you to Sawtooth Software for accepting our paper for presentation and publication, to Chris Chapman for his thoughtful advice, and to members of our team at Millward Brown who supported us through our efforts. Special thanks to Priya Pola for helping us crank through all the iterations.

Jing Yeh

Louise Hanlon

References

Diener, C, Narang, R, Shant, M, Chander, H, and Goyal, M. (2013). Making Conjoint Mobile: Adapting Conjoint to the Mobile Phenomenon. *Proceedings of the 2013 Sawtooth Software Conference,* Dana Point, CA. October 2013.

White, Joseph. (2013). Choice Experiments in Mobile Web Environments. *Proceedings of the 2013 Sawtooth Software Conference,* Dana Point, CA. October 2013.

A Forecaster's Guide to the Future: How to Make Better Predictions

David Bakken
Foreseeable Futures Group

Humans have been trying to predict the future at least since the beginning of recorded history and probably ever since we first noticed such natural regularities as day following night and the moon passing through phases of illumination. In fact, we appear to be uniquely equipped to make predictions about the future. We have what appears to be an innate ability to detect patterns and to infer causal relationships. These abilities do not, however, make our predictions accurate. We often believe we've discovered meaningful patterns and causal relationships in random data. Consider the belief in the "hot hand" performance streaks that some athletes and gamblers seem to experience occasionally. Statistical analysis of actual streaks shows that they are indeed random, but widespread belief in the hot hand persists. In a series of experiments conducted by Andreas Wilke and H. Clarke Bennett (2009), subjects had to predict which of two images would next appear on a computer screen. The order was random but the subjects were more likely to guess that the next image would be the same as the current image. In other words, they expected a streak. Of course, we are just as likely to fall victim to the gambler's fallacy, the naïve belief that a short term losing streak of random events will very soon reverse itself. In other words, the "law of averages" requires things to correct themselves to make up for past runs.

We are also susceptible to explanations and predictions that "make sense" whether or not they have any grounding in reality. Remember the boom in all stocks Internet-related in the late 1990's? Many investors bought into the narrative that the economics of the Internet were different and the future prospects of an Internet company could not be determined using "old" approaches. As things have turned out, the most successful Internet companies, Amazon, Facebook and Google, make most of their money in old-fashioned ways.

Of course, market researchers understand randomness and have methods to separate the signal from the noise in a set of observations. Regression analysis and related tools have become the workhorses of prediction precisely because they quantify the degree of randomness ("error") in our data. But even statistical modeling can mislead us. Nassim Nicholas Taleb (2007) argues that we underestimate the probability of "rare" or "impossible" events at least in part because our assumptions about the underlying probability distributions are wrong and that, in fact, the "tails" of these distributions are fatter than we think.

Moreover, we tend to have selective memory when it comes to remembering the accuracy of our forecasts. We remember the bold or audacious predictions that happen to come true while forgetting most of the predictions that turn out to be wrong.

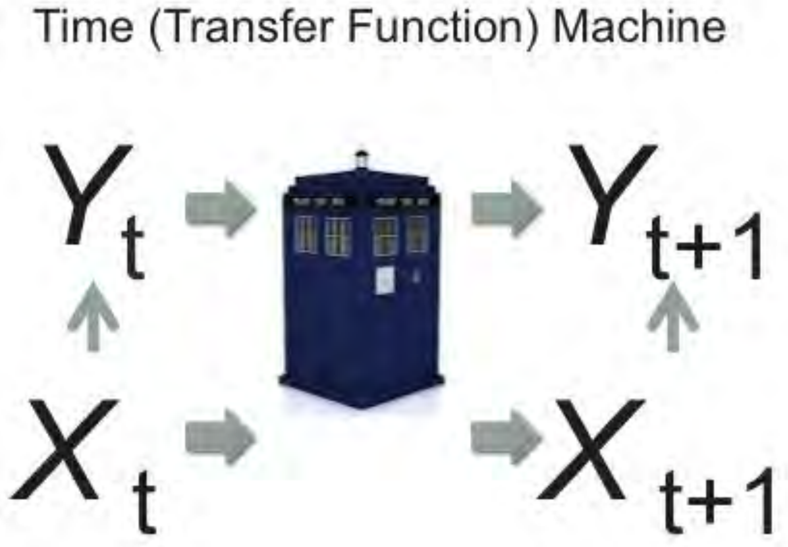
This paper looks at common causes of prediction failure and offers some ideas—along with examples—about how to make better predictions about the future.

The Prediction Paradigm

When it comes to predicting futures that are relevant to businesses, whether the future behavior of an individual consumer, the aggregate future behavior of a group of consumers, the future behavior of competitors, or disruptive changes in an industry, we tend to follow a single paradigm which is illustrated in **Figure 1** below.

We start with a set of *something*, here designated as ***Y**_t_*, which contains the values of the *something* elements at time ***t*** (this set could be empty). We want to know what the elements of set ***Y*** will look like at time ***t*+1**. In order to make this prediction we need some idea of the mechanism or *transfer function* that causes the elements in set ***Y*** to change over time (Dr. Who's Tardis time machine in Figure 1). While in some cases the transfer function might be intrinsic to the elements of ***Y*** (such as the rate of decay of a radioisotope), in many cases we look for an association between the elements of ***Y*** and the elements of some other set, ***X***. This trick depends, of course, on having a better idea of what the elements of ***X*** will look like in the future than we do for ***Y***. Sometimes the link between ***X*** and ***Y*** is causal; more often the link is merely an observed correlation.

Figure 1.



One common prediction task in marketing is forecasting sales for a new product. Market research might reveal the overall appeal of the product, often captured by asking a representative sample of customers how likely they are to buy the product, based on either a description or the actual product, and this *purchase intent* measure is translated into an estimate of *trial potential* (***Y***). Since consumers cannot purchase a product that they do not know about, *awareness* (***X***) should be correlated with the rate of *trial* or *adoption*. Because a marketer has some control over awareness through the amount spent on advertising and the effectiveness of the advertising execution, we can first state the expected level of awareness at ***t*+1**, **t+***2,* ***t*+3**, and so forth for as many time periods as we like (or until awareness reaches 100%). A really simple transfer function for the trial rate involves multiplying the total trial potential by the proportion of the target market that becomes aware by ***t*+1**, then subtracting that number from the total trial potential and repeating the calculation using expected awareness at ***t*+2,** a process that can be repeated either until the maximum projected awareness or the maximum trial potential is achieved.

The prediction paradigm for what we call "predictive analytics" or "predictive modeling" is subtly different from this paradigm for predicting the future. Instead of predicting the values of ***Y*** at some point in the future, predictive analytics focuses on predicting the values of ***Y*** for some set of unobserved units of observation under an assumption that the passage of time is irrelevant. As an example, a common use of predictive analytics is the assignment of consumers to a particular market segment based on a derived segmentation based on a different set of consumers. It does not matter if we perform that assignment today, next month, or next year. The assignment algorithm predicts segment membership based only on the endogenous relationships among the variables that define the segments, without reference to a specific point in time.

How the Future Happens

The time machine/transfer function in the prediction paradigm reflects our understanding of the way in which the particular future we are interested in will come about. The passage of time thus becomes an important element in our forecast, and we must specify how things change as a function of time.

Many of our notions about processes that bring about change over time are rooted in observations of the physical world. These notions include:

- Newton's laws of motion
- Growth, aging, and collapse
- Darwinian evolution
- Dynamic systems and equilibrium
- "If-then" causality

The perception of stable trends, for example, has a corollary in the Law of Inertia (e.g., a body in motion remains in motion unless acted upon by an external force; an object at rest remains at rest unless acted on by an unbalanced force). Similarly, the product life cycle (a sort of generic forecast model for any new product) has a metaphorical relationship to growth and aging.

Darwinian evolution defines a transfer function where external forces act on objects to "select" for certain traits or behaviors. This model requires a distribution of those traits or behaviors across the population of objects, a mechanism by which those characteristics undergo random changes (i.e., mutation), and varying external forces that interact with those traits and behaviors to determine if an individual object "survives" from one time period to the next. In the biological world, survival refers to an organism's ability to transfer its genetic code (in the form of DNA) to its offspring and their offspring. For a new product, survival is usually determined by profitability.

The transfer function for a dynamic system with equilibrium assumes that an initial instability or imbalance in the system is resolved by moving parts of the system around until equilibrium is achieved.

Finally, the "if-then" causal transfer function describes a dependency chain in which one event produces (with some observed or hypothesized likelihood) a second event which in turn may produce a third event, in the way that flipping a light switch closes a circuit which causes electricity to flow through a filament in a light bulb which heats up and emits visible and infrared radiation. This type of causality has to satisfy three important criteria: causes must precede effects, the cause and effect must co-vary, and other plausible explanations for the effect must be ruled out. These criteria mean “if-then” transfer functions must jump a relatively high hurdle.

There is one additional way in which the future comes about that may include elements of these other prototypical transfer functions. *Emergence* is a property of complex adaptive systems that arises in non-obvious or non-intuitive ways and reflects the way in which different parts of a system that include autonomous decision-making agents (such as buyers and sellers) interact to produce complex behaviors.

WHY FORECASTS ARE OFTEN INACCURATE AND PREDICTIONS OFTEN FAIL

Despite our inherent tendencies toward predicting the future—an adaptive mechanism that must have increased our ancestors’ chances of survival—we are not particularly good at predicting the future. Consider that every new product introduction reflects a prediction (albeit, some more scientific than others) that this product will “beat the odds” and become a market success, yet somewhere between 35% and 90% (depending on the industry) of new products end up failing (either not achieving the predicted success or disappearing from the market altogether).

Fortunately for us, we can learn (in theory, at least) by looking at prediction failures in hope of finding patterns that explain those failures. In the author’s experience with forecasting and prediction in marketing, six patterns stand out.

- The prediction or forecast relies entirely on historical correlations among the predictor variables and the predicted variables.
- The prediction or forecast requires too many assumptions (often unsupported) relative to factual inputs.
- The forecast model has many stochastic inputs relative to deterministic inputs.
- There is no underlying causal model or transfer function to explain how the particular future in question will come about.
- The forecasting model is too simple.
- The forecasting model is too complex.

1. Relying on Historical Correlations

Regression analysis and related methods are the primary tools for finding patterns of association in historical data that can be used to predict the values of some variable for unobserved instances, which includes the future as well as out-of-sample units of observation. Perhaps the best known methods are econometric or “causal” modeling and time-series analysis. Econometric modeling relies on ordinary least squares (OLS) regression, linking some set of “explanatory” variables (***Xs***) to an outcome variable (***Y***). The future value of ***Y*** is conditional on the future values of the ***Xs***, as in the basic paradigm described above. Time-series analysis assumes that all or most of the information needed to predict the future of a variable is contained in the historical trend for the variable, which can be decomposed into a time trend, a seasonal factor, a cyclical element, and an error term. Time-series models are really elaborate extrapolations from historical data.

Forecasts based on regression analysis of historical data are susceptible to multiple sources of error, including specification error (model likelihood has the wrong functional form). There are at least two additional ways in which reliance on correlations in historical data can lead to forecasting and prediction problems. The first is over-fitting, where the model estimated by the analysis (essentially, the coefficients and intercept in a regression model) fits the sample data extremely well but does a poor job of predicting out-of-sample cases. Over-fitting is an inherent problem for all models based on sampled observations, and various statistical methods exist for minimizing the impact of over-fitting.

The second is known as conditioning error, wherein the value of x_{t+1} on which the forecast is conditioned may be wrong. This problem becomes intractable when the historical data does not include all possible manifestations of the explanatory variables or the variable we are trying to predict. If a particular manifestation or value does not occur in the historical data, we have no reliable way to predict the occurrence of those events or their impact on the dependent variable.

2. Too Many Assumptions

Almost all forecasting and prediction problems require that we make assumptions about the likely future values of at least a few factors. For example, in forecasting the volume for a new prescription drug we need to establish the size of the indication market—the number of individuals in the population who will be diagnosed with the condition that the drug treats—in the future. The size of that market may well depend on the availability of a good diagnostic test, so we might undertake a forecast under the assumption that such a test will enter the market at a specific future date. Similarly, in forecasting trial and repeat rates for new fast moving consumer goods we might make assumptions about the level of awareness from advertising in each 4-week time period following the launch of the product.

These assumptions cause problems when they are unsupported, when we test only one or a few values out of a larger set of likely or plausible values, and when they outnumber the factual or evidence-based inputs to the forecast.

One of two things might happen when we have many assumptions relative to factual inputs. First, the assumptions will have more impact on forecast variability (and accuracy) than the factual inputs. This has been observed in new product forecasts based on concept evaluations and simulated test markets, where as much as two-thirds of forecast variability may be due to the marketing plan assumptions. Additionally, having too many assumptions tends to reduce the realism (and hence, validity) of the forecasting model. An example is the original Bass (1969) model of adoption and diffusion for new consumer durable goods. This model has a number of limiting assumptions, such as no other innovations in the market, that limit its realism.

Finally, when we make forecasts to help decide among different options, the assumptions may either add no information (if we have the same assumptions across the board) or incorrectly differentiate the options unless we have strong (i.e., evidence-based) reasons for having different assumptions for different options.

3. Too Many Stochastic Inputs

One way to handle forecasting inputs that cannot be determined in advance is to define a probability distribution for each of these inputs and then randomly draw a value for that input. The usual practice is to repeat this process for many iterations and then average across the many forecasts. With even a few stochastic inputs the forecast can quickly become dominated by randomness. Stochastic inputs should be considered where they reflect truly random processes in the system we are trying to forecast, such as the probability that a given individual will become aware of a new product from any source in a given time period, but when used to reflect our ignorance of key model features, they can overwhelm what we do know.

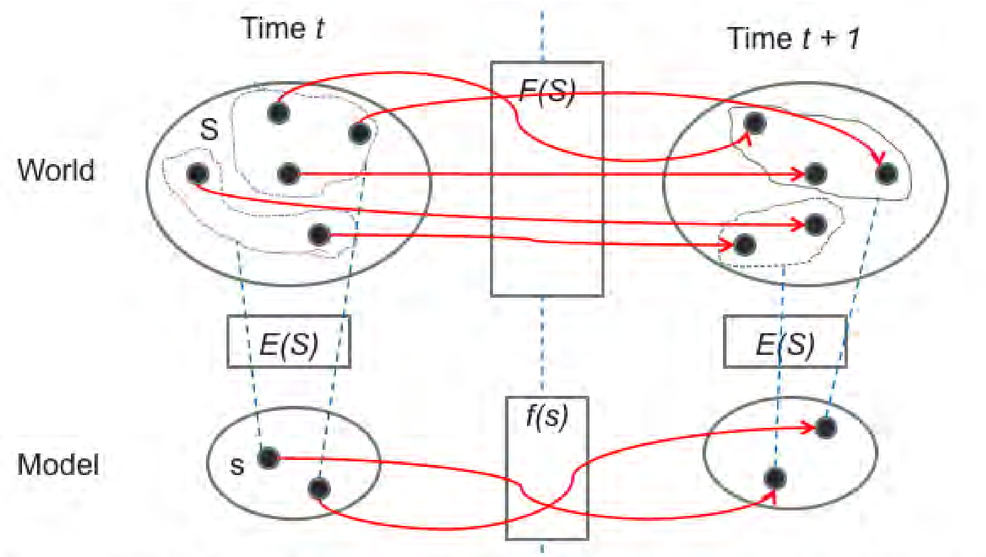
4. No Underlying Causal Model or Transfer Function

This is related to the problem of relying solely on historical data but may also apply to forecasts that rely heavily on Monte Carlo simulation (see point 3 above). It is very difficult to validate a forecasting approach without some idea of the underlying mechanism that creates the future of interest. Some predictive analytic approaches (e.g., path analysis and structural equation modeling) at least require that we think about the causal model and most experienced modelers realize that finding the right model specification requires some hypothesizing about how the predictor variables influence the outcome of interest. One problem with the machine learning approaches used in data mining and predictive analytics is that they usually are “model agnostic.” A neural network, for example, typically produces one or more “hidden layers” of weights or coefficients that link the input variables to the output states, making it difficult to both specify and test hypotheses about underlying causal models.

5. The Model Is Too Simple

Even when we have some hypothesis about the underlying transfer function or causal model we can run into trouble if we have over-simplified the model. As a general rule, a model should be as complex as it needs to be to represent the system of interest but not more complex. **Figure 2** below illustrates the way in which a forecast model might be simplified relative to the world it represents. In the real world we have five input variables, with transfer functions that describe how these variables change from time ***t*** to time ***t*+1**. The model aggregates or simplifies these five variables into only two factors, through the function designated ***E(S).*** At time ***t*+1** we need to reverse this function to get from the prediction back to the real world.

Figure 2.



Miller, J.H. and S.E. Page (2007). *Complex adaptive systems: An introduction to computational models of social life.*

The validity of the model depends crucially on the degree to which the simplification captures the important elements of the system.

6. The Model Is Too Complex

While leaving an important element out of a forecast model can be catastrophic, including too many elements has drawbacks as well. While it is a good idea to identify all the parts of the system before developing a forecast model, we seldom will have all the information we need to set values for all of the inputs in a model that completely represents the world. If we did, we would not need a "model" of the world.

Adding detail to a forecast model adds noise—each additional input has some associated error—as well as complexity that can make it difficult to determine which inputs have the biggest impact on the outcomes of interest. More detail also makes model validation difficult because of the number of factors that must be matched to actual observations.

What's a Forecaster to Do?

A crucial initial step in tackling any forecasting or prediction problem should be an attempt to describe the system we are trying to model in as much detail as possible with the aim of identifying the transfer functions that link the current state of the system to the future state. This is true whether our forecasting method will be a regression model based on historical data, a Monte Carlo simulation, or even a subjective, qualitative approach such as the Delphi method (an early, structured form of "crowdsourcing").

For example, imagine that we want to predict the demand for in-patient hospital stays in the United States ten years from now. We know, almost without thinking, that this demand will be some function of the characteristics of the population at that future date. In particular, because older people are more likely to be hospitalized for a variety of conditions, our model of the system will include the aging of the population over that time period. In this case we are likely to have good historical data and we can estimate a regression model of the association between age, hospital admissions, and perhaps some other covariates, such as individual health states (in case today's 65-year-olds tend to be healthier than their 75-year-old counterparts were at age 65, for example). In order to make our forecast, however, we must make some guesses about the future values of the predictor variables in our model. This might be as simple as projecting current demographic trends (age, mortality, etc.) ten years into the future. More likely is a situation where at least some of the variables our forecast is conditioned on are not completely predictable from their current states. This is especially true when the mechanism that will bring about the future involves random variables, evolutionary selection, or emergence.

Computer simulation is an effective way to deal with forecasting uncertainty arising from randomness in the predictor variables, evolution, and emergence. In the following sections we will look at examples of computer simulation techniques that incorporate these processes.

Simulating the Future

If we are working with historical data, a simple regression model that predicts hospital admissions from age and (for simplicity) health state at age 65 still does not account for the time-dependent aging of the population. We need some process to estimate how today's population of 65-year-olds might change on the way to age 75. One way to do this is to simulate the aging process. Some of today's 65-year-olds will not survive to age 75. Some will get cancer or other diseases that are likely to increase the odds of hospitalization, so our simulation will need to take these factors into account.

There are two basic methods for implementing this type of aging simulation: micro-simulation and systems dynamics modeling. In this case, both methods would start by grouping individuals into age cohorts. Micro-simulations represent a population as a group of heterogeneous individuals. Systems dynamics simulations treat all individuals in a cohort as identical. Both methods are similar in that transition probabilities act on the members of a cohort to determine what happens at the next time period. For example, there is a certain probability that an individual who is 65 years old will die before reaching age 66. In a systems dynamic model, all 65-year-olds have the same probability of dying before age 66; in a micro-simulation, each individual 65-year-old can have a different probability (which might be a function of other characteristics such as the presence of a particular medical condition).

Simulations with Interactions and Decision-Making Agents

While micro-simulations and systems dynamics models may have many components, they are relatively simple in that the individuals do not interact in any way with each other or with an external environment. If we are trying to forecast in-patient hospital stays this may not be a problem, but for many other forecasting problems understanding and capturing these interactions may be critical to accurate forecasting.

In the aging micro-simulation described above, the probability of one individual dying in a particular time period has no impact on the probability of another individual dying in that time period or a subsequent time period. But suppose that there is such an impact. Maybe the death of a spouse increases the likelihood of an individual dying in the same or subsequent time period. Now we have an interaction between individuals and events that we need to account for, and the future behavior of the system will be determined by the strength of these interactions.

The following simulation was designed to examine the impact of endogenous factors relative to external shocks on the rate of extinction in a simulated world inhabited by 100 species that occupy different ecological niches. This simulation is based on a model described by Ormerod (2005) and allows us to compare two different models of the extinction process:

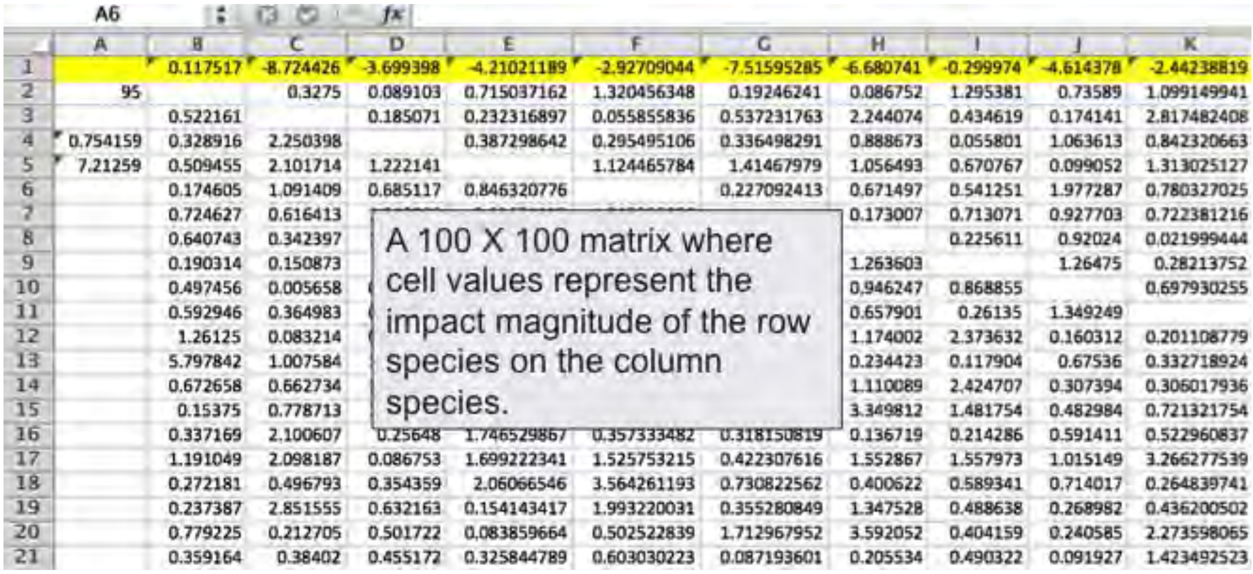
- An external shock model (proposed by Mark Newman, as cited by Ormerod, 2005) in which a species becomes extinct only when some random external shock (such as a drought) causes a stress that exceeds the species' ability to cope with that stress.

- A model (Ormerod, 2005) that combines endogenous effects with these external shocks. In this model, the outcome for one species might have a negative or positive effect on some other species. For example, if a predator dies out, the prey species benefits; if a prey species dies off, the predator suffers.

Each of the 100 species has one important characteristic, resilience, which determines the species' ability to withstand negative impacts. The higher this number, the more likely it is that the species will survive the effects of negative impacts. Because these are hypothetical species, resilience can be assigned simply by drawing values at random from a normal distribution. The mean and standard deviation of this distribution can be varied to determine the effect of diversity in the population. To avoid oversimplifying by having resilience be constant for a given simulation run, we have included a mutation process that changes the resilience of a small number of randomly selected species at each time step of the simulation. The mutation rate is a variable input.

We also establish pairwise connections among a proportion (another variable input) of species, and we vary the potential magnitude and direction of these impacts. **Figure 3** shows the worksheet in an Excel workbook that contains the potential magnitude of impact between each pair of species (the cell values contain the impact of the row species on the column species). These values are drawn at random from an exponential distribution so that most impacts will be of small magnitude but there will be a few large impacts. However, most pairwise connections are not enabled, and for those that are, the impact can be either positive (increasing the chances of a species' survival) or negative (reducing its chances). **Figure 4** shows the pairwise interactions we generate. Every pairwise species combination is determined by evaluating a random value drawn from a standard uniform distribution. If the value drawn is *less than* the expected incidence of negative impacts (a variable input), the pairwise interaction is assigned a value of -1. If the random value is *greater than* 1 minus the incidence of *positive* impacts, the pairwise interaction is assigned a value of +1. All other pairwise interactions are set to zero impact. The resulting matrix is multiplied by the corresponding matrix of *impacts* (Figure 3) to create the worksheet illustrated in Figure 4.

Figure 3. Pairwise Species' Impact Magnitude

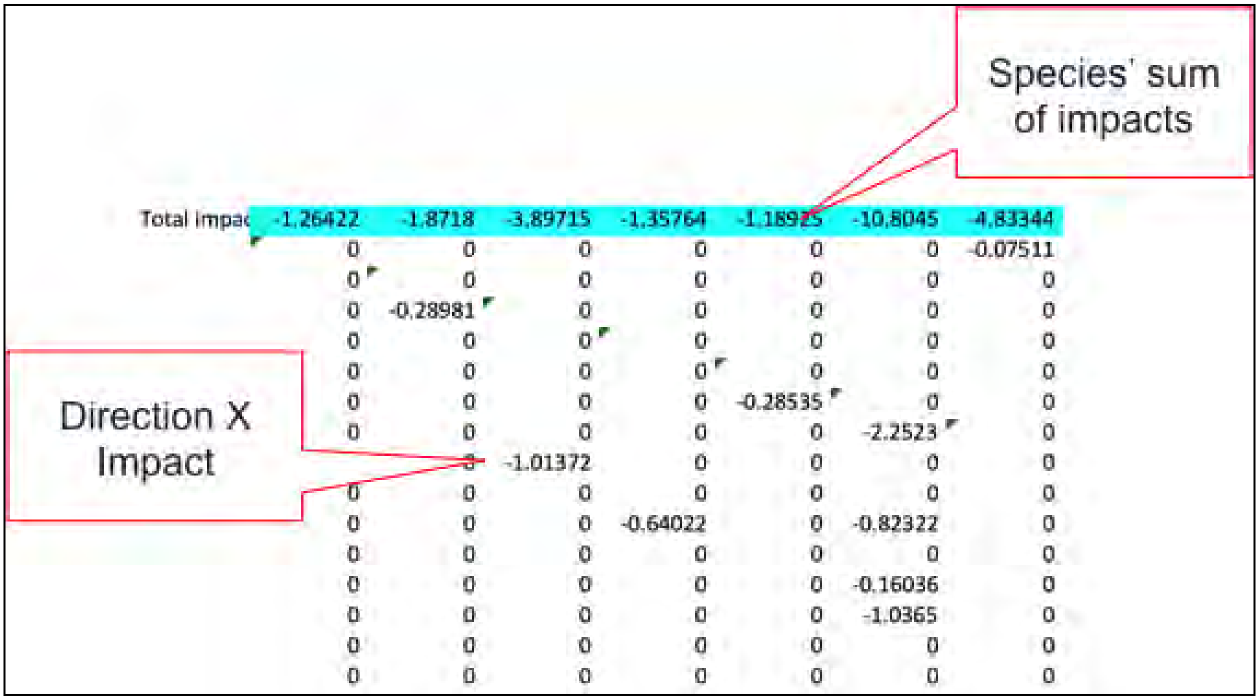


A6

	A	B	C	D	E	F	G	H	I	J	K
1		0.117517	-8.724426	-3.699398	-4.21021189	-2.92709044	-7.51595285	-6.680741	-0.299974	-4.614378	-2.44238819
2	95		0.3275	0.089103	0.715037162	1.320456348	0.19246241	0.086752	1.295381	0.73589	1.099149941
3		0.522161		0.185071	0.232316897	0.055855836	0.537231763	2.244074	0.434619	0.174141	2.817482408
4	0.754159	0.328916	2.250398		0.387298642	0.295495106	0.336498291	0.888673	0.055801	1.063613	0.842320663
5	7.21259	0.509455	2.101714	1.222141		1.124465784	1.41467979	1.056493	0.670767	0.099052	1.313025127
6		0.174605	1.091409	0.685117	0.846320776		0.227092413	0.671497	0.541251	1.977287	0.780327025
7		0.724627	0.616413					0.173007	0.713071	0.927703	0.722381216
8		0.640743	0.342397						0.225611	0.92024	0.021999444
9		0.190314	0.150873					1.263603		1.26475	0.28213752
10		0.497456	0.005658					0.946247	0.868855		0.697930255
11		0.592946	0.364983					0.657901	0.26135	1.349249	
12		1.26125	0.083214					1.174002	2.373632	0.160312	0.201108779
13		5.797842	1.007584					0.234423	0.117904	0.67536	0.332718924
14		0.672658	0.662734					1.110089	2.424707	0.307394	0.306017936
15		0.15375	0.778713					3.349812	1.481754	0.482984	0.721321754
16		0.337169	2.100607	0.25648	1.746529867	0.357333482	0.318150819	0.136719	0.214286	0.591411	0.522960837
17		1.191049	2.098187	0.086753	1.699222341	1.525753215	0.422307616	1.552867	1.557973	1.015149	3.266277539
18		0.272181	0.496793	0.354359	2.06066546	3.564261193	0.730822562	0.400622	0.589341	0.714017	0.264839741
19		0.237387	2.851555	0.632163	0.154143417	1.993220031	0.355280849	1.347528	0.488638	0.268982	0.436200502
20		0.779225	0.212705	0.501722	0.083859664	0.502522839	1.712967952	3.592052	0.404159	0.240585	2.273598065
21		0.359164	0.38402	0.455172	0.325844789	0.603030223	0.087193601	0.205534	0.490322	0.091927	1.423492523

A 100 X 100 matrix where cell values represent the impact magnitude of the row species on the column species.

Figure 4. Final Pairwise Species' Impacts



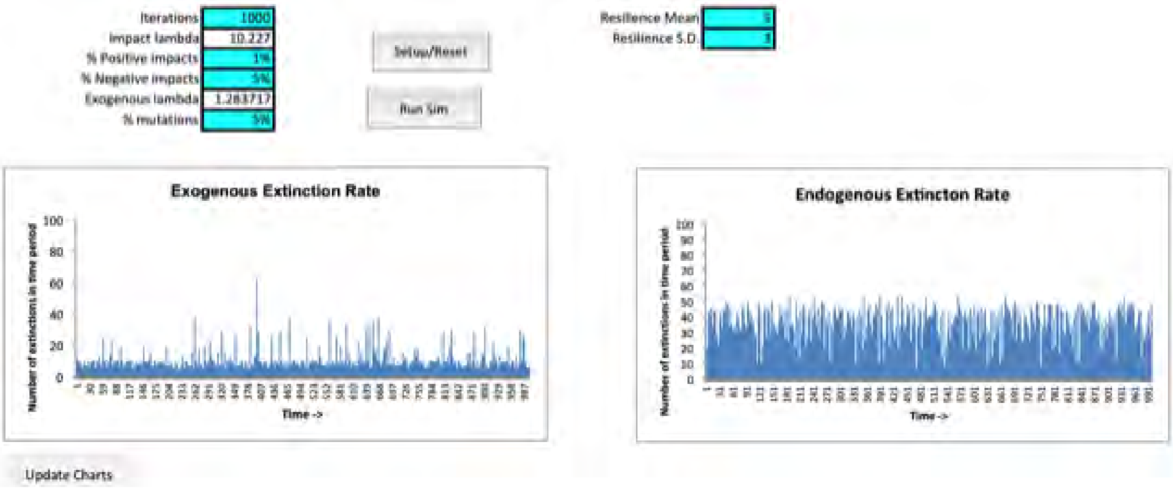
Species' sum of impacts

Direction X Impact

Total impac	-1.26422	-1.8718	-3.89715	-1.35764	-1.18925	-10.8045	-4.83344
	0	0	0	0	0	0	-0.07511
	0	0	0	0	0	0	0
	0	-0.28981	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	-0.28535	0	0
	0	0	0	0	0	-2.2523	0
		0	-1.01372	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	-0.64022	0	-0.82322	0
	0	0	0	0	0	0	0
	0	0	0	0	0	-0.16036	0
	0	0	0	0	0	-1.0365	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0

This simulation is designed to allow comparison of the two different extinction models described above. The simulator includes a user interface (**Figure 5**) that permits setting of variable inputs such as the incidence of positive and negative impacts and the mutation rate, and displays the output from both models.

Figure 5. Extinction Simulator Interface



Iterations 1000
impact lambda 10.227
% Positive impacts 1%
% Negative impacts 5%
Exogenous lambda 1.283717
% mutations 5%

Resilience Mean 5
Resilinnce S.D. 1

The actual simulation process is straightforward. At each time step a value for a systemic external shock is drawn at random from an exponential distribution. For the external effects model, each species' inherent resilience is evaluated against this random shock. If the shock is greater than a species' resilience, the species dies off in that time step. If not, the species survives to the next time step.

For the endogenous model, the sum of positive and negative impacts for each species (i.e., those illustrated in **Figure 4**) is evaluated against that species' resilience. If the sum of impacts plus resilience is less than zero, the species dies off in that time step. Then, before the next step, the impact matrix (i.e., **Figure 4**) may be updated in various ways, with details we won't get into in this paper.

We can also examine a combined model in which both external shock and the impact of connected species determine the survival of each species at each time step. **Figures 6, 7** and **8** illustrate the results of three simulations: external shock, endogenous impacts, and combined external and endogenous effects.

Figure 6. External Shock Model

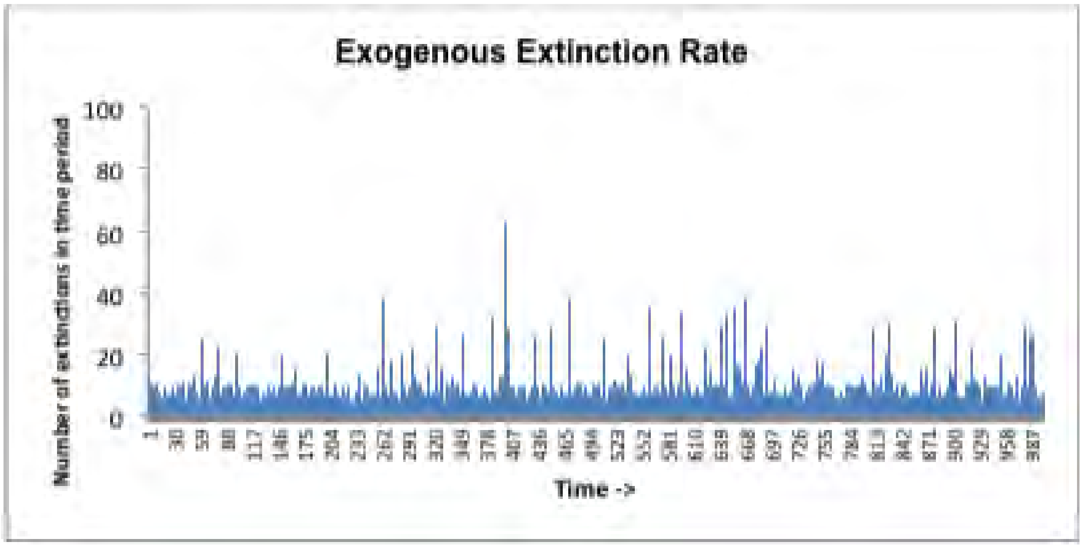


Figure 7. Endogenous Shock Model

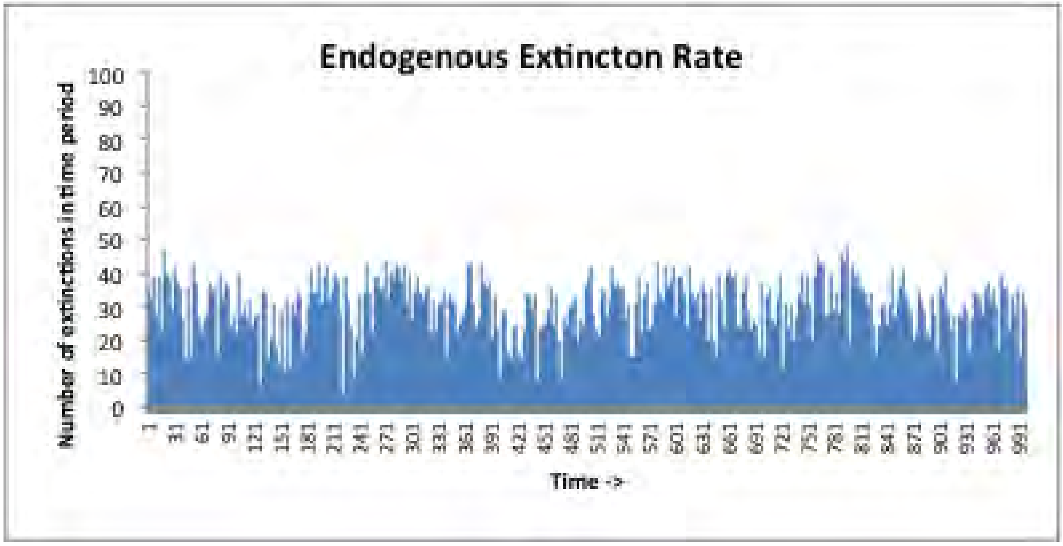
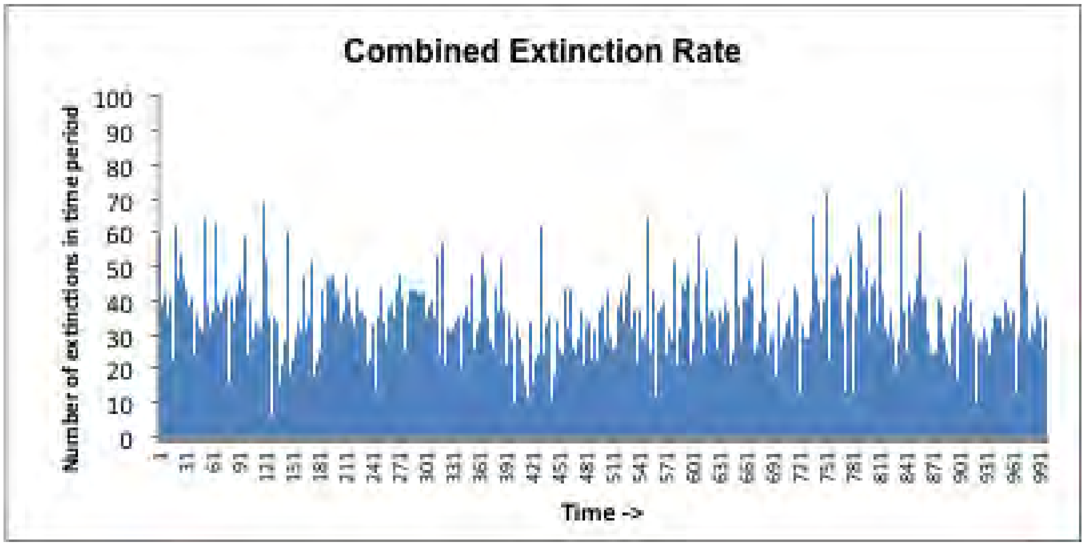


Figure 8. Combined External and Endogenous Shocks



In the external shock model (Figure 6) we observe a fairly low steady background rate of extinction (around 10% of species) and, occasional significant extinctions (20–25%) and rare massive extinction (>35% of species). When there are only endogenous impacts (Figure 7), we see a different picture, with a higher average rate of extinction but no extreme extinctions. Finally, in the combined model (Figure 8) we have an even higher background rate (we added another negative impact in each time step) and more extreme extinctions, including one that eliminates more than 90% of species.

Our extinction simulator is a vastly simplified model of the world but demonstrates the potential impact of incorporating heterogeneity in agent characteristics and interactions among agents (the species in this case). Our mechanism of inter-species impact is relatively crude and does not allow the species to take any actions (such as migrating) that might increase their chances of survival. The simulation has a simple species replacement model that insures there are always 100 species at every time step. How might we develop simulations that are more realistic but still manageable?

Simulations with Autonomous Decision-Making Agents

Agent-based simulation (ABS) has emerged over the last twenty years from the intersection of biology, social science, and computer science as a tool for understanding the behavior of complex adaptive systems. The defining characteristic of an agent-based simulation is the autonomous decision-making behavior of the agents. As noted above, micro-simulations and systems dynamic models are governed by an overall structure and process that determines the state of individual agents or subunits at any point in time. In contrast, agents in an agent-based simulation possess sensing mechanisms and decision rules that govern their behavior in response to changes in their environment and the behavior of other agents. Agents in an ABS must have at least these four essential attributes: autonomy, asynchrony, interaction and bounded rationality. Adaptation, or the ability to learn, is another trait that is present in many agent-based simulation models.

Autonomy refers to the characteristic that agents act independently of one another. While the actions of other agents may affect their behavior, agents are not guided by some central control authority or process. Asynchrony stems from autonomy and means that the sequencing and time required for an action by any one agent is independent of the state of any other agents. Bounded rationality means that agents make decisions without complete knowledge, with limited computational resources and time (just like real consumers).

Agent-based simulation addresses some of the key reasons for forecast failure identified above, primarily through the discipline required to develop and evaluate an agent-based simulation. Agent-based simulations also are well suited for situations where the underlying transfer function is based on growth, on evolution, on “if-then” causality or on emergence.

As a way to illustrate the application of an agent-based approach to an important marketing problem, consider the challenge of predicting the future sales of a new consumer durable product prior to launch. Predicting or forecasting sales for new consumer durable goods is a challenge for market researchers. Bayus, Hong and Labe (1989) cite a few specific reasons that such forecasts are difficult, including variability in the timing of purchases and fluctuations in consumer spending (as a function of other factors). Durables are also highly influenced by the dissemination of information about the performance of the products, often by word of mouth, since by their very nature it is difficult to “sample” a product like a washing machine prior to purchase.

Bass (1969) proposed an analytical model for forecasting the adoption rate of a new durable as a function of two parameters reflecting *innovation* (buyers who make a decision independently of the decisions of other buyers) and *imitation* (buyers who are influenced by the choices of others, such as the innovators). The model is simple and elegant but requires a number of assumptions. For example, the Bass model assumes that maximum market potential remains constant over time, that the diffusion of one innovation is independent of all other innovations, that marketing actions do not affect the diffusion process, and that there are no supply constraints. Additionally, the model does not incorporate any consumer heterogeneity (e.g., budget constraints, timing of purchase triggers, susceptibility to word of mouth).

In the Bass model, imitation is solely a function of the number of previous adopters. While we can estimate an aggregate imitation effect retrospectively, if we are going to predict or project imitation, we probably should capture the impact of social networks in some way to account for word of mouth. For example, an early adopter who has a bad experience with a new product and possesses a large social network can have a larger impact on imitation that a consumer with a small social network. This difference can be significant, determining whether negative experiences are widely and rapidly disseminated or relatively contained to a few consumers. Agent-based simulation is well suited for modeling the impact of social networks on the rate of diffusion.

The formal expression for the Bass model in its original form is:

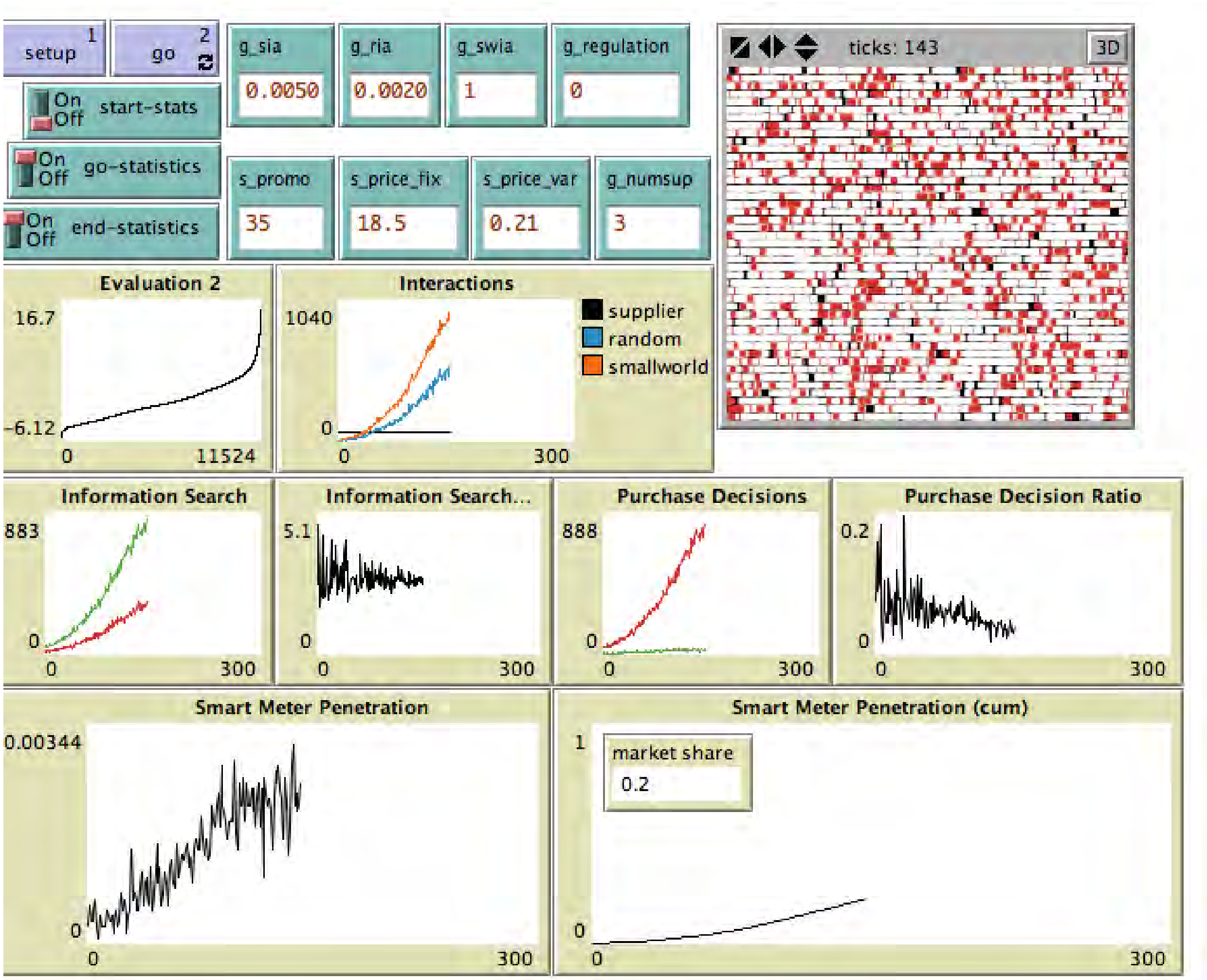
$$\boldsymbol{Q_t = [p + r\,(N_t\,/\,Q')]\,(Q'-N_t)}$$

Where:

$\boldsymbol{Q_t}$ = the number of adopters at time $\boldsymbol{t}$
$\boldsymbol{Q'}$ = the ultimate number of adopters
$\boldsymbol{N_t}$ = cumulative number of adopters to date
$\boldsymbol{r}$ = effect of each adopter on each non-adopter (coefficient of internal influence)
$\boldsymbol{p}$ = individual conversion ratio in the absence of adopters' influence (coefficient of external influence)

Various agent-based realizations of the Bass model exist. These realizations typically either create a heterogeneous population with respect to ***r*** and ***p***, a population that is connected via a network structure, or some combination of heterogeneity and network structure. For example, Rixin (2011) created a simulation (using NetLogo, a popular and free agent-based toolkit from Northwestern University) that incorporates individual preferences and other individual variables as well as marketing variables (e.g., promotion). The ***r*** and ***p*** are implemented by stochastic interactions between consumers (for ***r***) and between sellers and consumers along with the individual's proclivity to adopt (for ***p***). **Figure 9** displays a sample simulator interface from this model.

Figure 9. Bass Diffusion Simulator Interface



g_sia 0.0050

g_ria 0.0020

g_swia 1

g_regulation 0

s_promo 35

s_price_fix 18.5

s_price_var 0.21

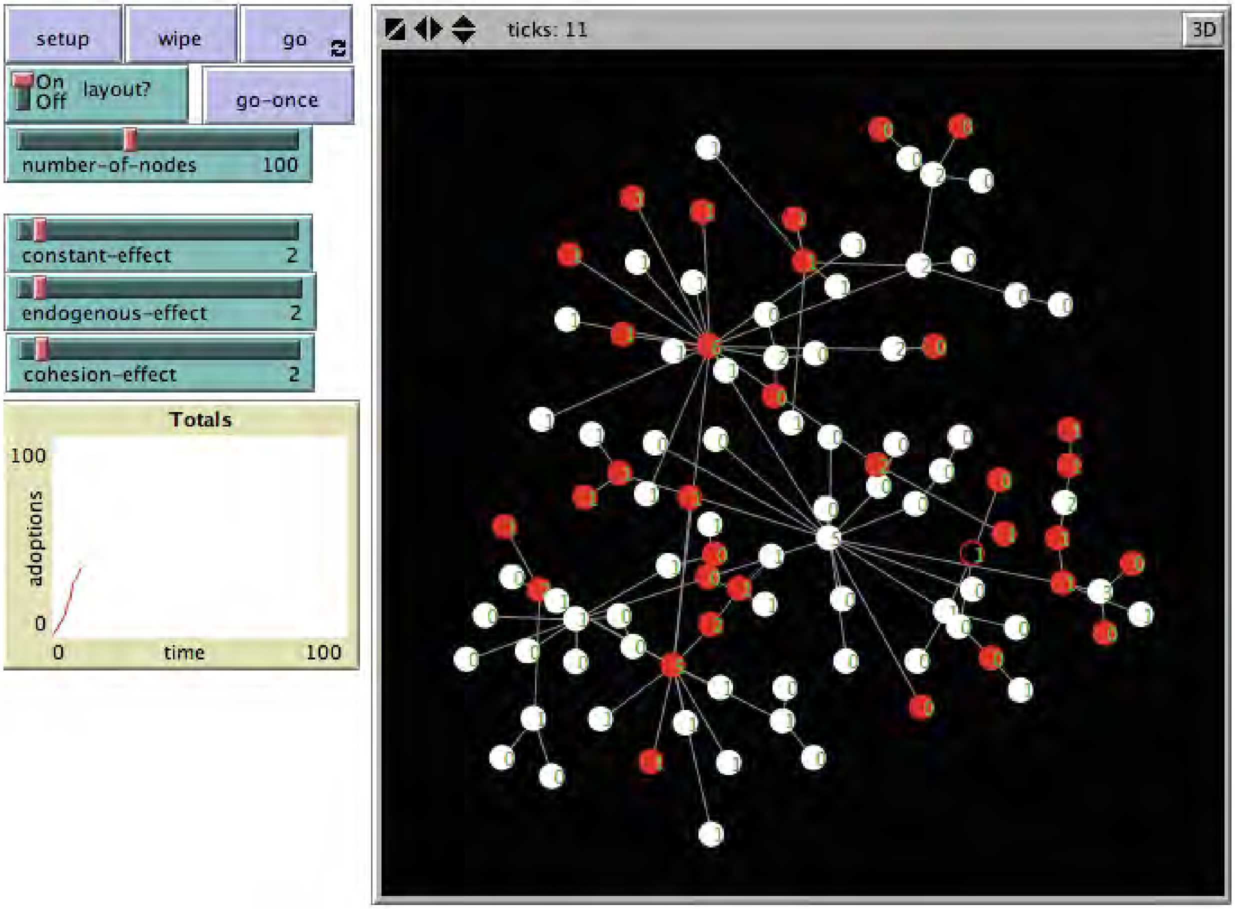
g_numsup 3

ticks: 143

market share 0.2

A second example based on the Bass model (Rossman, 2010) employs population level values for the coefficients of internal and external influence but introduces a preferential attachment network for simulating the diffusion of information due to adoption. A preferential attachment follows a power law, where most nodes have only one or two connections but a few nodes have very many connections, as we can see in the simulator interface for this model in Figure 10.

Figure 10. Preferential Attachment Diffusion Model



ticks: 11

On Off layout?

number-of-nodes 100

constant-effect 2

endogenous-effect 2

cohesion-effect 2

In this example, the shaded circles represent adopters. We can see that adopters are clustered in various network neighborhoods in this snapshot.

Agent-based models of new product diffusion offer potential advantages over other analytic or heuristic approaches. For one thing, in cases where there is little or no historical data for comparable products to use in estimating the internal and external influence coefficients we can use simulation to discover likely values, as in the Rixin Bass model simulation, or we can explore the coefficient space by systematically varying the values, as in the diffusion simulation with the preferential attachment network. Second, these simulations make it fairly easy to determine the sensitivity of a forecast to variations in the key input variables.

Five Guidelines for Making Better Predictions

As Nobel laureate physicist Niels Bohr quipped, “Prediction is very difficult, especially about the future.” We do not know what the future will be; at best we can make guesses about the likelihood of different possible futures. Our aim should be to understand which of those possible futures are more likely and which are less likely. Here are five guidelines that I believe will help us to make better estimates (guesses) for those likelihoods.

1. **Start with the most comprehensive description that you can imagine of the system you want to forecast.**

This description should identify the transfer function that produces the future from the current starting conditions. In most systems the transfer functions will be analogous to at least one of the physical world mechanisms identified previously: Newtonian motion, thermodynamics, evolution, systems dynamics, if-then causality, or emergence. This is the most important step in getting to better predictions and forecasts.

2. **Adopt an "agent-based" mindset when formulating your prediction problem.**

Not every forecasting or prediction problem requires (or fits) an agent-based simulation, but there are certain aspects of the agent-based mindset that can lead to better forecast models of all types:

- Agent-based models employ a bottom-up, disaggregate approach to describing the system of interest.
- Agent-based thinking requires that we consider the processes (events occurring in a sequence over time) that govern the system of interest.
- In agent-based modeling we try to find the simplest model that captures the important behavior of the system of interest (by starting with very simple models and gradually adding detail).

3. **Remember that the futures we are interested in are stochastic, not deterministic.**

Even in the presence of a fairly strong if-then causal transfer function many of the variables that shape the future will result from random processes. The "art" of forecasting often requires that we find the right balance between the deterministic and stochastic components of the system of interest. We must run our simulations many times to generate a distribution of likely futures.

4. **Use computer simulations to help validate our causal models of the future and reveal possible outcomes that we did not anticipate.**

I believe one reason we rely so much on predictive models based on historical data is that we have some way (in theory, at least) to validate the model (e.g., by predicting holdout cases) relative to existing data. If we take some other approach, validation becomes a challenge because we pretty much have to wait for the future to happen in order to confirm that our model is valid.

With a simulation model we can adapt a different approach that is less dependent on historical data. When we start by describing and implementing a causal model via simulation, we can test that model by using existing data to set input values and compare simulation results to the current (or past) real world.

A good example of this process can be found in Epstein and Axtell (1996) who describe the process of "growing" an artificial society using agent-based simulation. They clearly demonstrate how they develop a model of a simple process and design agents and the environments they inhabit. They compare the results of their simulations with historical information about human societies to validate their hypotheses.

5. Get comfortable with subjective prior beliefs.

Now that market researchers are being exposed to Bayesian thinking we are beginning to understand the role that our so-called "subjective" beliefs about the state of nature can play in making predictions about uncertain events. When it comes to forecasting, "subjective" prior beliefs (which may be partly rooted in evidence) have at least two uses. First, for forecast inputs that are "unknown unknowns," our best guess for a value may be our only option. In some cases this guess may be uninformed, as when we start with the expectation that each of two possible outcomes is equally likely. In other cases we may be able to make a more informed guess.

Second, our prior beliefs should be the basis for reasonableness checks on our forecasting model. If our forecasting model consistently produces results that just seem, based on our subjective beliefs or our intuitions, "too good to be true," it may well be that they are too good to be true and we need to re-examine our forecasting model.

Forecasting and prediction is at least as much art as it is science, and forecasters who keep that it mind will, on average, have better luck than those who do not.

David Bakken

References

Bass, F. M. (1969). "A new product growth model for consumer durables," *Management Science*, Vol. 15, No. 5, Theory series, pp. 215–227.

Bayus, B. L., Hong, S., and Labe, Jr., R. P. (1989). "Developing and using forecasting models of consumer durables: the case of color television," *Journal of Product Innovation Management*, 6, pp. 5–19.

Epstein, J., and Axtell, R. (1996). *Growing artificial societies: social sciences from the bottom up.* Washington, D.C., Brookings Institution.

Miller, J. H. and Page, S. E. (2007). Complex adaptive systems: an introduction to computational models of social life. Princeton: Princeton University Press.

Ormerod, P. (2005). Why most things fail: evolution, extinction, economics, New York: John Wiley and Sons.

Rixin, Martin (2011, August 18). "A consumer-demand simulation for Smart Metering tariffs (Innovation Diffusion)" (Version 1. *CoMSES Computational Model Library*. Retrieved from: https://www.openabm.org/model/2592/version/1

Rossman, G. (2010). Diffusion simulation, NetLogo User Community Models. Retrieved from ccl.northwestern.edu/netlogo/models/community/diffusion.

Taleb, N. N. (2007). The black swan: the impact of the highly improbable. New York: Random House.

Wilke, A. and Barrett, H. C. (2009). “The hot hand phenomenon as a cognitive adaptation to clumped resources,” *Evolution and Human Behavior*, 30, pp. 161–169.

WALLET ECONOMICS?: CREDIT CARD CHOICE-BASED CONJOINT— BEYOND PREFERENCE AND APPLICATION

DEMITRY ESTRIN
MICHELLE WALKEY
VISION CRITICAL
VIDYA SUBRAMANI
CLIENT BANK
CARLA WILSON
VISA
JANE TANG
ROSANNA MAU
VISION CRITICAL

ABSTRACT

Credit cards have become an important part of our financial world. Choice Based Conjoint (CBC) is a natural fit to analyze this category as it accurately reflects how consumers make tradeoff decisions among various credit card offers. However, the standard CBC outputs (preference share and simulated likelihood that the card will be applied for) are of limited use for card issuers. To accurately assess the potential profitability of a credit card offer, issuers need to know not only whether the card offer will be accepted into the wallet, but also how the credit card will be used once there. The amount of revenue earned on a card is directly tied to the amount spent on the card. Further the costs associated with rewards and features are often tied to specific spend categories (i.e., gas, groceries, restaurants, etc.). The standard CBC output doesn't shed light on either of these areas.

We propose a framework that extends the usual CBC deliverable to include card usage projections, including spend by category. This extension allows us to more accurately estimate the profit potential of one credit card configuration versus another. We model spend on the preferred credit card using volumetric estimation proposed in Eagle (2010). The amount spent is assumed to be proportional to the appeal of the credit card offer. We also employ a series of general questions on a consumer's intention towards spending in each category given features of the new credit card. This approach allows us to take the respondent's total spending estimate, from the volumetric estimation, and allocate it into the different merchant categories given the features of this credit card.

The analysis of the credit card offer is built up from these individual results, aggregating them to obtain revenue and the cost of rewards for the entire sample and population projection. This bottom-up approach is very flexible and is able to accommodate different types of credit card offerings—cash back, points based rewards, loyalty rewards, etc. It is also reliable, and has been validated in multiple checkpoints across multiple projects.

Most of the research conducted in this space revolves around enhancing or evolving an existing card portfolio toward higher acquisition and usage. As we project the profitability of

each possible product structure, we are careful to validate our results with what we know to be true of the current portfolio. As we design our product matrix, we ensure that all of the features and levels associated with the current card are included. This allows us to estimate the acquisition rate, revenue, cost and profitability of the existing card offer. Over the course of several projects, the team has been able to validate these estimates to actual in-market performance of the current card. This validation confirms that our modelled projections are grounded in a realistic base that reflects consumer preferences and usage patterns.

As is often the case, what consumers want (great features with richest rewards) is at odds with what the issuers can deliver. The ability to accurately estimate the revenue and cost potential of credit card offerings allows the card issuer to optimize their product. With some work, they can find the optimal intersection between consumer preferences and profitability.

1. Introduction

The payments industry is complex and constantly evolving. Credit cards represent the largest segment of the industry and in North America the credit card market is mature, consolidated and highly competitive. To provide context, here are some statistics on the US credit card space:

$470 billion	-	purchase volume in Q3‘14 from top 7 issuers
1.4 billion	-	number of credit cards in circulation
72%	-	of US consumers have a credit card
3–4	-	cards in wallet among those who use credit cards
$719	-	average monthly spend dedicated to primary card
$260	-	average monthly spend dedicated to secondary card in wallet

The persistent challenge for card marketers, in today’s environment, is to design and deliver the right product, to the right customer at the right time.

Acquisition efforts are in part hampered by the simple fact that consumers are oversaturated. The relatively low cost of digital marketing generates a deluge of touch points, leaving consumers feeling bombarded by unsolicited content. And when we turn our attention to direct mail, the bedrock for acquisition marketing, the statistics from U.S. Postal Service are staggering. An average household receives approximately 13 direct mail pieces a week. And those households that are fortunate enough to earn more than a $100,000 a year receive 21 pieces a week. That's 84 offers a month, on average, for affluent households. With credit card offers taking up the lion’s share of our weekly clutter, offers from banks account for approximately 30% of all advertising mail.

On the surface, the challenge for banks is to create a compelling and relevant offer that can break through all of this clutter. However, for the card manager responsible for the overall health of the portfolio, the business objective is rarely that simple. After all, the goal is to present the *right offer* to the *right customer* and this translates into three complementary objectives of *acquisition*, *usage* and *profitability.* A successful offer has to deliver on all three of these objectives to ensure a healthy and viable portfolio for the bank.

Coming out of the financial downturn, the payments industry has put more emphasis on innovation and product design than ever before. Regulatory changes and the reality of massive losses incurred during the credit crisis have forced the industry to abandon the lure of a term-driven offer in favor of enhanced value propositions with tailored rewards and ancillary benefits that uniquely meet the needs of target consumer segments. In fact, in an otherwise commoditized product space, rewards have become a crucial point of differentiation for issuers and a coveted source of value for consumers. Taking terms out of the equation, recent studies have shown that rewards can drive as much as 20% to 30% of the decision to sign-up for a new card and 40% to 50% of the monthly spend allocated to the new card. With that much impact on usage, depending on the value of the reward offered, it is very possible to create an offer that is very attractive for consumers but at the same time highly unprofitable for the bank.

Choice Based Conjoint (CBC) is a natural fit to analyze this category as it accurately reflects how consumers make tradeoff decisions among various credit card offers. However, the standard CBC outputs (preference share and simulated likelihood that the card will be applied for) provide only a partial view of the information banks need. To accurately measure the potential success of a credit card, revenue and costs need to be considered. Hence, banks need to know, not only whether the card offer will be accepted into the wallet, but also how the credit card will be used once there. The amount of revenue earned on a card is directly tied to the amount spent on the card. Further the costs associated with rewards and features are often tied to specific spend categories (i.e., gas, groceries, restaurants, etc.). The standard CBC output doesn't shed light on either of these areas.

We propose an evolved framework that extends the usual CBC deliverable to include card usage projections, including spend by category. This extension allows us to more accurately estimate the profit potential of one credit card configuration versus another. This evolved application of CBC allows the card marketer to define a "win-win" strategy that balances consumer preferences with product profitability.

2. Our Approach

We use the following simplified case study to illustrate our approach.

2.1 Study Design & Questionnaire Elements

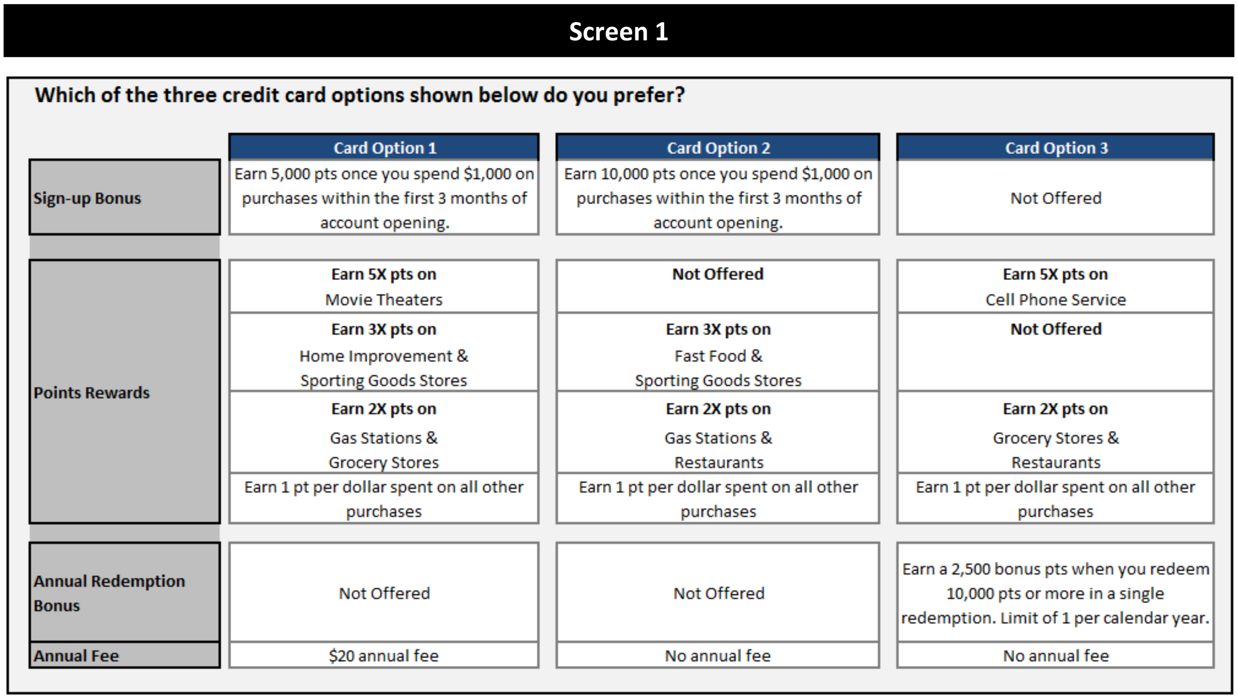
Our client at Bank ABC is interested in refining the value proposition in their XYZ credit card offer. In particular, current cardholders, while they like the existing card features and benefits, have low awareness of different product features included in the card. The prospects have little or no awareness of the product and do not find the current rewards program attractive. The goal of this research initiative is to guide the refinement of XYZ Card to make it more attractive to prospects and more engaging to existing cardholders. More specifically, the objectives of this research are as follows:

- Understand features and benefits (both existing and incremental) of the XYZ card that resonate with cardholders and card prospects.
- Evaluate the impact of all tested features on the product's profitability (i.e., what do customers want vs. what is affordable).

The inputs for the conjoint model are as follows. The underlined level in each factor represents the feature in the existing card offer.

Optimization Features		**Levels**			
Earn points for every $1 spent	**5 points for….**	Not Offered	Movie Theaters	Bookstores	Cell Phone
	3 points for…	Not Offered	Home Improvement + Fast Food	Home Improvement + Sporting Goods Stores	Fast Food + Sporting Goods Stores
	2 points for…	Not Offered	Gas Stations + Grocery Stores	Gas Stations + Restaurants	Grocery Stores + Restaurants
	1 point for…	all other purchases			
Sign-up Bonus (once you spend $1,000 within the first 3 months of account opening)		Not Offered	5,000 points	10,000 points	
Annual Redemption Bonus (when you redeem 10,000 points or more in a single redemption.)		Not Offered	2,500 points		
Annual Fee		No Annual Fee	$20 Annual Fee		

We set up the standard CBC design using SAS PROC FACTEX/OPTEX, showing 3 card options per screen and asking respondents to choose their preferred option.

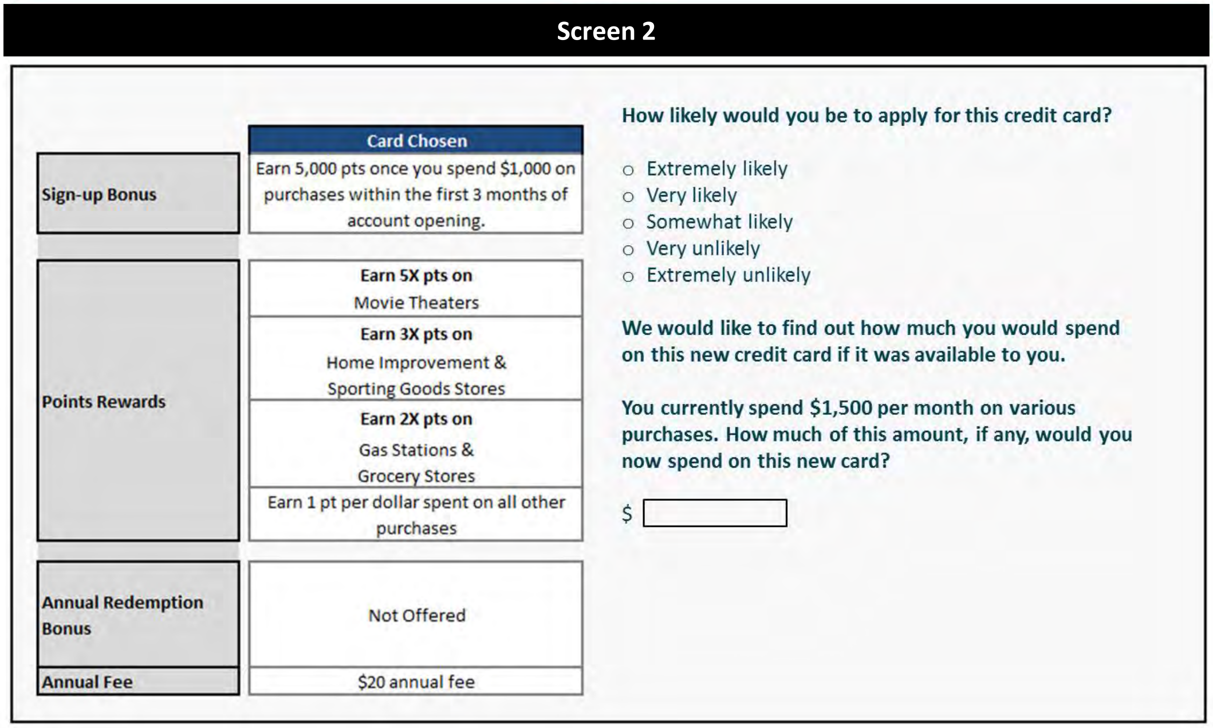


Screen 1

Which of the three credit card options shown below do you prefer?

	Card Option 1	Card Option 2	Card Option 3
Sign-up Bonus	Earn 5,000 pts once you spend $1,000 on purchases within the first 3 months of account opening.	Earn 10,000 pts once you spend $1,000 on purchases within the first 3 months of account opening.	Not Offered
Points Rewards	**Earn 5X pts on** Movie Theaters	**Not Offered**	**Earn 5X pts on** Cell Phone Service
	Earn 3X pts on Home Improvement & Sporting Goods Stores	**Earn 3X pts on** Fast Food & Sporting Goods Stores	**Not Offered**
	Earn 2X pts on Gas Stations & Grocery Stores	**Earn 2X pts on** Gas Stations & Restaurants	**Earn 2X pts on** Grocery Stores & Restaurants
	Earn 1 pt per dollar spent on all other purchases	Earn 1 pt per dollar spent on all other purchases	Earn 1 pt per dollar spent on all other purchases
Annual Redemption Bonus	Not Offered	Not Offered	Earn a 2,500 bonus pts when you redeem 10,000 pts or more in a single redemption. Limit of 1 per calendar year.
Annual Fee	$20 annual fee	No annual fee	No annual fee

A dual-response setup is used in collecting data for application intent (on his preferred card) and intended spend on the card on a second screen.



Screen 2

	Card Chosen
Sign-up Bonus	Earn 5,000 pts once you spend $1,000 on purchases within the first 3 months of account opening.
Points Rewards	**Earn 5X pts on** Movie Theaters **Earn 3X pts on** Home Improvement & Sporting Goods Stores **Earn 2X pts on** Gas Stations & Grocery Stores Earn 1 pt per dollar spent on all other purchases
Annual Redemption Bonus	Not Offered
Annual Fee	$20 annual fee

How likely would you be to apply for this credit card?

- ○ Extremely likely
- ○ Very likely
- ○ Somewhat likely
- ○ Very unlikely
- ○ Extremely unlikely

We would like to find out how much you would spend on this new credit card if it was available to you.

You currently spend $1,500 per month on various purchases. How much of this amount, if any, would you now spend on this new card?

$ □

We repeat both screens of questions (preference, followed by intent to apply and spend) 8 times for each respondent.

Before the CBC exercise, we also ask respondents about their general spending habits:

Q1 Current Spend on credit cards
Q2 Current Spend on debit, check card tied to a checking account, cash, checks, etc.

Additionally, we ask the current spend in the various merchant categories that we are interested in testing.

Q3

Movie Theaters
Cell phone
Bookstores

Home Improvement
Fast Food
Sporting Goods

Gas Stations
Grocery Stores
Restaurants

Following the CBC exercise, we ask respondents to estimate the amount they would charge to the new card for these various merchant categories if the reward levels were at 5Xs, 3Xs, and 2Xs.

Q100.

5Xs	3Xs	2Xs
Movie Theaters	Home Improvement	Gas Stations
Cell phone	Fast Food	Grocery Stores
Bookstores	Sporting Goods	Restaurants

2.2 CBC Modeling & Volume Estimation

We use the standard Hierarchical Bayes Multinomial logit model to model the preference data as the first step. Only the dual-response portion of the choice data is used in this stage of modeling. The application intent data is dichotomized (extremely/very likely = buy) in the modeling. Sawtooth Software's CBC/HB product is used in the modeling estimation.

We then use the predicted application intent on a respondent's preferred card as the input for the second step volume model. The simulated likelihood of applying for the respondent's preferred card in each choice is based on using only two options in the simulation, the preferred card and the "none" alternative.

This approach was described in Eagle (2010). Specifically, the dependent variable is the proportion of intended spend in each choice task out of the current spend. In other words, spend projection is capped at the current spend, i.e., a respondent can never spend more than what he or she currently spends. For each respondent i, volume (i.e., card spend as proportion of total spend) is assumed to be directly proportional to the appeal of the card offer, expressed by the simulated likelihood of application for the card. Sawtooth Software's HB-REG product is used in the estimation.

$$Volume_i = \beta_i \widehat{Pr_i}$$

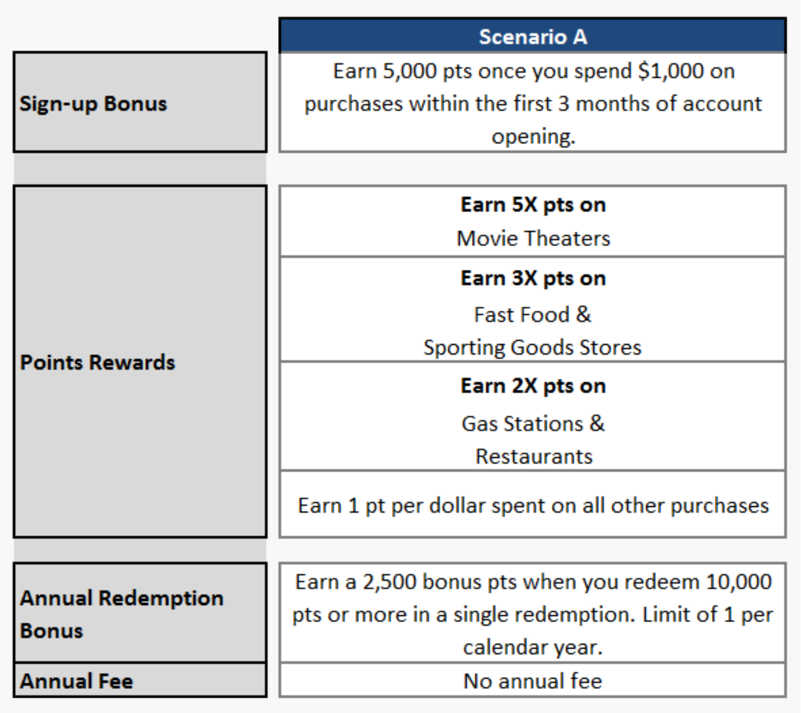
The R-square values in the model are generally good for most respondents. The exception to this are respondents who show no variation in their spend data. As it is not possible to estimate a slope for these respondents, they are generally excluded from this part of the estimation. During simulation, an arbitrary cutoff point is used to determine spend on the card offer. If a respondent has a 50% or higher likelihood of applying for the card, the full spend amount (constant through all the choice tasks) is used as the spend estimate. Otherwise, a value of 0 is used (as the respondent is unlikely to apply for the card and hence cannot actually use the card).

It may also be possible to expand the estimation of the model here by including additional terms related to specific features of the card that impact volume (i.e., spend) independently of the card preference. For example, a consumer may equally prefer a card with accelerated rewards for travel spend and a card with accelerated rewards for grocery spend. As travel inherently costs more than groceries (unless you are feeding multiple teenage boys), the accelerated travel reward card should result in more spend. Additional terms in the volume model may be useful in this case.

In the simulation, only two values are then carried forward into the revenue/cost calculation for each respondent: the application likelihood value (i.e., probability of applying for the card), and the predicted total spend on the card (i.e., the sum of predicted credit card spend and non-credit card spend).

2.3 Revenue and Cost Calculation

Consider the following card offer in Scenario A:



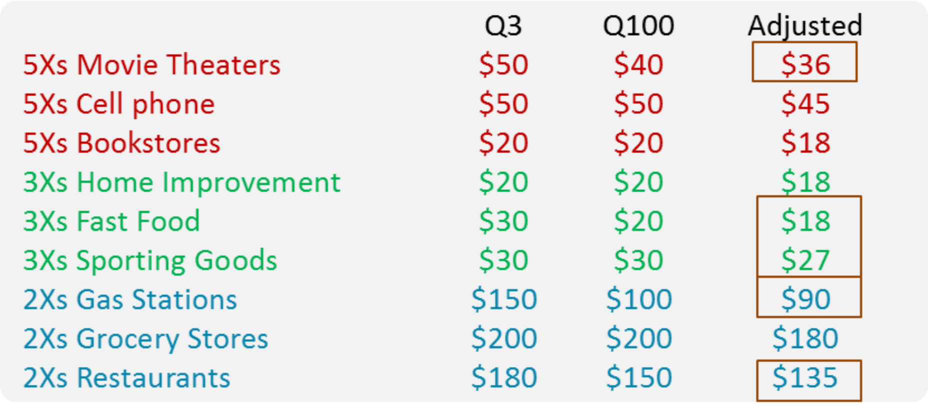
	Scenario A
Sign-up Bonus	Earn 5,000 pts once you spend $1,000 on purchases within the first 3 months of account opening.
Points Rewards	**Earn 5X pts on** Movie Theaters
	Earn 3X pts on Fast Food & Sporting Goods Stores
	Earn 2X pts on Gas Stations & Restaurants
	Earn 1 pt per dollar spent on all other purchases
Annual Redemption Bonus	Earn a 2,500 bonus pts when you redeem 10,000 pts or more in a single redemption. Limit of 1 per calendar year.
Annual Fee	No annual fee

A respondent named Amanda tells us that she typically spends a total of $1,200 per month. Based on the model, we estimate that Amanda has a 90% probability of applying for this card offer, and will spend a total of $720 per month on this card.

The bank derives revenue from Amanda's usage of this card from both the annual fee she pays on the card and the interchange revenue (i.e., a fixed % of every dollar spent on each card in the portfolio). In this case, the card offer has no annual fee, the revenue from that is 0. With the interchange rate of 2%, the annual interchange revenue is approximately $173.

On the other hand, year 1 cost related to Amanda's card is the sum of sign-up bonus, cost of rewards, and redemption bonus. Based on the $720 monthly spend, Amanda would be entitled to the sign-up bonus of 5,000 points. Adjusted by the 90% likelihood of application, we expect Amanda will earn 4,500 points as her sign-up bonus.

If we adjust the category spend information Amanda gives us by her likelihood of applying for the card, we arrive at the adjusted category spend information.



	Q3	Q100	Adjusted
5Xs Movie Theaters	$50	$40	$36
5Xs Cell phone	$50	$50	$45
5Xs Bookstores	$20	$20	$18
3Xs Home Improvement	$20	$20	$18
3Xs Fast Food	$30	$20	$18
3Xs Sporting Goods	$30	$30	$27
2Xs Gas Stations	$150	$100	$90
2Xs Grocery Stores	$200	$200	$180
2Xs Restaurants	$180	$150	$135

A bit of math shows that Amanda is expected to earn a total of **14,148 points** in a year:

5Xs - $36 (Movies) * 5 = 180 points

3Xs - ($18 + $27 (Fast Food+Sporting Goods)) * 3 = 135 points

2Xs - ($90 + $135 (Gas+Restaurants)) * 2 = 450 points

1X: $720 (Predicted Spend) - $306 (Movies+Fast Food+SportingGoods+Gas+ Restaurants) = $414 = 414 points

Total reward earned = (180 + 135 + 450 + 414) points * 12 months =14,148 points

In Amanda's case, her spend in each of the accelerated rewards categories (i.e., Movie Theaters, Fast Food, Sporting Goods, Gas and Restaurants) sums up to only $306 out of the total $720 predicted spend, which leaves $414 as her spend on all other categories. In cases where the spend in these accelerated rewards categories exceeds the total predicted spend, the total predicted spend estimate is retained and distributed proportionally to each of the accelerated reward categories, leaving 0 spend in the non-accelerated categories.

Amanda's accumulated annual rewards (14,148 points) exceeds the threshold (10,000 points). We expect Amanda will also receive 2,250 points in redemption bonus. The total amount of reward she will earn (in year 1) is expected to be 20,898 points (4,500 points + 14,148 points + 2,250 points). At a standard rate of 1 cent per point, this translates to approximately $209.

As the cost of Amanda's reward program exceeds the revenue the bank can derive from Amanda, when Amanda applies for a card outlined in Scenario A, we expect a year 1 loss of $36.

Similarly, we estimate that the bank can expect a profit of $24 from Amanda for the existing card offer. See Appendix A for a detailed calculation.

It is important to note that within the confines of this research we focus on the cost of rewards only to derive our estimate of profitability. After the research is conducted the business adds their own projections and modelling to incorporate other elements of revenue (e.g., revenue from fees, interest, etc.) and cost (e,g., marketing, operational, etc.).

2.4 Adjustment & Aggregation

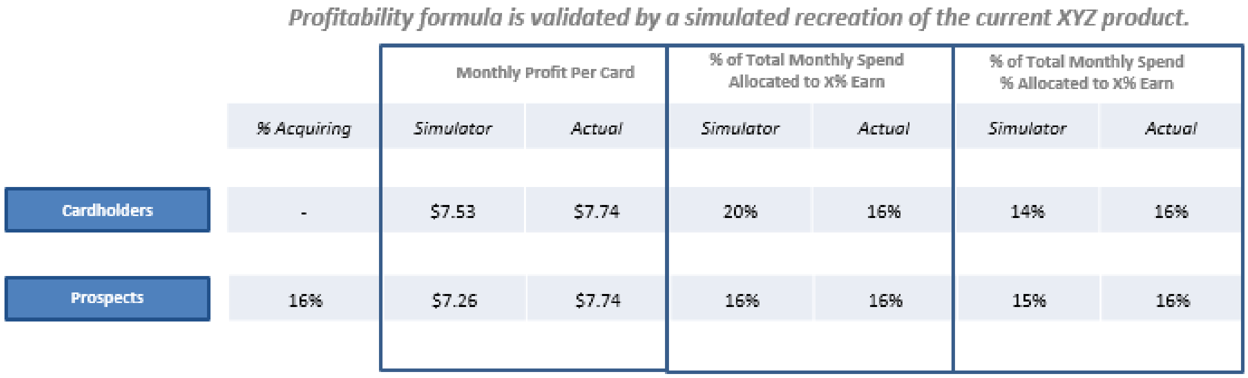
When we add up Amanda's numbers with those of everyone else in the sample, and project the results out to the population, we get a profit estimate for the market as a whole. To improve accuracy, the underlying revenue and cost estimates are adjusted for portfolio attrition (i.e., the proportion of cardholders cancelling the card over the course of the year) as well as reward breakage (i.e., proportion of earned rewards that go unredeemed). These statistics along with basic inputs such as interchange revenue are provided by the bank's product management group at the outset of the project.

2.5 Validation

Most of the time, this research is focused on refining an existing portfolio. As a result, we are able to include the features and levels of the current card in the design matrix.

This allows us to recreate the current card and compare how close our modeling estimates are to what we know to be true of the existing portfolio.

Over the course of several projects, we have yielded results that fall within +/- 10% of the actual portfolio. The exhibit below depicts a validation output from one of our recent studies. As shown, our monthly profitability estimate was within $.20–$.50 of the actual profit witnessed in the portfolio.



Profitability formula is validated by a simulated recreation of the current XYZ product.

		Monthly Profit Per Card		% of Total Monthly Spend Allocated to X% Earn		% of Total Monthly Spend % Allocated to X% Earn	
	% Acquiring	*Simulator*	*Actual*	*Simulator*	*Actual*	*Simulator*	*Actual*
Cardholders	-	$7.53	$7.74	20%	16%	14%	16%
Prospects	16%	$7.26	$7.74	16%	16%	15%	16%

2.6 Challenges

Every design and project has unique challenges. With this approach, we have seen overstatement for anticipated merchant category spend with accelerated rewards. We believe that the overstatement is partially due to the fact that the merchant spend exercise lives outside of the choice task. Theoretically, the data would be more accurate if we asked respondents to re-allocate merchant category spend for each preferred card. However, the exercise would become too tedious and respondent fatigue would become another factor to adversely impact accuracy.

3. Conclusion

As is often the case, what consumers want (great features with richest rewards) is at odds with what a bank can deliver. Every consumer would love 6% cash back on groceries, gas and dining on a credit card with no annual fee and a sizeable intro bonus. While a card like this would be popular, it would also be highly unprofitable for the business.

The dilemma with typical conjoint applications is that they stop at preference and acquisition intent. As a result, the output from a standard CBC is limited to the top 5 or 10 product designs that consumers like most, but that banks may not be able to afford. The implication with a standard CBC is that the product manager needs to limit the initial choice design to include only features and levels (and all possible combinations) that are viable and profitable for the business. If they don't, the recommended product is destined for ignominy. While this seems practical, the reality is that with a standard CBC we are missing out on one of the inherent benefits of conducting this type of research and that is the opportunity to see how far we can push the boundaries of the products and services that we are designing. Our proposed methodology allows us to create choice tasks that include features and levels that push and often transcend the limits of what a product manager would consider to be a viable product lever. Because we look at usage and then convert that data into a proxy of profitability, we no longer have to constrain our design at the outset of the project and instead can rely on our analysis to find the intersection of consumer and bank preferences. This middle ground creates a product that is both highly rewarding for the customer and viable for the business. And this is the essence of why this approach is so important and relevant for the business.

While at the outset a feature may seem too expensive, our analysis can identify the importance of this feature in driving choice as well as usage and then identify if a trade-off exists where the expensive, and highly attractive feature can actually be offered if the bank pulls back on some of the other elements that are bundled into the product.

Ultimately, this approach allows the bank to design a profitable and a highly competitive offer. Considering the mature, consolidated and highly competitive nature of the credit card market, this evolved methodology is an important tool that allows banks and co-brands to create and protect market share.

We believe that our approach can be successfully leveraged in any area where the aim is to go beyond product or service adoption. In fact, any setting where the degree and type of product usage may result in variable business outcomes should be a good fit for this application. We don’t have to stop with credit cards or banking for that matter. Online loyalty programs, online media stores (music, movies, etc.), app design (in-app purchases), retail loyalty programs, online gaming and gambling, and SaaS offerings are just some examples of areas where this evolved CBC design may yield incremental and profitable business insights.

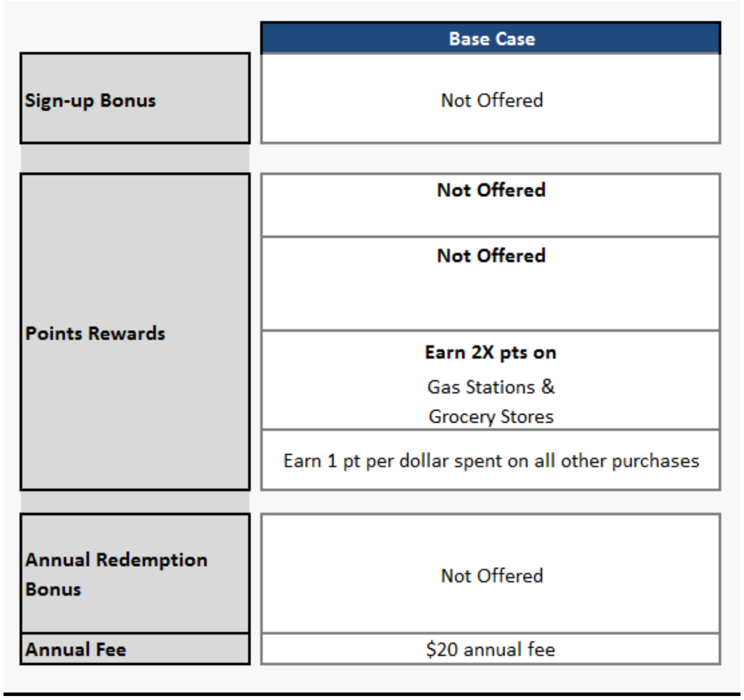
Demitry Estrin

Carla Wilson

Jane Tang

Rosanna Mau

Appendix A: Existing Card Offer/Base Case—Calculation for Amanda



	Base Case
Sign-up Bonus	Not Offered
Points Rewards	**Not Offered** **Not Offered** **Earn 2X pts on** Gas Stations & Grocery Stores Earn 1 pt per dollar spent on all other purchases
Annual Redemption Bonus	Not Offered
Annual Fee	$20 annual fee

From the CBC and volumetric model, we estimate that Amanda has a 30% probability of applying for the card, and a monthly spend of $240.

Revenue: Annual Fee = $20 * 30% = $6 Interchange Revenue = $240 * 2% * 12 = $58 Annual Revenue = $6 + $58 = $64	No sign-up bonus. No redemption bonus. Cost of Rewards: 2Xs: ($30 + $60 (Gas+Grocery))* 2 times = 180 pts 1X: $240 (Predicted Spend) - $90 (Gas+Grocery) = 150 pts Cost of rewards (annual) = (180 + 150) pts * 12 months = 3,960 pts Annual Cost (year 1) = 0 pts + 3,960 pts + 0 pts = 3,960 pts Converting Pts to Cash (1 pt = $0.01) = 3,960 pts * $0.01 = $40
Profit = $64 - $40 = +$24	

References:

Eagle, T. (2010), "Modeling Demand Using Simple Methods: Joint Discrete/Continuous Modeling" Sawtooth Software Conference Proceedings.

Estrin, D. (2013) "Making Card Marketing All About the Consumer" PaymentsSource.

Flamme, M. Grieve, K. (2014), "7 Trends Impacting Retail Payments," ABA Banking Journal.

Team, T. (2014) "A Look At The Country's Largest Card Lenders: Credit Card Payment Volumes" Forbes.

CONJOINT FOR FINANCIAL PRODUCTS: THE EXAMPLE OF ANNUITIES

SUZANNE B. SHU
ROBERT ZEITHAMMER
UCLA
JOHN PAYNE
DUKE UNIVERSITY

ABSTRACT

We propose and estimate a model of individual preferences for life annuity attributes using a choice-based stated-preference survey. Annuities are presented in terms of consumer-relevant attributes such as monthly income, yearly adjustments, period certain guarantees, and company financial strength. We find that attributes directly influence preferences beyond their impact on the annuity's expected present value. The strength of the direct influence depends on how annuities are described: when represented only via basic attributes, consumers undervalue inflation protection and preferences are not monotonically increasing in duration of period certain guarantees. When descriptions are enriched with cumulative payment information, consumers no longer undervalue inflation protection, but nonlinear preferences for period certain options remain.

INTRODUCTION: A SUMMARY OF A FULL-LENGTH PAPER

This paper is a summary of our full-length paper titled "Consumer Preferences for Annuity Attributes: Beyond NPV" published elsewhere and available on our websites. The full-length paper sets up the context (life annuities and decumulation of retirement savings), and provides detailed results, as well as managerially important simulations. In this summary, we focus on the novel way we specify the utility model for analyzing data from a choice-based conjoint (CBC) analysis survey. Specifically, we adapt CBC to the domain of financial products. The domain of financial products is distinct from other applications of CBC because the product attributes jointly imply an expected present financial value of the product, and normative economic models suggest that the financial value should be the main driver of choice. Knowing the financial value of each product in our conjoint space allows us to see whether an attribute influences demand only through its contribution to the financial value of the product or whether it also has psychological worth "beyond the financial impact."

We find that a typical consumer choosing from a set of annuities does not merely maximize the expected financial value, but also reacts to several product attributes directly—expressing preferences beyond the effect of attributes on the financial value. For example, most consumers overvalue medium (10–20 years) levels of period-certain guarantees relative to their financial impact, but generally undervalue inflation protection via annual increases in payments.

Our second goal is to understand how annuity attribute valuations are affected by changes in information presentation. Varying information presentation has long been part of the toolkit available to marketers, and is increasingly seen as a tool available to policy makers in their efforts to "nudge" consumers toward purchases that may increase consumer welfare (Thaler and

Sunstein, 2008). We predict that the strength of the influence of attributes on consumer preferences beyond their impact on NPV will depend on how the annuity products are described. In one of the presentation conditions of our study, we describe each annuity product in terms of its basic attributes as per current industry norms. In another presentation condition, we enrich the product description with non-discounted cumulative payment information for a few representative "live-to" ages. Note that this "enriched information" condition does not provide consumers with additional information—it merely helps them get a sense of possible payoffs given exactly the same underlying attributes. Not surprisingly, we find that consumers in the enriched information condition undervalue inflation protection attributes less than consumers in the basic information condition. In contrast to this partial de-biasing effect of the enriched information, preferences for period-certain guarantees continue to exhibit very similar under- and over-valuation as in the basic information condition. We also find that enrichment of information increases the baseline preference of annuitization over self-management.

In each information condition, we also find significant individual differences in preferences for annuity attributes correlated with consumer characteristics such as amount saved for retirement, subjective life expectancy, numeracy, and perceived fairness of annuities. Most of these characteristics are correlated with preferences in a qualitatively similar manner regardless of the product description condition, with the exception of subjective life expectancy which is positively correlated with a preference for annual increases only in the enriched information condition.

Our findings provide several insights regarding consumer annuity choice and ways that marketers can improve consumers' acceptance of annuitization without paying out more money in expectation. For example, a marketer can increase demand for an annuity of a fixed expected present value by reducing the amount of an annual increase and using the resulting savings to fund an increase in the duration of the period-certain guarantee up to 20 years. Which products the issuer should offer depends on the way they will be described (shorter period-certain guarantees are optimal under enriched information than under basic information). Regardless of the information presentation, we find that such "repackaging" of the payout stream can have a large effect on demand, sometimes even doubling the take-up rate of annuities in the population we study. Before presenting the detailed methods and results of the conjoint analysis of annuity product features we next turn to brief review of the role of annuities in the retirement journey.

A Conjoint Study of Consumer Preferences for Annuity Attributes

Our discrete choice experiment consists of 20 choice tasks. In every choice task, we asked participants, "If you were 65 and considering putting $100,000 of your retirement savings into an annuity, which of the following would you choose?" They then saw three annuity options and a fourth no-choice option that read, "None: If these were my only options, I would defer my choice and continue to self-manage my retirement assets."

Attribute Selection

The attributes we use include starting income, insurance company financial strength ratings, amount and type of annual income increases, and period-certain guarantees. Each attribute can take on several levels selected to span the levels commonly observed in the market today (see

Table 1). To understand why we selected these attributes and how we selected their levels, please see our full-length paper.

Individual Differences

The multiple responses per individual allow us to estimate each individual's indirect utility of an annuity contract as a function of the contract's attributes, both directly and via their contribution to the expected payout (calculated using the Social Security Administration's gender-specific life expectancy tables). To try and explain some of the population heterogeneity we observe, we collect several key demographic and psychographic measures from each participant: age, gender, subjective life expectancy, numeracy, perceived fairness of annuities and loss aversion. To understand why we selected these demographics and psychographics, please see our full-length paper.

Table 1: Attribute Levels Used in the Conjoint Analysis

Level	Starting monthly income	Company financial strength rating	Annual increases in payments	Period-certain guarantee
1	Monthly payments start at $300 ($3,600/year)	Company rated AA (very strong)	Fixed payments (no annual increase)	No period-certain option
2	Monthly payments start at $400 ($4,800/year)	Company rated AAA (extremely strong)	3% annual increase in payments	5-year period-certain
3	Monthly payments start at $500 ($6,000/year)		5% annual increase in payments	10-year period-certain
4	Monthly payments start at $600 ($7,200/year)		7% annual increase in payments	20-year period-certain
5			$200 annual increase in payments	30-year period-certain
6			$400 annual increase in payments	
7			$500 annual increase in payments	

Information Presentation Treatment

To test how presentation of information about annuity choices affects attribute valuation, our study tests two versions of the annuity choice task, between subjects. In the basic condition, each annuity is described based only on its primary attributes of starting monthly (and annual) payments, annual increases, period certain options, and company rating. This presentation is modeled on typical presentations of annuity attributes by issuers in the market today. Our second "information enriched" condition provides the same information but also includes a table of cumulative payout per annuity conditional on living until the ages of 70, 75, 80, 85, 90, and 95. These cumulative tables do not provide any additional information beyond what the participant could calculate directly using the provided attributes. However, we predict that by "doing the math," participants will be able to more clearly see the joint cumulative impact of all attributes on expected payouts and hence align their choices with it better. Sample presentations for each condition are shown in Figures 1a and 1b.

Model Specification

Each of the 20 choice sets in our study consists of K=3 alternatives (annuities), with the k-th alternative in the n-th choice set characterized by a combination of the attributes presented in Table 1. Our baseline utility specification is based on the variables that should theoretically drive annuity choice, namely, the expected payout and the financial strength rating of the issuer. We denote the expected payout of the annuity V, and calculate it from the monthly income, period certain, and the annual increase (if any) of the k-th annuity in the n-th choice set as follows:

$$V_{n,k} = \underbrace{\sum_{age=65}^{65+pc_{n,k}} \delta^{(age-65)}\left(12\times income_{n,k,age}\right)}_{\text{guaranteed income during the period certain } pc_{n,k}} + \underbrace{\sum_{age=66+pc_{n,k}}^{120} \delta^{(age-65)}\Pr\left(\text{alive at } age\right)\left(12\times income_{n,k,age}\right)}_{\text{uncertain income conditional on living until a given age}} \qquad (1)$$

(1)

where $pc_{n,k}$ is the length of the period-certain guarantee (if any), $\Pr\left(\text{alive at } age\right)$ is the probability of being alive at a given *age* past 65 (conditional on being alive at 65)[1] based on the gender-specific life expectancy Social Security tables (Social Security Administration 2006)[2], δ is an annual discount factor set to 0.97 following 2011 OMB guidelines (OMB Circular A-94), and $income_{n,k,age}$ is the monthly income provided by the k-th annuity in the n-th choice set when the buyer reaches the given *age*. The latter is in turn determined by the starting income and the annual increases (if any). Note that for annuities with the period-certain guarantee, we implicitly assume that the annuity buyer cares equally about payout to himself/herself, and the payout to beneficiaries in the case of an early death. In our choice model, we assume that the buyer cares about the expected net present gain over the purchase price $V_{n,k} - price_{n,k}$. Since all annuities in our study cost p=\$100,000, the variation in expected gain is driven completely by the variation in $V_{n,k}$, so the model specification is almost identical to assuming consumers care about $V_{n,k}$. A rational buyer should also care about the financial strength of the company as measured by the AAA versus AA ratings. We include both the main effect of financial strength and its interaction with expected gain in our model. To motivate the interaction, note that the same expected gain is more certain when provided by an AAA versus AA company, so a rational buyer should value it more, ceteris paribus.

In addition to the effect of the total expected gain and the company's financial strength suggested by normative theory, we let several attributes enter utility directly to capture the "beyond NPV" idea discussed above. Specifically, we include the type and amount of annual increase and the level of the period-certain guarantee. All levels of these additional attributes are

[1] Note the study participants are asked to imagine they are already at age 65 when they are buying the annuity, and thus no adjustment should be made for actual current age or the chance of living until 65.

[2] Annuity issuers often maintain their own mortality tables which are adjusted for possible adverse selection among annuity purchasers. The effect on our estimates of using SSA mortality tables rather than issuer specific rates is a possible underestimation of the expected NPV per annuity. Thus, any estimates of under valuation per attribute should be considered conservative.

dummy coded and contained in a row attribute vector $X_{k,n}$.[3] We exclude starting income from $X_{k,n}$ to avoid strong collinearity: we find that the expected gain is too correlated with starting income for the model to separately identify the impact of starting income on utility beyond its impact on the expected payout. However, we did analyze an alternative specification of our model that replaces the expected net present gain with starting income, keeping the rest of the same (the detailed estimates of the alternative specification are available in the Online Appendix). Comparing our estimates with those from the alternative specification will be useful in interpreting our results.

Given the expected payout $V_{n,k}$, the dummy variable $AAA_{n,k}$, the price of the annuity p (which we fixed to \$100,000 throughout the study by design) and the $X_{k,n}$ variables, we model the respondent j's utility of the k-th annuity in the n-th choice set as a linear regression:

$$U_{n,k,j} = \underbrace{\alpha_j + \beta_j\left(V_{n,k} - p\right) + \gamma_j AAA_{n,k} + \delta_j AAA_{n,k} \times \left(V_{n,k} - p\right)}_{\text{normative model}} + \underbrace{X_{n,k}\theta_j}_{\substack{\text{direct effect}\\ \text{beyond NPV}}} + \varepsilon_{n,k,j} \tag{2}$$

where $\varepsilon_{n,k,j} \sim N(0,1)$ and we normalize the utility of the outside ("none of the above") alternative k=0 to zero to identify the parameters: $U_{n,0,j} = 0$. This normalization implies that the utility of inside alternatives should be interpreted as relative to self-management of a \$100,000 investment. Together with a simplifying assumption that $\varepsilon_{n,k,j}$ are independent, our model becomes a constrained version[4] of the multinomial probit model (Hausman and Wise 1978). The A individual-level parameters to be estimated are $\{\alpha_j, \beta_j, \gamma_j, \delta_j, \theta_j\}_{j=1}^{J}$, where θ_j is a column vector of the same length as $X_{k,n}$, and the rest are scalars.

To pool data across respondents j=1,2, . . ., J while allowing for heterogeneity of preferences, we use the standard hierarchical approach following Lenk et al. (1996)

[3] We do not include interactions of these direct effects with AAA for two reasons: 1) the normative effect of a risk-reduction due to stronger financial health is already captured in the interaction between AAA and expected net present gain and 2) the limited number of questions a survey respondent can answer before wearing out makes us unable to estimate such interactions in addition to all the other parameters of interest.

[4] The restriction of *one* of the scalar elements of the covariance of the $\varepsilon_{n,j}$ vector to unity is standard. The restriction of the entire covariance matrix to identity simplifies estimation and reflects our belief that the unobserved shocks associated with the individual annuity profiles are not heteroskedastic and not mutually correlated. The resulting model is sometimes called "independent probit" (Hausman and Wise 1978).

Table 2: Respondent Demographic and Psychographic Characteristics

Demographic or psychographic characteristic	Baseline treatment (334 respondents)			Enriched info treatment (323 respondents)			Same for both treatments	
	mean	median	std. dev	mean	median	std. dev	min	max
Age (years)	52.87	53	6.83	52.80	53	7.02	40	65
Male	0.41	0	0.49	0.40	0	0.49	0	1
Retirement savings 75to150K	0.13	0	0.34	0.17	0	0.38	0	1
Retirement savings over 150K	0.18	0	0.38	0.21	0	0.41	0	1
Perceived fairness of annuities	0.59	0.67	0.22	0.57	0.67	0.22	0	1
Loss aversion	0.66	0.7	0.29	0.68	0.7	0.29	0	1
Numeracy	0.50	0.5	0.16	0.50	0.5	0.15	0.125	1
Life expectancy (age at death)	85.77	87	8.03	84.80	86	9.01	59	99

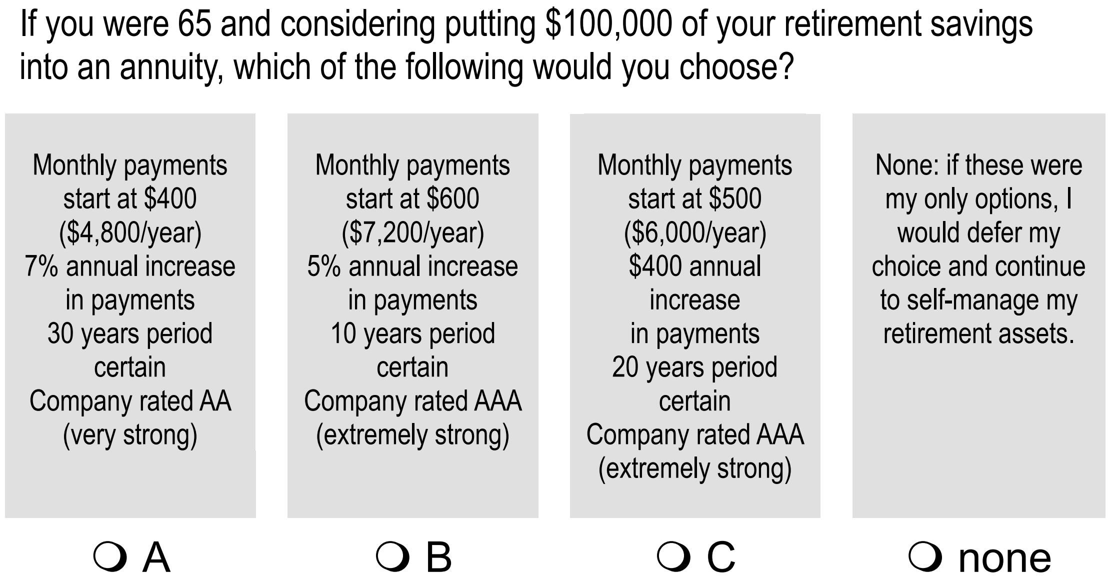
Statistical Design Optimization

Given the attribute levels in Table 1 and the model described above, we used SAS software (an industry standard) to generate the optimal choice-based survey design. We created the 20 choice sets using the %ChoicEff macro in SAS (Kuhfeld 2005), which finds utility-balanced efficient designs for choice-based conjoint tasks (Kuhfeld et al. 1994, Huber and Zwerina 1996). Because the design of the choice tasks is not intended to be the main contribution of our study, we merely strive to follow current practice and arrive at a reasonable design. Note that the design cannot be orthogonal by construction: the expected NPV is a combination of the other attributes. The non-linearity of the NPV formula allows us to still estimate the direct (beyond NPV) impact of each attribute other than starting income.

Estimation Methodology

To estimate the parameters of our choice model, we follow a standard Bayesian procedure to generate draws from the posterior distribution of all parameters using a Gibbs sampler. Please see Rossi et al. (2005) for a detailed description of setting up the Gibbs sampler for a hierarchical linear model. We ran the Gibbs sampler for 50,000 iterations, discarding the first 10,000 as burn-in iterations and using the remaining 40,000 draws to conduct our analysis of the results. As in the case of the experiment design, the estimation method is standard in the field.

Figure 1a: Sample Conjoint Choice Task



If you were 65 and considering putting $100,000 of your retirement savings into an annuity, which of the following would you choose?

Monthly payments start at $400 ($4,800/year) 7% annual increase in payments 30 years period certain Company rated AA (very strong)	Monthly payments start at $600 ($7,200/year) 5% annual increase in payments 10 years period certain Company rated AAA (extremely strong)	Monthly payments start at $500 ($6,000/year) $400 annual increase in payments 20 years period certain Company rated AAA (extremely strong)	None: if these were my only options, I would defer my choice and continue to self-manage my retirement assets.
○ A	○ B	○ C	○ none

Figure 1b: Sample Conjoint Choice Task with Cumulative Payouts In the enriched information treatment, the following table was shown directly under the task:

	Cumulative amount paid to you by different ages if you live to that age					
Age	70	75	80	85	90	95
Option A	$27,600	$66,300	$120,600	$196,800	$303,600	$453,400
Option B	$39,800	$90,600	$155,400	$238,100	$343,600	$478,400
Option C	$34,000	$78,000	$132,000	$196,000	$270,000	$354,000

Participants

We recruited participants through a commercial online panel from Qualtrics. Qualtrics does hundreds of academic research projects and also serves clients such as the US Army and government agencies. Panel members opt in to Qualtrics through various websites and are offered the opportunity to participate in surveys; Qualtrics does not actively solicit for its panel. For this project, we limited participation to individuals between the ages of 40 and 65 because this target group is the most appropriate for annuity purchases. We placed no limit on current retirement savings, but we collected data on savings as part of our demographic measures so that we could perform an analysis of how financial status affects preferences.

Because any survey attracts some respondents who either do not understand the instructions or do not pay attention to the task, we included an attention filter at the start of the survey and excluded participants who did not pass the filter. Our estimation sample consists of 334 respondents in the basic treatment and 323 in the enriched information treatment. Table 2 summarizes the respondent demographic and psychographic characteristics.

Procedure

We first presented participants with short descriptions of the annuity attributes being investigated (monthly income level, annual income increases, period-certain guarantees, and company ratings) as well as the full range of levels for each of these attributes. We told them the annuities were otherwise identical and satisfactory on all omitted characteristics. We also told them all annuities were based on an initial purchase price of $100,000 at age 65, consistent with prior experimental work on annuity choices (e.g., Brown et al., 2008). We then asked each participant to complete 20 choice tasks from one of the two conditions. To control for order effects, we presented the choice tasks in a random order. Figure 1 provides a sample choice task and illustrates the enriched information treatment. After completing all 20 choice tasks in their assigned condition, participants were asked to fill out the additional demographic and psychographic measures.

Results: Population Average Parameters and Their Interpretation

Although our experiment involved 20 choices between four options (three annuities and one outside option), a substantial proportion of respondents did not like any of the annuities on offer. Specifically, between 15 and 20 percent of respondents selected self-management in every task (see Table 2b for details). Some of the annuities in our design provided well over $200K in expected payout, in exchange for the $100K price of the annuity (held constant throughout). Therefore, we conclude that some people simply seem to dislike the idea of an annuity a priori, and are unwilling to consider these products. To be conservative in our analysis, we retain these “annuity haters” in the full estimation.

We omit the presentation of raw utility parameters $(\alpha,\beta,\gamma,\delta,\theta)$; please see the full-length paper for details. Because we are estimating a choice model, the raw utility parameters cannot be directly compared across treatments because of the well-known scaling problem (Swait and Louviere 1993). One transformation of the parameters that can be meaningfully compared is their ratio, and the most interesting ratio to consider is the ratio of “beyond NPV” parameters $(\alpha,\gamma,,\theta)$ to the expected gain parameter (β for AA annuity, $\beta+\delta$ for AAA annuity). Table 3 reports the standardized estimates for a AAA annuity, by treatment, setting the unit of currency to $100.

We call this ratio a “willingness to pay beyond NPV” (hereafter WTPbNPV) because for every attribute level, it measures the amount of expected present gain (delivered through changing starting income or other attributes) that would compensate for the presence of an attribute level relative to the baseline level of the same attribute. For example, the -$27.1 WTPbNPV of the “annual increase 3%” attribute means that on average, our respondents are indifferent between an annuity that includes a 3% annual increase and delivers an expected gain of $100, and another annuity that does not include annual increases and somehow (presumably via other attributes) delivers the same expected gain plus -$27.1, namely an expected gain of $72.9. Thus, WTPbNPV is willingness to pay while keeping the expected payout constant.

Table 3: Effect of Enriched Information: Average Beyond-NPV Willingness to Pay for Annuity features

	Proposed model specification					Alternative model specification				
Information treatment	**Basic**		**Enriched**			**Basic**		**Enriched**		
	Average beyond-NPV willingness to pay (WTPbNPV)	Posterior std. dev.of WTPbNPV	Average beyond-NPV willingness to pay (WTPbNPV)	Posterior std. dev.of WTPbNPV	Difference in average WTPbNPV (enriched-basic)	Average willingness to pay (WTP)	Posterior std. deviation of average WTP	Average willingness to pay (WTP)	Posterior std. deviation of average WTP	Difference in average WTP (enriched-basic)
Expected gain of $100 ($V_{n,k}$-$p=100$)	**$100.0**	*$0.0*	**$100.0**	*$0.0*	$100.0					
Starting monthly income of $100						**$100.0**	*$0.0*	**$100.0**	*$0.0*	$0.0
AAA rated issuer (vs. AA)	-$4.0	*$3.6*	-$1.9	*$2.5*	$2.1	-$23.1	*$45.3*	$49.6	*$39.7*	$72.7
Annual increase 3% (vs. 0)	**-$27.1**	*$4.5*	**-$9.6**	*$3.3*	**$17.5**	**$40.3**	*$16.1*	**$125.2**	*$15.6*	**$84.8**
Annual increase 5% (vs. 0)	**-$36.4**	*$4.1*	**-$9.7**	*$3.6*	**$26.6**	**$95.0**	*$15.4*	**$223.0**	*$17.8*	**$128.0**
Annual increase 7% (vs. 0)	**-$64.5**	*$4.7*	**-$34.1**	*$3.9*	**$30.4**	**$144.7**	*$17.2*	**$283.3**	*$19.3*	**$138.6**
Annual increase $200 (vs. 0)	**-$8.8**	*$4.4*	**-$7.8**	*$3.7*	$1.0	**$81.3**	*$16.7*	**$93.2**	*$15.1*	$11.9
Annual increase $400 (vs. 0)	**-$28.8**	*$4.1*	**-$13.7**	*$3.4*	**$15.1**	**$108.0**	*$17.5*	**$187.7**	*$16.1*	**$79.7**
Annual increase $500 (vs. 0)	**-$31.8**	*$4.6*	**-$15.8**	*$3.6*	**$16.0**	**$177.0**	*$19.6*	**$263.5**	*$20.0*	**$86.5**
Period certain 5 years (vs. 0)	**-$25.8**	*$6.1*	-$1.4	*$2.6*	**$24.4**	**-$109.0**	*$23.1*	-$11.3	*$11.5*	**$97.7**
Period certain 10 years (vs. 0)	$8.6	*$5.5*	**$7.4**	*$2.7*	-$1.2	**$59.0**	*$21.4*	**$46.9**	*$12.1*	-$12.2
Period certain 20 years (vs. 0)	**$26.6**	*$5.9*	**-$8.9**	*$3.3*	**-$35.4**	**$218.2**	*$25.5*	**$89.7**	*$13.8*	**-$128.5**
Period certain 30 years (vs. 0)	**-$39.8**	*$6.6*	**-$70.0**	*$5.1*	**-$30.3**	**$119.3**	*$25.8*	$29.9	*$16.5*	**-$89.4**

Note: The computations assume a AAA annuity. Average willingness-to-pay beyond NPV (WTPbNPV) parameters are derived from the individual parameters as follows: For each iteration of the Gibbs sampler, we divide the population average of all utility parameters by the population average of the coefficient on the expected payout (β+δ in equation 2 since we are considering a AAA annuity). The resulting draws of the population-average WTPbNPV are then used in computing both the posterior mean and the posterior standard deviation over all post–burn-in draws. In the alternative model specification, the same computations result in the more standard total willingness-to-pay (WTP). **Bold** indicates that 97.5% or more of the posterior mass has the same sign as the posterior mean—a Bayesian analogue of significance at the 5% level.

The WTPbNPV concept arising naturally from our proposed model specification can be contrasted with a more standard marginal willingness to pay (hereafter WTP) that results when the same ratio is calculated under the alternative model specification, in which the expected gain is replaced with starting income. Table 3 also contains all such “standard” WTP estimates; the raw parameter estimates of that specification (analogues of Tables 4a and 4b) are available in the Online Appendix. For example, the WTP of $40.3 for the 3% annual increase means that, on average, our respondents are indifferent between an annuity that includes a 3% annual increase and $100 of additional starting income and an otherwise identical annuity that does not include annual increases but involves $140.3=$100+$40.3 of starting income.

Comparing the WTPbNPV to WTP highlights the novelty of our model. Note that since WTPbNPV is measured in terms of expected gain and WTP is measured in terms of starting monthly income, the dollar quantities are not comparable between the two model specifications. However, one can safely compare their signs. In the case of 3% increase, the WTP is positive, meaning that 3% increase is more valuable than no increase while keeping initial monthly income and all other attributes the same. On the other hand, the WTPbNPV is negative, meaning that 3% is less valuable than no increase while keeping the expected payout the same.

Estimation Results: Average Preferences in the Basic Information Treatment

We first consider the results for the basic information treatment. Several conclusions can be drawn from the WTPbNPVs (in Table 3). As expected, the average coefficients on both the expected gain and its interaction with the AAA rating are positive. The insignificant coefficient on the AAA dummy shows that consumer preference for financially safe issuers manifests itself solely via an increased weight on expected gain, and not as a shift in the intercept of the utility function. A qualitative comparison with the alternative model specification rules out a simplistic theory about the antecedents of the significant interaction between AAA and expected gain: Under the alternative model specification (see full-length paper), neither the AAA dummy nor its interaction with starting income are significant at the population level, suggesting that the significant coefficient on *Expected_gainXAAA* is not merely capturing the respondents' higher valuation of starting income when it is provided by a AAA issuer. Instead, the respondents seem to value some NPV-like combination of the starting income with other attributes (annual increases and/or certainty guarantees) more when it is provided by an AAA issuer.

The coefficients on the annual-increase and period-certain dummies are mostly significant and often large, indicating consumer behavior is not well-captured by using only the expected payout and financial-strength variables. We discuss each of the "beyond NPV" influences from these different attributes in turn.

Annual Increases

The negative signs on all of the percentage increase coefficients suggest that consumers systematically undervalue the benefits of annual payment increases. From the WTPbNPV estimates, we can see that the magnitude of the undervaluation can be large, especially for the percentage increases. For example, the WTPbNPV of -$64.5 on the 7% annual increase means our respondents are indifferent between an annuity that generates an expected gain of $100 with a constant monthly income, and another annuity that generates $164.50 expected gain by starting at a lower monthly income level and adding 7% per year. In contrast, the WTPs under the alternative model specification are all positive. Together, these results indicate that consumers pay attention to increases and value them positively, but they systematically undervalue them relative to their true expected value.

The additive increases exhibit a similar pattern, but they are generally under-valued less, echoing the results of McKenzie and Liersch (2011). To see the difference in Table 3, recall that we selected the levels of annual increases as pairs matched across the type of increase (additive vs. percentage). Specifically, the $500/year increase results in approximately[5] the same expected payout as the 7% increase, and the ($300, 5%) and ($200, 3%) pairs are matched analogously. Therefore, we can compare the WTPbNPV numbers within these matched pairs, and conclude that the average consumer prefers additive increases to percentage increases, ceteris paribus. In

[5] The magnitude of the difference in expected payout depends on gender, starting income, and other attributes.

the full-length paper we quantify the difference in terms of demand by simulating the magnitude of the effect of various increases on total market demand using counter-factual experiments.

Period Certain

The positive average coefficient on the 20-year period-certain guarantee suggests consumers like this option beyond its financial impact on the expected payout. Conversely, the short (5-year) and very long (30-year) period-certain guarantees are undervalued. The WTPs under the alternative model specification reveal that consumers do not only under-value the 5-year period certain while keeping expected payout the same, they also undervalue it relative to no period certain while keeping other attributes the same. Moreover, the WTP for a 30-year period certain is about half the WTP of a 20-year period certain despite the much higher expected payout from the former. Therefore, the inverse-U pattern we find is not an artifact of our specification or our particular calculation of the expected gain.

Effect of the Enriched Information Treatment on Average Preferences

Recall that only the standardized coefficients (WTPbNPV, in Table 3) can be meaningfully compared across treatments. Table 3 provides both the WTPbNPVs for the enriched information treatment and the difference between treatments.

We offer three observations: First, the magnitudes of the WTPbNPVs for annual increases are much smaller in the enriched condition, thereby indicating the apparent disliking of the increases in the basic treatment may be due to the subjects' inability to "do the math" on compounding, and not to a more fundamental aversion. The WTPs under the alternative model specification all increase, supporting the interpretation that respondents now value increases more. At the same time, however, the WTPbNPVs are still negative, indicating that the respondents still undervalue increases even in the enriched information condition.

Second, the difference between additive and percentage increases mostly vanishes in the enriched treatment, with the exception of the (7%, $500) pair which still exhibits a larger under-valuation of the percentage increase. But even in that extreme pair, the dollar difference between the WTPBNPVs is reduced from about $33 to about $18. This finding agrees with prior work in the literature on individuals' difficulty with compounding in financial decisions (e.g., Wagenaar and Sagaria 1975, McKenzie and Liersch 2011). By seeing a table of cumulative payouts, individuals can better appreciate the impact of the percentage increases over time.

Finally, respondents in the enriched treatment continue to exhibit the inverse-U relationship between preferences and the duration of period certain guarantees (even under the alternative model specification), but the peak of the preference shifts towards shorter periods (10-year period becomes the most over-valued). The persistence of the inverse-U relationship across the two information treatments suggests the relationship is not fundamentally driven by miscalculation or inability to "do the math" when estimating guarantees' impact on payout.

Estimation Results: Population Heterogeneity of Preferences

We find a lot of heterogeneity in preferences, some of which can be explained by variance in demographics and psychographics, and some of which remains unexplained. One effect stands out as large: regardless of the information treatment[6], we find that perceived fairness of annuities is strongly correlated with their baseline liking. In the enriched information treatment, individuals with higher levels of perceived fairness also value expected gain more. In the basic information treatment, individuals with higher levels of perceived fairness also show increased liking of annual increases beyond NPV, but not increased enough to de-bias them.

Several other effects of demographics and psychographics also deserve a mention: As one would expect, more numerate individuals care more about the expected payoff regardless of treatment. More surprisingly, they also undervalue annual increases even more than less numerate people, especially in the basic information treatment.

Finally, as a rational model would predict, higher life expectancy increases the liking of annual increases, but this effect only exists in the enriched information treatment. To see how much longer than average a respondent needs to expect to live to eliminate the under-valuation of annual increases, one can calculate the ratios of the population-average beyond-NPV coefficients and the Δ coefficient on demeaned life expectancy. The result is between 8 and 17 years, i.e., between one and two standard deviations of life expectancy. Hence, we find that the enriched treatment de-biases the valuation of annual increases of people who expect to live more than a standard deviation longer than the average life expectancy in the population.

Discussion

This paper proposes a model of consumer preferences for attributes of immediate life annuities, and estimates the model using stated preferences in a discrete choice experiment with a national panel of people aged 40–65 years. Our main methodological contribution is a model specification that allows direct measurement of the direct influence of attributes on preferences beyond their impact via the expected net present value of the annuity, a.k.a "beyond NPV." We find that consumers value increases in the expected net present value of the payouts, but some annuity attributes also influence consumer preferences directly, beyond their impact on financial value.

One of the main managerial contributions of our model is the design of products that maximize demand without increasing the expected payout. The highest-demand products are good "smart defaults" (Smith, Goldstein, and Johnson 2013), candidates for policy makers interested in increasing annuitization. We find that careful "packaging" of a given net present value into the optimal mix of the attributes can more than double demand for annuity products relative to the poorest performing attribute mixes. Regardless of the information treatment, the demand-maximizing annuities do involve medium-length period-certain guarantees and no annual increases. The optimal length of the period certain guarantee depends on the information treatment: it is shorter when information is enriched.

[6] Recall that we cannot compare the coefficients between Tables 4a and 4b directly (Swait and Louviere 1993). We thus confine ourselves to broad qualitative observations of the effect of the enriched information on our estimates.

Suzanne B. Shu

Robert Zeithammer

John Payne

References

Benartzi, Shlomo, Alessandro Previtero, and Richard Thaler (2011), "Annuitization Puzzles," The Journal of Economic Perspectives, 25(4), 143–164.

Brown, J. R. (2007), Rational and behavioral perspectives on the role of annuities in retirement planning. (NBER Working Paper No. 13537). Cambridge, MA: National Bureau of Economic Research, Inc.

Brown, J. R., Kling, J.R., Mullainathan,S. & Wrobel, M.V. (2008), "Why don't people insure late-life consumption? A framing explanation of the under-annuitization puzzle," American Economic Review, 98(2), 304–09.

Goldstein, Daniel G. and Hershfield, Hal E. and Benartzi, Shlomo, 2014. The Illusion of Wealth and Its Reversal. Working paper.

Hausman, Jerry A. and David A. Wise (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," Econometrica, 46(2), 403–426.

Hu, W.Y., and Scott, J.S. (2007), "Behavioral obstacles to the annuity market," Financial Analysts Journal, 63(6), 71–82.

Huber, Joel and Klaus Zwerina (1996). "The Importance of Utility Balance in Efficient Choice Designs," Journal of Marketing Research, 33(3), 307–317.

Kahneman, D., Knetsch, J. & Thaler, R.H. (1986), "Fairness as a constraint on profit-seeking: Entitlements in the market." American Economic Review, 76(4), 728–741.

Kuhfeld, Warren F. (2005). "Marketing Research Methods in SAS." *Experimental Design, Choice, Conjoint, and Graphical Techniques*.

Lenk, P., W. DeSarbo, P. Green, and M. Young. (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs." *Marketing Science* 15 (2):173–91.

McCulloch, Robert, and Peter E. Rossi (1994) "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics*, 64(1), 207–240.

McKenzie, Craig and Michael J. Liersch (2011) "Misunderstanding Savings Growth: Implications for Retirement Savings Behavior." *Journal of Marketing Research* 48, S1–S13.

Orme, B. (2006). Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Madison, WI: Research Publishers LLC.

Payne, John, Namika Sagara, Suzanne B. Shu, Kirsten Appelt, and Eric Johnson (2013). "Life Expectancy as a Constructed Belief: Evidence of a Live-To or Die-By Framing Effect," *Journal of Risk and Uncertainty*, 46, 27–50.

Rossi, Peter, Greg Allenby, and Robert McCulloch. (2005), *Bayesian Statistics and Marketing*. Hoboken NJ: Wiley.

Scott, J. S. (2008). "The longevity annuity: An annuity for everyone?" *Financial Analysts Journal, 64*(1), 40–48.

Smith, Craig and Daniel G. Goldstein, and Eric J. Johnson (2013) "Choice Without Awareness: Ethical and Policy Implications of Defaults." Journal of Public Policy & Marketing 32 (2), 159–172.

Swait, Joffre and Jordan Louviere. (1993). "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research*, 30(3), 305–314.

Wagenaar, Willem A., & Sabato D. Sagaria (1975). "Misperception of exponential growth." *Attention, Perception & Psychophysics*, 18, 416–422.

Comparing Message Bundle Optimization Methods: Should Interactions Be Addressed Directly?

Dimitri Liakhovitski[1]
GfK
Faina Shmulyian
MetrixLab
Tatiana Koudinova
GfK

Abstract

Frequently used Message Bundle Optimization Methods (MBOMs) ignore the issue of semantic interactions among messages, or address it indirectly. This research compares the performance of several MBOMs in studies with real and simulated respondents. Findings suggest the need to address the issue of semantic interactions directly.

Introduction

Marketers want to communicate to the consumer/client as much information as possible about their product/service. However, the amount of space/time available for such messages is limited. Thus, they frequently formulate a number of claims or messages that could be used in marketing communications and then approach marketing scientists with the following question:

- What is the optimal (best) bundle of two messages; 2nd best bundle of two messages, 3rd best bundle . . . etc.—out of all the messages that we've formulated.

- What is the optimal bundle of three messages; 2nd best bundle of three messages, 3rd best bundle . . . etc.—out of all the messages that we've formulated.

- What is the optimal bundle of four messages; 2nd best bundle of four messages, 3rd best bundle . . . etc.—out of all the messages that we've formulated.

Marketing scientists have developed a number of MBOMs to address these questions. Study 1 compared the performance of several MBOMs in an empirical study. Study 2 was a computer simulation—it compared the performance of several techniques using synthetic respondents.

Study 1

MBOMs Tested

We catalogued several frequently used MBOMs. The tables below describe each method—how the data are gathered and how the final sample level metric is calculated for each bundle of messages.

[1] Dimitri.Liakhovitski@gfk.com, Faina.Shmulyian@metrixlab.com, Tatiana.Koudinova@gfk.com

MaxDiff Shares:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
-Traditional MaxDiff exercise with all messages -MaxDiff responses fed into a traditional HB estimation in SSI Web.	Sum of shares: Sum of the *rescaled probability shares* of the individual messages that comprise the bundle, averaged across respondents	With 4 messages per screen and the raw utility of Message 1 = x_1, the rescaled probability for that message for each respondent is: $\exp(x_1) / (\exp(x_1)+3) /$ $\sum_1^k \exp(xi) / (\exp(xi) + 3)$

MaxDiff Based TURFs:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
-Traditional MaxDiff exercise with all messages -MaxDiff responses fed into a traditional HB estimation in SSI Web.	% reach: % of respondents reached by the bundle	Reach is always dichotomous (reached vs. not reached) Two possible operationalizations of reach: - One SD above the mean: each respondent is reached by the messages whose probability of choice are >1 SD higher than his/her mean probability of choice across all messages - Top 3 messages: each respondent is reached by his/her best 3 messages

Anchored MaxDiff Shares:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
- MaxDiff exercise with all messages. On each MaxDiff screen the respondent answers the following question: *Considering just the messages shown, select one that applies:* *All messages are important* *None of the messages are important* *Some are important, some are not* - The anchored MaxDiff responses are fed into SSI Web's Anchored MaxDiff estimation; as a result, individually important messages have utilities >0 and non-important messages have utilities <0.	Sum of shares: Sum of the *rescaled probability shares* of the individual messages that comprise the bundle, averaged across respondents % reach: % of respondents reached by the bundle	A respondent is reached by all messages with utilities >0. A respondent is reached by a bundle of messages if s/he is reached by at least one message in the bundle.

Ratings-Based TURF:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
- Each respondent rates each message on a 1 to 7 scale.	% reach: % of respondents reached by the bundle	Two possible operationalizations of reach: -Top box -Top 2 box

DCM Shares: DCM with Bundles (Partial Profile Conjoint), No Message Categories:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
- A conjoint exercise; each alternative is a bundle of 3 or 4 messages; - Respondents pick the best bundle on each task; - CBC responses fed into a traditional HB estimation.	Sum of shares: Choice probabilities of the message bundles, averaged across respondents.	Estimation could be done: - Without constraints: only main effects - With constraints (message present > no message): main effects only - With constraints: main effects + significant interactions Individual message choice probability: Assuming the sum of applicable utilities for a given bundle is y_1, and for the other 3 bundles it is y_2, y_3, and y_4, the choice probability for that first bundle for each respondent is: $\frac{\exp(y_1)}{\exp(y_1)+\exp(y_2)+\exp(y_3)+\exp(y_4)}$

DCM with Categories Shares:

Data Elicitation & Analysis	Sample Level Bundle Metrics Calculated	Notes
- Respondents rank messages within Z message categories; - Only the best and the worst messages from each category are piped into the later CBC. - CBC with message bundles as alternatives; each bundle has Z messages—each from one of Z categories; - CBC responses fed into a traditional HB estimation.	Sum of shares: Choice probabilities of the message bundles, averaged across respondents.	CBC HB estimation has to be constrained: in each category, the best message has to be preferred to the worst message; Interaction effects between categories can be tested. Each respondent has utilities only for his/her best and worst message in each category; The "middle" messages in each category are assigned utilities in between the best and the worst one, equidistantly—based on rankings.

Among these MBOMs only DCM based methods allow to include interaction effects in the model. However sample size and number of messages restrict how many (if any) interactions can be included.

Study Background and Design

In Study 1 we used 16 unbranded painkiller-related messages selected from several previous GfK US studies. Study 1 was fielded in early December 2013, online, with 1018 US respondents from opt-in panels who had purchased non-prescription pain medications in the past 3 months.

Below are 16 messages arranged in 4 a-priori content categories:

Speed of action	**General Effectiveness**	**Trustworthiness**	**Long Lasting**
Relieves your pain so you quickly feel like yourself again. Gets you back on track before anyone knows you hurt. No other pain reliever relieves pain faster. Effectively relieves pain within 15 minutes.	Helps you control your pain. Relieves your pain so you can get back to your life. Is the most effective pain relief you can buy. Confronts your pain, so you can get on with your day.	Contains the medicine prescribed most for tough pain. Is the #1 choice for pain. Was developed by pain experts and is endorsed by doctors. Used most by doctors for fast, all day relief of tough pain.	Powerful pain relief that lasts longer than any other pain reliever. Just 2 doses provides pain relief for a full 24 hours. Works through the night so you can get a good night's sleep. The longest-lasting pain reliever available without a prescription.

Survey respondents were randomly assigned to one of 5 experimental cells. The survey did not take long; the flow for each cell is shown below:

Study Cells	**Sample Size**	**Median Survey Length**	**Section 1**	**Section 2**	**Section 3**	**Section 4**	**Section 5**
1. Anchored MD	201	11.13 min	Screeners	Anchored MaxDiff	**Holdout tasks**		
2. Category DCM	204	15.19 min	Screeners	Rankings by category	MaxDiff	Category DCM	**Holdout tasks**
3. Ratings	203	6.03 min	Screeners	Message ratings	**Holdout tasks**		
4. DCM	209	8.37 min	Screeners	DCM	**Holdout tasks**		
5. Tradit. MD	201	7.5 min	Screeners	MaxDiff	**Holdout tasks**		

The table below shows what MBOMs could be applied to analyze the data collected in each study cell:

Study Cells:	MaxDiff Shares	All MaxDiff Based TURFs	Anchored MaxDiff	All Anchored MaxDiff Based TURFs	Rating Based TURF	DCM Shares	DCM w Categories Shares
1. Anchored MaxDiff	+	+	+	+			
2. Category DCM[2]	+	+					+
3. Ratings					+		
4. DCM						+	
5. Tradit. MaxDiff	+	+					

Holdout Tasks, Thurstone Scores, and Metric for Judging Predictive Performance of MBOMs

To compare predictive performance of MBOMs, at the end of the survey respondents in all study cells were shown the same six holdout tasks. Each task had four message bundles. The first three holdout tasks had bundles of 3 messages each, the last three holdout tasks had bundles of 4 messages each (see Appendix). For each holdout task respondents were asked to rank bundles from 1 ("would most motivate to purchase a painkiller") to 4 ("would least motivate to purchase a painkiller"). The ranks were used to calculate Thurstone scores (see the table in the Appendix). Thurstone scores were consistent across five study cells.

For each holdout task bundle, a sample level metric was calculated based on the MBOM under consideration. Then, a Pearson correlation was calculated between four message bundles' Thurstone scores and their MBOM metric. A high correlation implied a good prediction of the respondents' true preferences by the MBOM under consideration.

Study 1 Findings

The table below shows correlations between each method's scores for 4 holdout message bundles and their respective Thurstone scores for Study Cell 1 (Anchored MaxDiff). MaxDiff Sum of Shares outperformed all MaxDiff based TURFs:

[2] MaxDiff data was also collected in Cell 2, but it was not used for Rankings + DCM method

	TURFs						
Holdout Screens	**Anch. MD Utilities** (reached if >0)	**Anch. MD Shares** 1 SD > mean	**Anch. MD Shares** Top 3 Messages	**MD Shares** 1 SD > mean	**MD Shares** Top 3 Messages	**Anch. MD Shares**	**MD Shares**
1	0.74	0.96	0.95	0.96	0.94	0.95	0.95
2	0.93	0.73	0.76	0.72	0.82	0.79	0.78
3	0.90	0.96	0.93	0.97	0.92	1.00	1.00
4	-0.19	0.18	0.09	0.29	0.26	0.99	0.98
5	0.99	0.98	0.98	0.98	0.98	0.99	0.99
6	0.85	0.72	0.64	0.66	0.70	0.70	0.69
Average across 6 screens	**0.70**	**0.76**	**0.73**	**0.76**	**0.77**	**0.90**	**0.90**

The table below shows correlations between each technique's scores for 4 holdout message bundles and their respective Thurstone scores for Study Cell 2 (Category DCM). MaxDiff Sum of Shares outperformed other techniques[3]. DCM with Categories came in close second.

	TURFs			
Holdout Screens	**MD Shares** 1 SD > mean	**MD Shares** Top 3 Messages	**MD Shares**	**DCM w. Categories** (w Constr.): **Shares**
1	0.98	0.97	1.00	0.95
2	0.75	0.79	0.54	0.74
3	0.97	0.96	0.98	0.96
4	-0.27	-0.09	0.72	0.55
5	0.76	0.64	0.93	0.78
6	0.91	0.93	0.81	0.80
Average across 6 screens	**0.68**	**0.70**	**0.83**	**0.80**

The table below shows correlations between each technique's scores for 4 holdout message bundles and their respective Thurstone scores for Study Cell 3 (Ratings). Unimpressive performance:

[3] We could test MaxDiff's performance because this cell's respondents completed a regular MaxDiff as well.

	Ratings-based TURFs	
Holdout Screens	Top Box = reached	Top 2 Box = reached
1	0.71	0.96
2	0.15	0.43
3	0.97	0.90
4	0.31	0.11
5	-0.37	0.92
6	0.74	0.12
Average across 6 screens	**0.42**	**0.57**

The table below shows correlations between each technique's scores for 4 holdout message bundles and their respective Thurstone scores for Study Cell 4 (DCM). DCM with constraints, and especially when significant interactions were taken into account, performed best.

Holdout Screens	**DCM Shares** no Constraints	**DCM Shares** w Constraints	**DCM Shares** w Constraints & Interactions
1	0.98	0.95	0.99
2	0.93	0.98	1.00
3	0.92	0.96	1.00
4	0.38	0.63	0.79
5	0.36	0.91	0.95
6	0.93	0.97	0.94
Average across 6 screens	**0.75**	**0.90**	**0.95**

The table below shows correlations between each technique's scores for 4 holdout message bundles and their respective Thurstone scores for Study Cell 5 (Traditional MaxDiff). Again, Sum of Shares outperformed the TURFs.

Correlations	**MaxDiff TURFs**		
Holdout Screens	MD Shares 1 SD > mean	MD Shares Top 3 Messages	**MD Shares**
1	0.94	0.94	0.99
2	0.71	0.74	0.62
3	0.80	0.76	0.99
4	-0.12	-0.12	0.39
5	0.70	0.63	0.90
6	0.80	0.87	0.76
Average across 6 screens	**0.64**	**0.64**	**0.77**

A comparison of correlations across study cells and across MBOMs (table below) shows that MaxDiff Sum of Shares and DCMs demonstrated superior performance while TURFs performed poorly.

	TURFs								
Study Cell:	**MD Shares:** 1 SD > mean	**MD Shares:** Top 3 Messages	**Ratings** TB	**Ratings** T2B	**MD Shares**	**DCM Shares** no Constr.	**DCM Shares** w. Constr.	**DCM Shares** w. Interact**.**	**DCM w. Cat.: Shares**
1. Anchored MaxDiff	.76	.73			.90				
2. Category DCM	.68	.70			.83				.81
3. Ratings			.42	.57					
4. DCM						.75	.90	.95	
5. MaxDiff	.64	.64			.77				

Study 1 Conclusions

Techniques based on straightforward summation of individual message metrics (shares) have performed best. The top performers were:

- DCM without categories (with constraints)—sum of shares—especially with interactions
- MaxDiff (regular or anchored)—sum of shares
- Category DCM—sum of shares

Both MaxDiff based TURFs and ratings based TURFs performed considerably worse. Why did the TURFs performed so poorly?

- The ultimate objective of MBO research is to answer a bundle focused question: On average, how much do people like a specific message bundle?
- TURF based methods answer a sample focused question: How many people like at least one message in a bundle?
- The response to the latter question might not be a successful proxy for answering the former.

At the same time, neither MaxDiff (regular or anchored) nor TURF can capture semantic interactions between messages. How would these methods perform in the presence of interactions between messages?

Study 2 was supposed to help us address the questions raised above.

Study 2

Study 2 Motivation

Potential semantic interactions between messages in a bundle seem an important issue that needs to be addressed. However, most MBO techniques we tested in Study 1 ignored semantic interactions between messages. The objective of Study 2 was to use simulated data to examine how presence and strength of interactions affect accuracy of TURF and MaxDiff estimates.

Underlying Assumption

The basic additive interaction model was used for the analysis. A utility of a combination of messages *i* and *j* for a particular respondent was presented as following:

$$\textbf{Utility of a Pair} = \beta_i + \beta_j + \beta_{ij},$$

where β_i and β_j are individual utilities of messages *i* and *j*—**main effects**, and β_{ij} is the *additional* utility of the combination of messages *i* and *j*—**interaction effect.**

The interaction could be positive or negative and different in strength. An interaction between messages in bundles is called “weak” if the absolute value of the interaction term is small relative to the individual utilities. If the absolute value of the interaction term is higher than most of the sums of individual utilities and if the interaction utility can “override” these sums, it will represent “strong” interactions between messages.

Two additional assumptions were made in the simulations:

- Message order (i and j) has no impact on interaction effect;
- Higher order interaction effects (3+ messages) are negligible.

The assumptions are realistic for most of the MBO studies. Also, they can be released or ignored, but it will result in significant increase of computational complexity of the estimations.

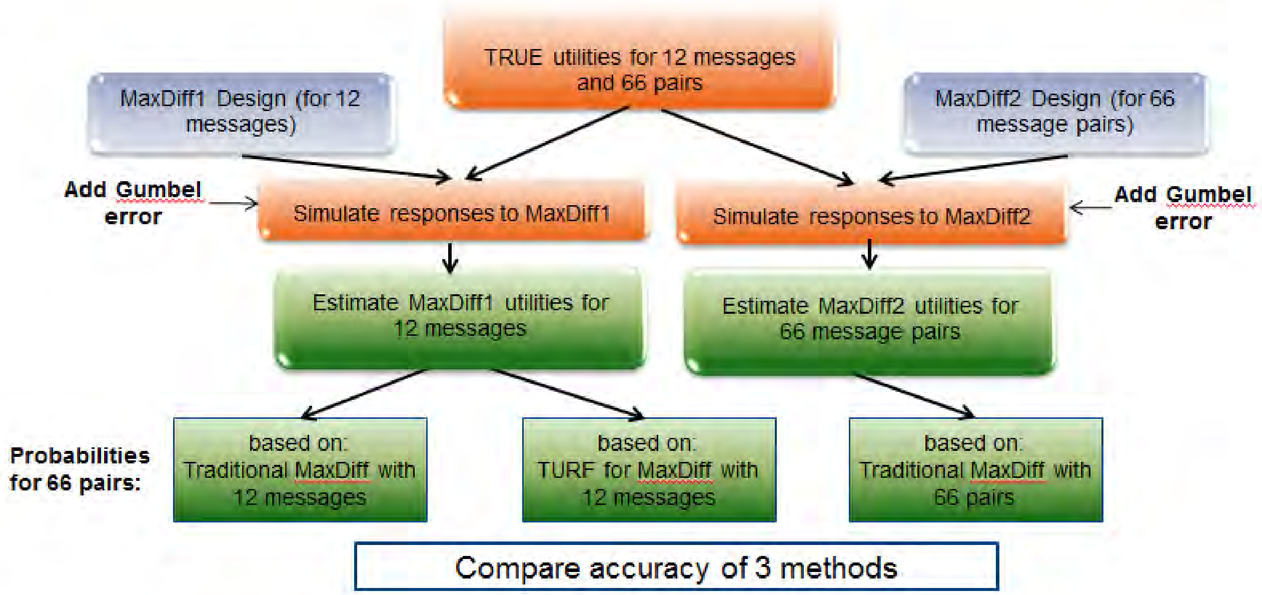
Study 2 Design

The flow of the study is presented on Exhibit 1. First, a set of artificial utilities for 12 messages was generated using Multinomial Normal Distribution with parameters (means for each utility, average variance across utilities and variance across respondents) taken from a real study. 200 “artificial respondents” were generated. Variance of 12 utilities/respondent ranged from 0.9 to 4.7.

Multiple sets of utilities for all 66 pairs of 12 messages for the same “artificial respondents” were produced using individual utilities generated for 12 messages and simulating interaction terms of different strength and sign.

The designs for two MaxDiffs—for 12 individual messages (MaxDiff1) and for 66 pairs (MaxDiff2)—were generated in Sawtooth Software’s SSI Web. For MaxDiff1, the design consisted of 9 sets of 4 messages. For MaxDiff2, two different designs were tested—19 sets of 7 pairs and 28 sets of 7 pairs. Using artificial utilities and applying Gumbel error, responses were simulated for both MaxDiffs. Individual utilities were estimated in CBC HB.

The focus of the study was on bundle optimization, therefore the utilities for 12 individual messages were used to derive probabilities for 66 pairs mimicking approaches considered in Study 1—Traditional MaxDiff to estimate probability of pairs based on individual probabilities for each message and different variations of TURF with individual utilities for 12 messages. The probabilities estimated using the MaxDiff1 utilities were compared with the ones estimated directly for 66 pairs in MaxDiff2. The flow of the study is in the next exhibit.


TRUE utilities for 12 messages and 66 pairs
MaxDiff1 Design (for 12 messages)
MaxDiff2 Design (for 66 message pairs)
Add Gumbel error
Simulate responses to MaxDiff1
Simulate responses to MaxDiff2
Add Gumbel error
Estimate MaxDiff1 utilities for 12 messages
Estimate MaxDiff2 utilities for 66 message pairs
Probabilities for 66 pairs:
based on: Traditional MaxDiff with 12 messages
based on: TURF for MaxDiff with 12 messages
based on: Traditional MaxDiff with 66 pairs
Compare accuracy of 3 methods

To facilitate reading and interpretation of the results in Study 2, an actual message was assigned for each of the 12 items tested in the study. The messages and average artificial (true) utilities across respondents are presented in the table below. These messages are similar to the ones tested in a real study.

	Message	β (Utility)
1	99% of Cats Can't Resist!	2.00
2	99% of Cats Love!	1.60
3	America's #1 Cat Treat	1.20
4	9 out of 10 Cats Love!	0.90
5	95% of Cats Can't Resist!	0.60
6	95% of Cats Love!	0.25
7	9 out of 10 Cats Can't Resist!	-0.10
8	40% More!	-0.45
9	Cats Can't Resist!	-0.85
10	Picky Cats Love!	-1.20
11	Turns any time into playtime!	-1.60
12	40% More Love!	-2.00

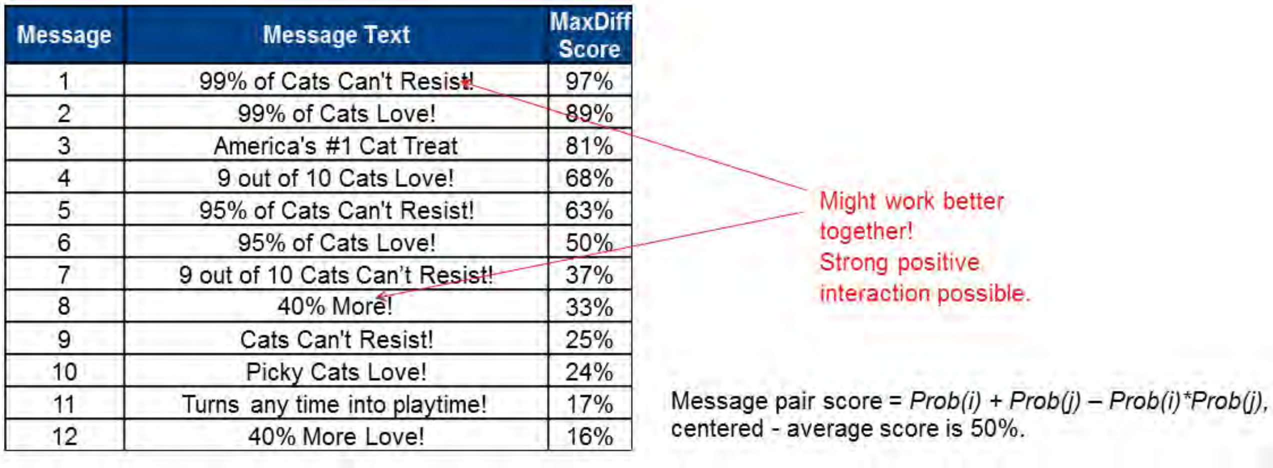
In the study, artificial utilities for pairs were tested assuming different strength of semantic interactions between the messages. In eight sets of utilities for pairs, absolute values of the interaction terms were gradually increased starting from zero up to the ones that correspond to a very strong interaction.

Study 2 Results

First, individual utilities for 12 messages were analyzed following the logic of standard MBO methods. The MaxDiff1 scores (average estimated probabilities) for each message are presented in the table below.

If the solution for MBO is based on MaxDiff1 scores for individual messages and semantic interactions are not taken into account, the winning pair will consist of two very similar messages—Message 1 and Message 2.

In presence of synergies, Message 1 and Message 8 could be a much better pair.



Message	Message Text	MaxDiff Score
1	99% of Cats Can't Resist!	97%
2	99% of Cats Love!	89%
3	America's #1 Cat Treat	81%
4	9 out of 10 Cats Love!	68%
5	95% of Cats Can't Resist!	63%
6	95% of Cats Love!	50%
7	9 out of 10 Cats Can't Resist!	37%
8	40% More!	33%
9	Cats Can't Resist!	25%
10	Picky Cats Love!	24%
11	Turns any time into playtime!	17%
12	40% More Love!	16%

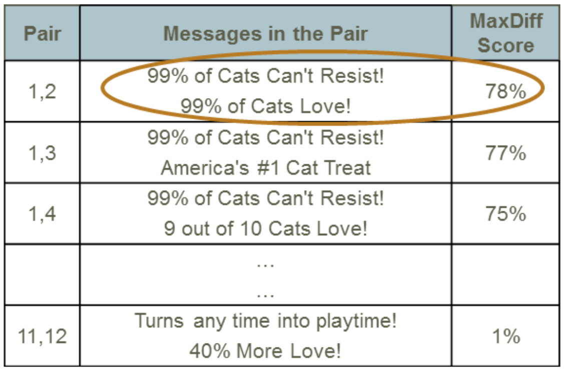
Might work better together!
Strong positive interaction possible.

Message pair score = *Prob(i)* + *Prob(j)* – *Prob(i)*Prob(j)*, centered - average score is 50%.

The next three tables summarize the difference in MBO results based on MaxDiff for individual messages and MaxDiff for pairs (MaxDiff1 and MaxDiff2) in presence of strong semantic interactions with signs pre-selected for each pair. The signs were preselected taking into account actual meaning of the messages in the pair. For example, it was assumed that the messages "99% of Cats Can't Resist!" and "99% of Cats Love!" are too similar and have a negative interaction of different amplitude for all respondents. Messages "99% of Cats Can't Resist!" and "40% More!" work well together have a positive interaction for all respondents.

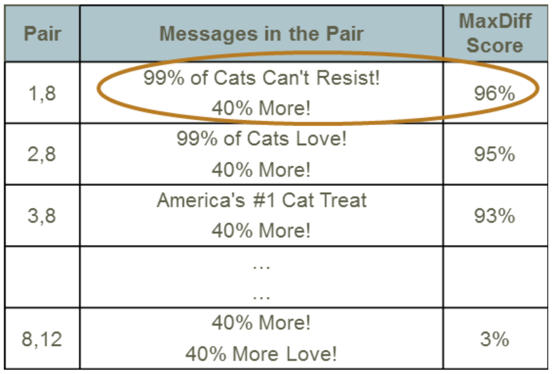
The tables below illustrate that MaxDiff for 66 pairs, where we take into account semantic interactions produces more consistent, meaningful, and actionable results.

Table: MaxDiff for 12 messages



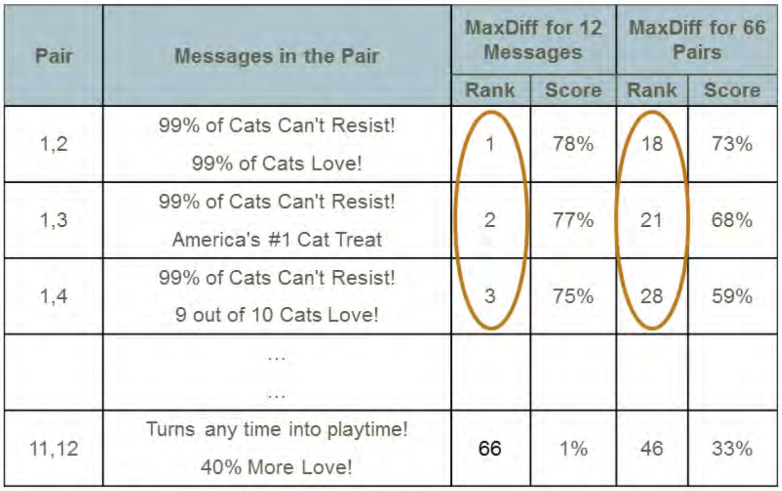
Pair	Messages in the Pair	MaxDiff Score
1,2	99% of Cats Can't Resist! 99% of Cats Love!	78%
1,3	99% of Cats Can't Resist! America's #1 Cat Treat	77%
1,4	99% of Cats Can't Resist! 9 out of 10 Cats Love!	75%
	… …	
11,12	Turns any time into playtime! 40% More Love!	1%

Table: MaxDiff for 66 Pairs



Pair	Messages in the Pair	MaxDiff Score
1,8	99% of Cats Can't Resist! 40% More!	96%
2,8	99% of Cats Love! 40% More!	95%
3,8	America's #1 Cat Treat 40% More!	93%
	… …	
8,12	40% More! 40% More Love!	3%

Table: Rank order for pairs in MaxDiff1 and MaxDiff2

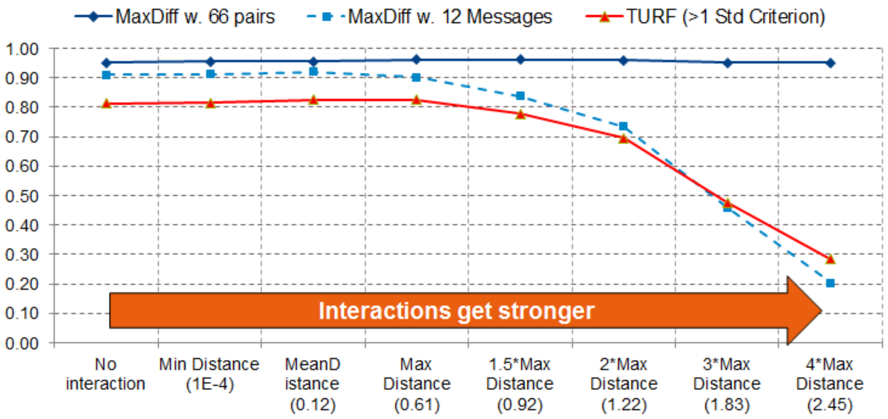


Pair	Messages in the Pair	MaxDiff for 12 Messages		MaxDiff for 66 Pairs	
		Rank	Score	Rank	Score
1,2	99% of Cats Can't Resist! 99% of Cats Love!	1	78%	18	73%
1,3	99% of Cats Can't Resist! America's #1 Cat Treat	2	77%	21	68%
1,4	99% of Cats Can't Resist! 9 out of 10 Cats Love!	3	75%	28	59%
	… …				
11,12	Turns any time into playtime! 40% More Love!	66	1%	46	33%

In Study 2, TURF was applied to generate best pairs of messages using the same methodology as in Study 1. Different criteria for reach used with TURF are described earlier within this article. The performance of the TURF-like methods was poor especially with strong semantic interactions simulated in the data. TURF performance did not strongly depend on the reach criteria chosen for the model.

Exhibit 2 presents performance of different methods—MaxDiff1, TURF and MaxDiff2—in presence of semantic interactions of different strength. To measure the performance, we use Pearson correlation between average "True" utilities for 66 pairs of messages and the scores for each pair estimated in MaxDiff1 and MaxDiff2 and also the shares of each pair estimated in TURF with "more than one standard deviation" criteria.

Exhibit 2: Pearson Correlations between Mean "True" Utilities for 66 Pairs and Metrics in the 3 Methods with Sign of the Interactions Manually Selected.



The results are in line with Study 1. MaxDiff2 for 66 pairs is able to accurately estimate utilities independently of the interaction strength. Performance of MaxDiff1 and TURF is not significantly worse than MaxDiff2 performance for weak or moderate interactions, but becomes much worse for strong interactions. For weak and moderate interactions, MaxDiff1 shows better results than TURF.

The Spearman correlations with "True" utilities for the same experiment setup as above and Pearson and Spearman correlations with "True" utilities for the three methods in case of a randomly chosen sign for the interactions are shown in the Appendix.

Conclusions:

- It is important to take into account semantic message interactions to accurately optimize message bundles.
- Using MaxDiff for individual messages to optimize message combinations might work fine under conditions of no interactions or weak interactions; but its conclusions might be erroneous when non-trivial interactions are present.
- Methods that elicit reactions to message pairs and model preferences for pairs directly (e.g., MaxDiff with message pairs) should be preferred for MBO purposes when interactions are reasonable to assume.
- TURFs (ratings-based or MaxDiff-based) are not an adequate solution for MBO purposes.

Dimitri Liakhovitski

Faina Shmulyian

Tatiana Koudinova

APPENDIX

Message Bundles for Holdout Tasks 1–3:

Holdout Task 1			
Bundle 1	Bundle 2	Bundle 3	Bundle 4
Helps you control your pain	Gets you back on track before anyone knows you hurt	Relieves your pain so you quickly feel like yourself again	Effectively relieves pain within 15 minutes
Was developed by pain experts and is endorsed by doctors	Is the #1 choice for pain	Contains the medicine prescribed most for tough pain	Relieves your pain so you can get back to your life
Works through the night so you can get a good night's sleep	Just 2 doses provides pain relief for a full 24 hours	Powerful pain relief that lasts longer than any other pain reliever	The longest-lasting pain reliever available without a prescription
Holdout Task 2			
Bundle 1	Bundle 2	Bundle 3	Bundle 4
Gets you back on track before anyone knows you hurt	No other pain reliever relieves pain faster	Effectively relieves pain within 15 minutes	Relieves your pain so you quickly feel like yourself again
Is the most effective pain relief you can buy	Confronts your pain, so you can get on with your day	Helps you control your pain	Is the #1 choice for pain
Powerful pain relief that lasts longer than any other pain reliever	Used most by doctors for fast, all day relief of tough pain	Just 2 doses provides pain relief for a full 24 hours	The longest-lasting pain reliever available without a prescription
Holdout Task 3			
Bundle 1	Bundle 2	Bundle 3	Bundle 4
Relieves your pain so you can get back to your life	Effectively relieves pain within 15 minutes	Gets you back on track before anyone knows you hurt	No other pain reliever relieves pain faster
Used most by doctors for fast, all day relief of tough pain	Is the #1 choice for pain	Confronts your pain, so you can get on with your day	Is the most effective pain relief you can buy
Works through the night so you can get a good night's sleep	Powerful pain relief that lasts longer than any other pain reliever	Was developed by pain experts and is endorsed by doctors	Contains the medicine prescribed most for tough pain

Message Bundles for Holdout Tasks 4–6:

<table>
<tr><th colspan="4">Holdout Task 4</th></tr>
<tr><td>Bundle 1</td><td>Bundle 2</td><td>Bundle 3</td><td>Bundle 4</td></tr>
<tr><td>Gets you back on track before anyone knows you hurt</td><td>Effectively relieves pain within 15 minutes</td><td>No other pain reliever relieves pain faster</td><td>Relieves your pain so you quickly feel like yourself again</td></tr>
<tr><td>Relieves your pain so you can get back to your life</td><td>Confronts your pain, so you can get on with your day</td><td>Is the most effective pain relief you can buy</td><td>Helps you control your pain</td></tr>
<tr><td>Used most by doctors for fast, all day relief of tough pain</td><td>Is the #1 choice for pain</td><td>Contains the medicine prescribed most for tough pain</td><td>Was developed by pain experts and is endorsed by doctors</td></tr>
<tr><td>The longest-lasting pain reliever available without a prescription</td><td>Just 2 doses provides pain relief for a full 24 hours</td><td>Works through the night so you can get a good night's sleep</td><td>Powerful pain relief that lasts longer than any other pain reliever</td></tr>
<tr><th colspan="4">Holdout Task 5</th></tr>
<tr><td>Bundle 1</td><td>Bundle 2</td><td>Bundle 3</td><td>Bundle 4</td></tr>
<tr><td>No other pain reliever relieves pain faster</td><td>Relieves your pain so you quickly feel like yourself again</td><td>Gets you back on track before anyone knows you hurt</td><td>Effectively relieves pain within 15 minutes</td></tr>
<tr><td>Relieves your pain so you can get back to your life</td><td>Confronts your pain, so you can get on with your day</td><td>Is the most effective pain relief you can buy</td><td>Helps you control your pain</td></tr>
<tr><td>Is the #1 choice for pain</td><td>Used most by doctors for fast, all day relief of tough pain</td><td>Was developed by pain experts and is endorsed by doctors</td><td>Contains the medicine prescribed most for tough pain</td></tr>
<tr><td>Powerful pain relief that lasts longer than any other pain reliever</td><td>Works through the night so you can get a good night's sleep</td><td>Just 2 doses provides pain relief for a full 24 hours</td><td>The longest-lasting pain reliever available without a prescription</td></tr>
<tr><th colspan="4">Holdout Task 6</th></tr>
<tr><td>Bundle 1</td><td>Bundle 2</td><td>Bundle 3</td><td>Bundle 4</td></tr>
<tr><td>Gets you back on track before anyone knows you hurt</td><td>No other pain reliever relieves pain faster</td><td>Effectively relieves pain within 15 minutes</td><td>Relieves your pain so you quickly feel like yourself again</td></tr>
<tr><td>Confronts your pain, so you can get on with your day</td><td>Helps you control your pain</td><td>Relieves your pain so you can get back to your life</td><td>Is the most effective pain relief you can buy</td></tr>
<tr><td>Contains the medicine prescribed most for tough pain</td><td>Used most by doctors for fast, all day relief of tough pain</td><td>Was developed by pain experts and is endorsed by doctors</td><td>Is the #1 choice for pain</td></tr>
<tr><td>Powerful pain relief that lasts longer than any other pain reliever</td><td>Just 2 doses provides pain relief for a full 24 hours</td><td>Works through the night so you can get a good night's sleep</td><td>The longest-lasting pain reliever available without a prescription</td></tr>
</table>

Thurstone Scores by Study Cell, by Holdout Task:

Study Cell	N	Holdout Task	Bundle1	Bundle2	Bundle3	Bundle4
1 Anchored MD	201	1	-0.29	-0.21	0.11	0.39
		2	-0.17	-0.13	0.25	0.05
		3	0.01	0.35	-0.35	-0.01
		4	-0.23	0.28	0.12	-0.17
		5	-0.03	-0.09	-0.19	0.31
		6	-0.23	0.01	0.22	0.00
2 Category DCM	204	1	-0.15	-0.24	0.10	0.29
		2	-0.17	-0.07	0.19	0.05
		3	0.03	0.24	-0.34	0.07
		4	-0.11	0.06	0.11	-0.07
		5	0.04	0.01	-0.27	0.22
		6	-0.18	-0.03	0.16	0.06
3 Ratings	203	1	-0.15	-0.18	-0.03	0.36
		2	-0.26	-0.09	0.31	0.04
		3	-0.10	0.39	-0.25	-0.04
		4	-0.16	0.26	0.02	-0.12
		5	-0.01	-0.05	-0.14	0.19
		6	-0.29	0.06	0.17	0.06
4 DCM	209	1	-0.12	-0.13	-0.03	0.28
		2	-0.28	-0.11	0.25	0.14
		3	0.12	0.34	-0.31	-0.14
		4	-0.18	0.20	0.07	-0.09
		5	-0.05	0.05	-0.18	0.17
		6	-0.31	0.00	0.36	-0.05
5 Traditional MD	201	1	-0.18	-0.25	0.18	0.25
		2	-0.12	-0.05	0.16	0.02
		3	-0.01	0.32	-0.39	0.08
		4	-0.20	0.26	0.11	-0.17
		5	0.06	-0.11	-0.12	0.17
		6	-0.20	-0.01	0.16	0.05

Exhibit: *Spearman* Correlations between Mean "True" Utilities for 66 Pairs and Metrics in the 3 Methods with Sign of the Interactions Manually Selected

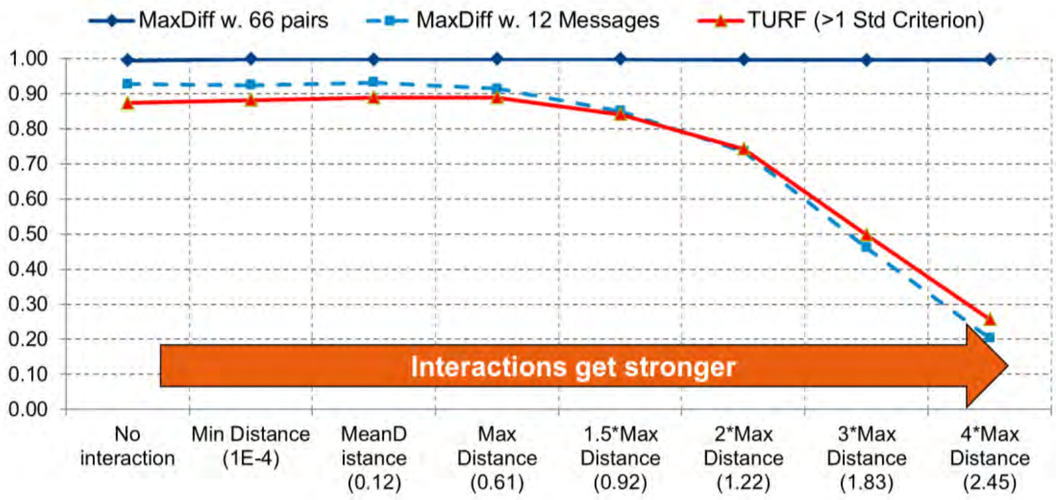


Exhibit: *Spearman* Correlations between Mean "True" Utilities for 66 Pairs and Metrics in the 3 Methods with Randomly Selected Sign of the Interactions

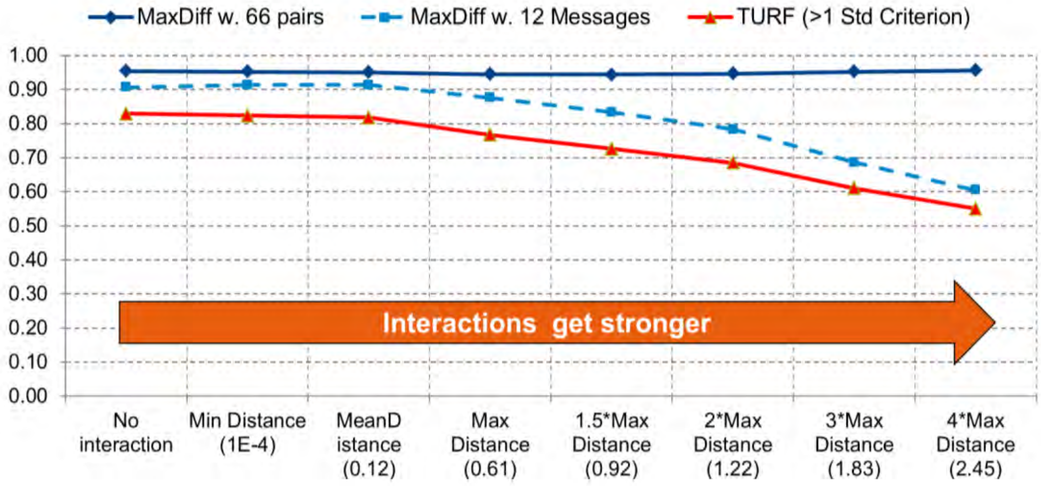
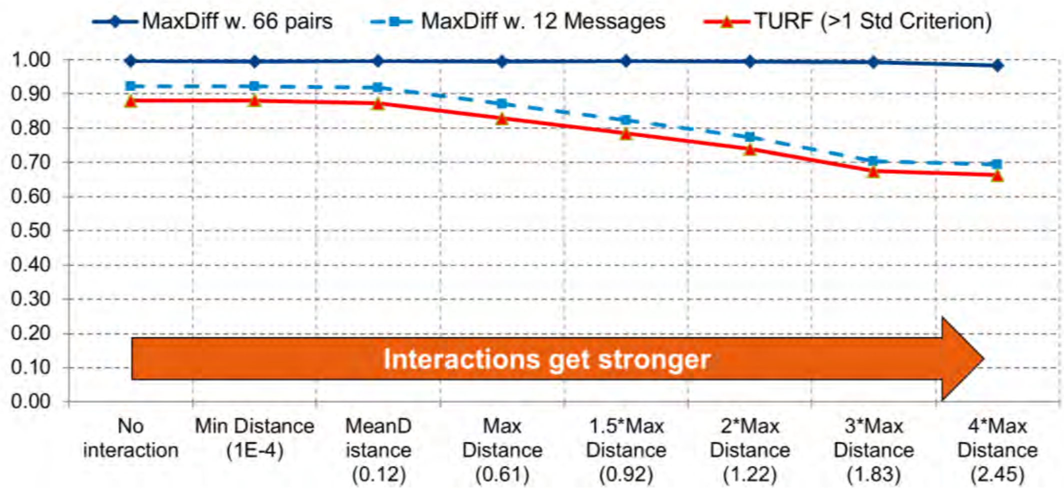


Exhibit: *Pearson* Correlations between Mean “True” Utilities for 66 Pairs and Metrics in the 3 Methods with Randomly Selected Sign of the Interactions.



USING TURF ANALYSIS TO OPTIMIZE REWARD PORTFOLIOS

PAUL JOHNSON
KYLE GRIFFIN
SURVEY SAMPLING INTERNATIONAL

ABSTRACT

Within the online market research space, creating and maintaining high engagement levels for survey communities is of paramount importance; however, keeping people active and engaged has never been more difficult. This is very evident in that many online communities experience high churn rates among their members. While people decide to leave online communities for many different reasons, the importance of incentives can't be overlooked. For instance, Leverage Saliency Theory suggests that providing meaningful rewards can improve cooperation rates. Therefore, panel companies are faced with a need to understand how they can optimize reward offerings to maximize value to their members.

An optimized rewards portfolio can build loyalty and improve activity rates. Reward programs can be used to leverage the brand recognition of popular retailers and link it to panel communities in the hope that the two become inseparable in the minds of the community members. We explore using total unduplicated reach and frequency (TURF) analysis to come up with an optimal rewards portfolio. In addition to a chip allocation question to fuel the TURF analysis we explore using anchored-scaled MaxDiff and a self explicated model to inform the TURF analysis.

INTRODUCTION

Survey Sampling International (SSI) operates online market research communities in thirty-five countries around the world. As a panel company, our largest asset is our panel members; therefore, keeping panel members engaged and active in our communities is critical. Our experience has shown that this important task is very difficult to do well. There are a multitude of reasons why survey participants become disengaged and no longer participate in market research (Callegaro & DiSogra, 2008). Whatever the reason, replacing good people becomes one of the top expenses for the company. For example, last year SSI replaced over one million panelists in the United States alone.

Leverage Salience Theory indicates that rewards are critical to successful attraction and retention of people taking surveys (Groves et al., 2000). Therefore, incentive strategy and management has a very important role in the success or failure of a panel company. Providing targeted incentives to panelists in thirty-five countries requires international cooperation and local knowledge. There are thousands of potential rewards that can be offered to meet the expectations of people all over the world. Therefore, there needs to be a data driven process for understanding what the ideal rewards portfolio looks like in each country to maximize value to panel community members while also being financially advantageous to the panel company.

SSI has previously conducted testing of new rewards options using conjoint analysis in order to optimize the amount and types of rewards offered at the survey level (Fawson & Johnson, 2009). This research project isn't about the survey level rewards, but rather the composition of reward options available to redeem points they have collected. Our research has the primary goal of discovering if TURF analysis can be used to provide the necessary data for making correct decisions concerning the optimal allocation of our global rewards portfolio. In each case, we also know the cost structure of the portfolio items so we can optimize not just for reach but also for cost savings. Like in the 2009 study, we then compare the stated preference to the actual preference we have seen through the panel redemptions.

BACKGROUND

Our research will focus primarily on two of our panel communities: Opinion Outpost United States and Opinion World Argentina. We have chosen these two communities in order to see how TURF analysis can be used in vastly differing market environments.

Opinion Outpost United States

The United States is a very developed market for panel research. Opinion Outpost US is our flagship panel and has over 525,000 members. This panel produces over five million survey completes each year resulting in over 450,000 rewards fulfillment transactions. There are currently four primary rewards offered as part of this panel community. Each reward provider is offered in the local currency and panelists' contentment with the portfolio is very good. One way to gauge the success of the rewards offering is by performing research-on-research within our panel. For example, we have seen that 85% of all incentives earned by panelists on Opinion Outpost are redeemed. This is a very good indicator of panel satisfaction with the rewards offered. Additionally, the incentive market in the United States is very competitive. Rewards providers are often willing to offer volume based discounts to panel communities to have their products included inside the rewards portfolio.

OpinionWorld Argentina

Argentina is an emerging market for panel research. OpinionWorld Argentina currently has 40,000 members and produces fewer than 100,000 completes per year. Our current rewards portfolio only includes a single provider with eleven redemption options resulting in less than 1,500 fulfillment transactions per year. These options are not in local currency and panelist contentment is low. For example, only 8% of earned incentives are redeemed for these options. The incentive market in Argentina is not well established and volume discounts are not available.

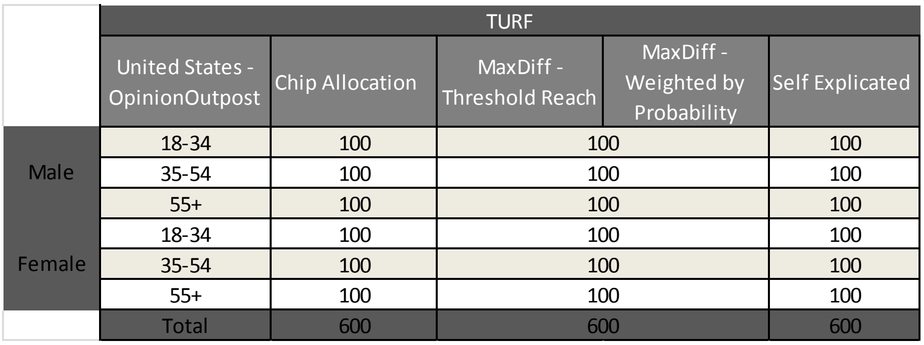
Over the past few years, we have identified twenty-two potential reward options for each market. We will use these twenty-two options to drive our TURF analysis. Our primary goal in the United States is to maximize our panelist satisfaction level (reach) while also maximizing the volume discounts for managing the portfolio. Our primary goal in Argentina is to discover the ideal portfolio for maximizing panelist satisfaction level with the least amount of options possible.

RESEARCH DESIGN/METHODOLOGY

TURF analysis was originally used for media planning (Zufryden, 1975), but has since been adapted to optimize product line extensions (Miaoulis et al., 1990). More recently, Sawtooth Software published a paper on how TURF analysis could be repurposed with MaxDiff utilities to get a proportion weight for reach rather than a binary result for reach (Howell, 2012). There has also been increased use of a technique Jordan Louviere used to anchor the MaxDiff utilities around a threshold which would be a natural input for a TURF analysis (Orme, 2010). We propose to compare four different implementations of TURF analysis to gauge success of our reward portfolio optimization. The four methods we will use are: Chip Allocation, MaxDiff–Threshold Reach, MaxDiff–Weighted by Probability, and Self Explicated.

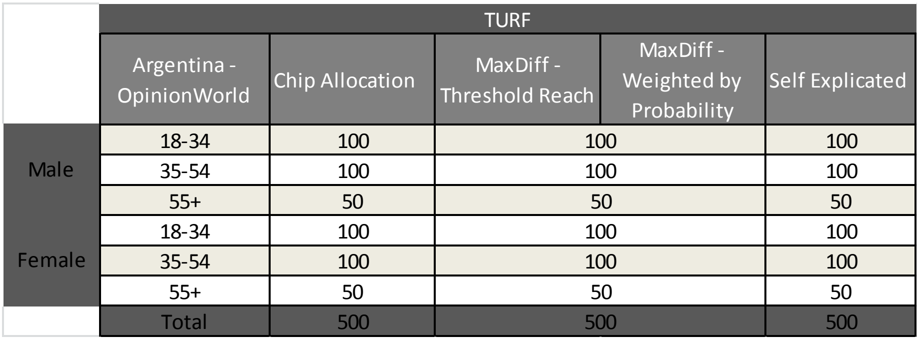
The total sample size in each treatment of the research in the United States was 600 while in Argentina it was 500 (Figures 1 & 2). The sample was distributed across age and gender groups evenly with the exception that the Argentina panel can't support older quota groups.

Figure 1. Design Quotas for the United States Test



	TURF				
	United States - OpinionOutpost	Chip Allocation	MaxDiff - Threshold Reach	MaxDiff - Weighted by Probability	Self Explicated
Male	18-34	100	100		100
	35-54	100	100		100
	55+	100	100		100
Female	18-34	100	100		100
	35-54	100	100		100
	55+	100	100		100
	Total	600	600		600

Figure 2. Design Quotas for the Argentina Test

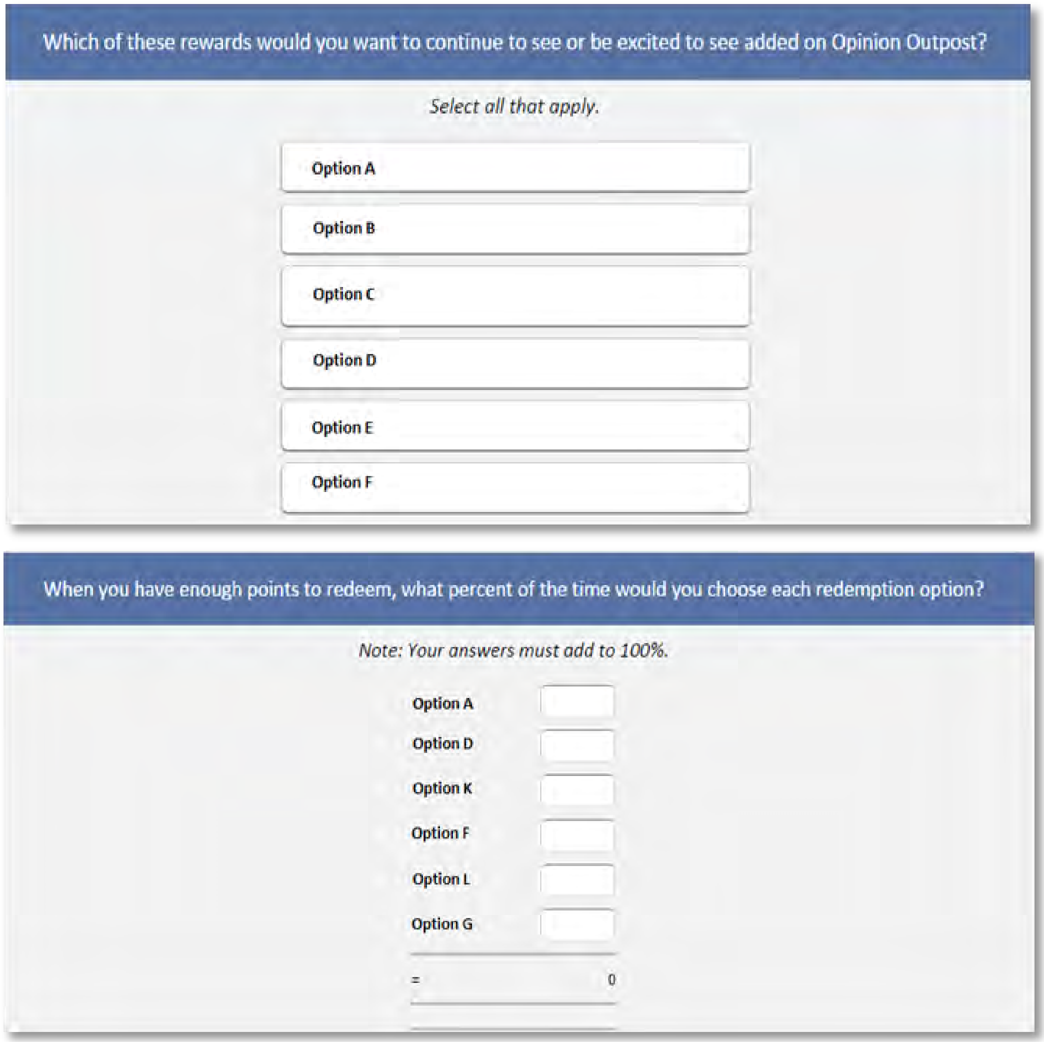


	TURF				
	Argentina - OpinionWorld	Chip Allocation	MaxDiff - Threshold Reach	MaxDiff - Weighted by Probability	Self Explicated
Male	18-34	100	100		100
	35-54	100	100		100
	55+	50	50		50
Female	18-34	100	100		100
	35-54	100	100		100
	55+	50	50		50
	Total	500	500		500

Chip Allocation

This method starts with a multi-select question. Then, using data from the multi-select, we ask respondents to allocate their points among the options they would use to redeem (Figure 3). The multi-select is used to measure reach by each product while the allocation question is used to then estimate volume.

Figure 3. Visual of Chip Allocation Process



Which of these rewards would you want to continue to see or be excited to see added on Opinion Outpost?

Select all that apply.

Option A

Option B

Option C

Option D

Option E

Option F

When you have enough points to redeem, what percent of the time would you choose each redemption option?

Note: Your answers must add to 100%.

Option A

Option D

Option K

Option F

Option L

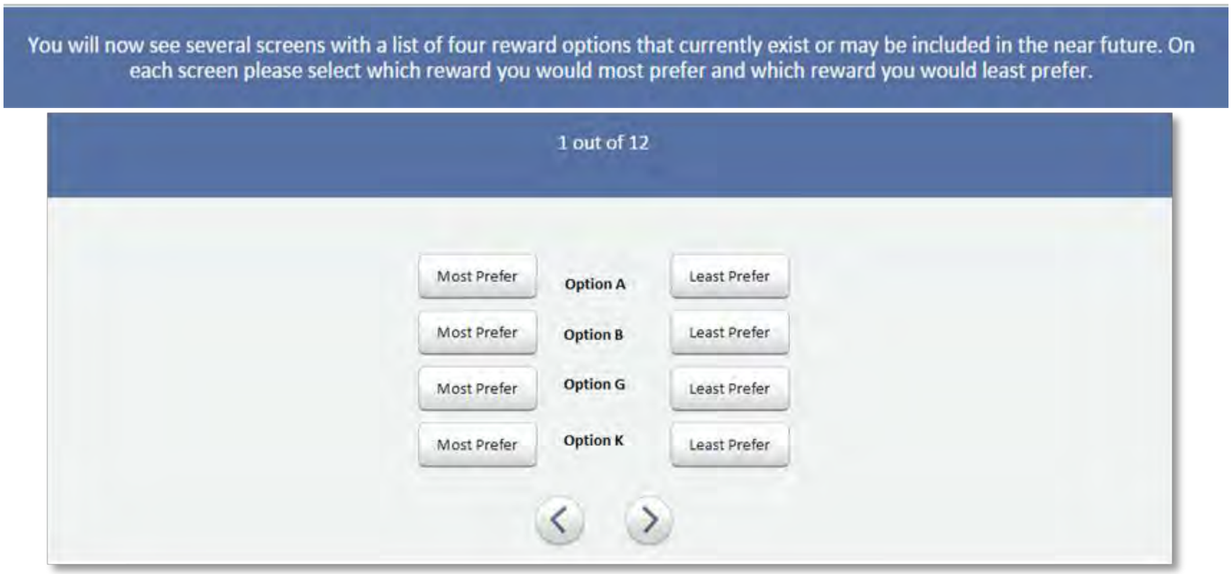
Option G

= 0

MaxDiff–Threshold Reach & MaxDiff–Weighted by Probability

Methods two and three use an Anchored-Scaled MaxDiff with the Direct Binary method developed by Kevin Lattery (Lattery, 2011). The MaxDiff question is the standard one and the multi-select question at the end was used for the threshold estimation (Figure 4). In the Threshold Reach Analysis, we also utilize a chip allocation question to estimate volume while the MaxDiff question is used to estimate reach. In the Weighted by Probability Method, we ignored the chip allocation question and just used the utility scores weighted by probability to get volume estimates.

Figure 4. Visual of Anchored-Scale MaxDiff Process



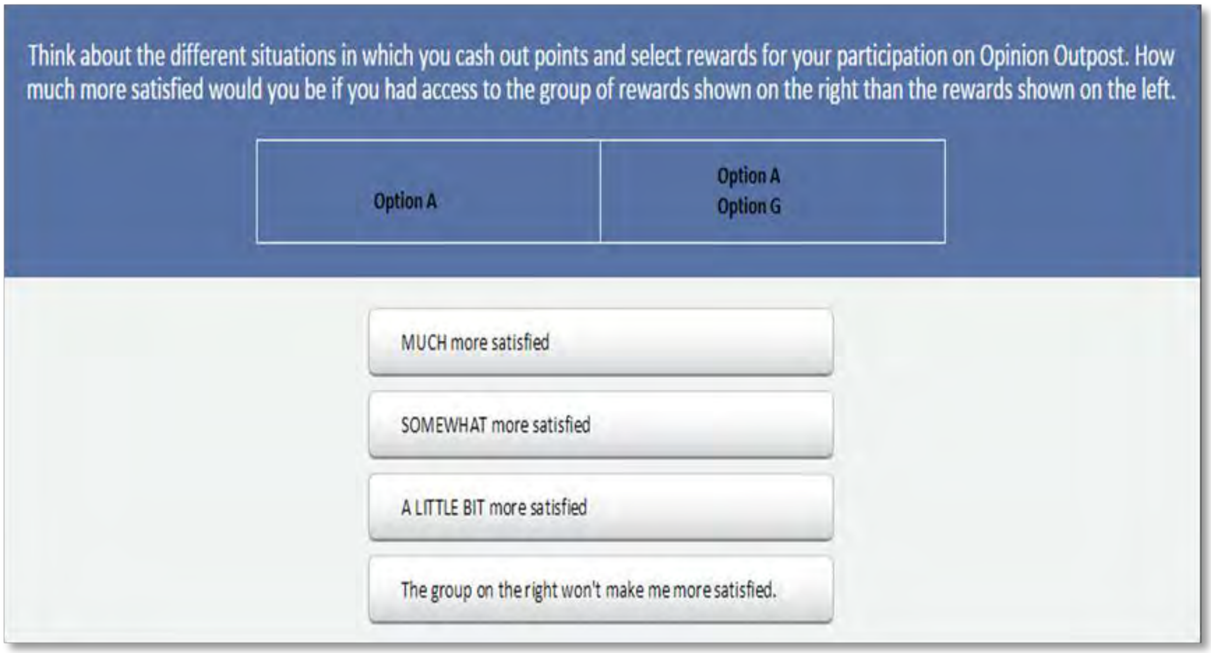
You will now see several screens with a list of four reward options that currently exist or may be included in the near future. On each screen please select which reward you would most prefer and which reward you would least prefer.

1 out of 12

Self-Explicated

The fourth method was suggested to us by Bryan Orme. It is called Self-Explicated because it mimics a build-your-own approach where panelists start with the top two options and compare the benefit of adding additional options to not adding additional options (Figure 5). If the respondent is more satisfied with the portfolio with more options, we loop back to ask the next most exciting option and add it to the portfolio until they are equally satisfied with the one on the right and on the left. Once again, we use a chip allocation question to get volume estimates on the resulting options selected.

Figure 5. Visual of Anchored-Scale MaxDiff Process



Think about the different situations in which you cash out points and select rewards for your participation on Opinion Outpost. How much more satisfied would you be if you had access to the group of rewards shown on the right than the rewards shown on the left.

Option A	Option A Option G

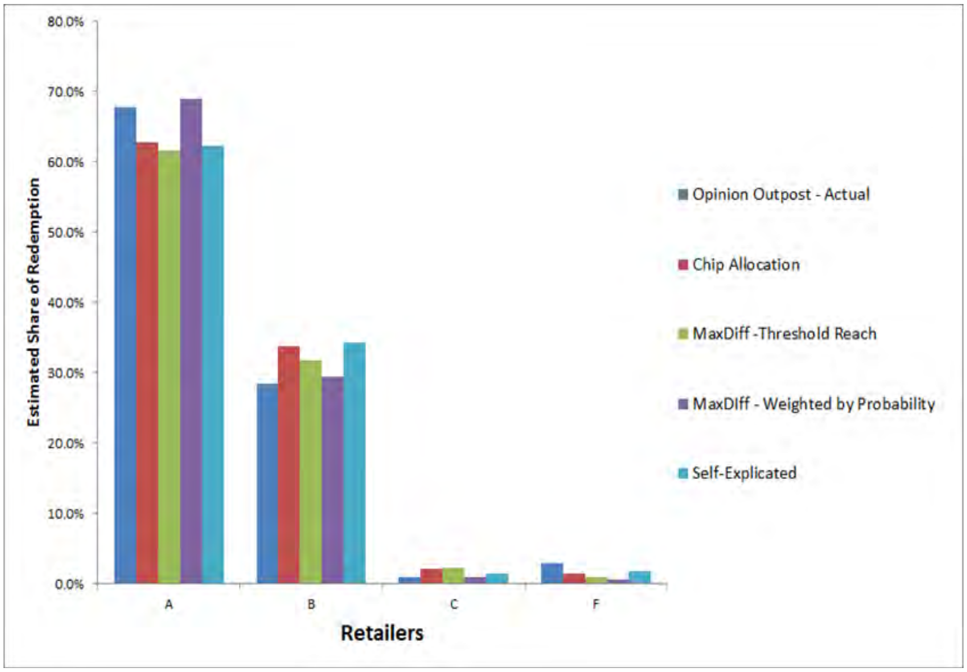
MUCH more satisfied
SOMEWHAT more satisfied
A LITTLE BIT more satisfied
The group on the right won't make me more satisfied.

Research Results

Testing Model Accuracy & Stated Versus Actual Preference

Unfortunately the portfolio provider in Argentina was never able to deliver on the reward options they said they could, so we have no actual data in the developing country. We were able to compare the model accuracy in the United States though. Opinion Outpost currently has four retailers. We used those four retailers in our TURF analysis to see which method best accurately predicted actual behavior (Figure 6).

Figure 6. Estimates by Model Compared to Actual Behavior with Current Portfolio

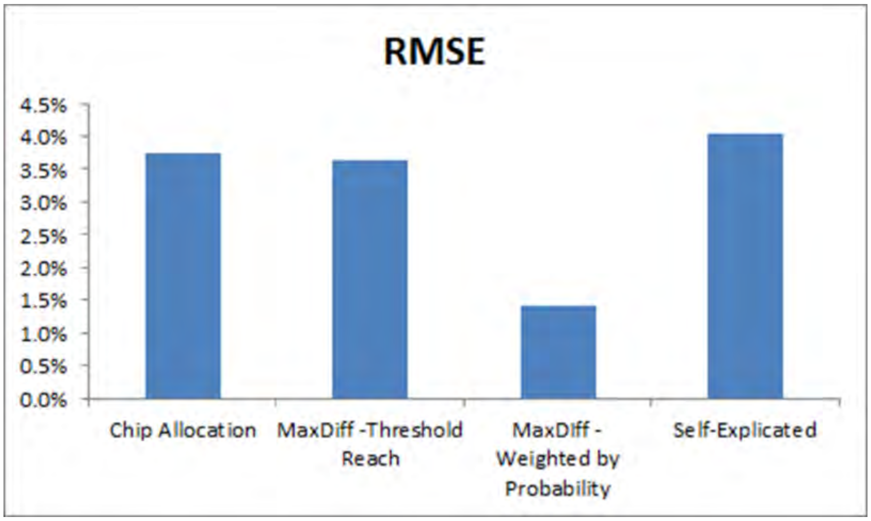


We were very pleased to see that each of the TURF models was very similar to the actual usage recorded on the panel. The consistency across models increases our trust that TURF models can actually provide reasonable forecasts on future behavior. That being said, it is very important to understand the reasons that our modeling turned out to be so accurate. We had the following advantages:

1. We limited the actual data to just those specific survey respondents who took any of the treatment surveys. Other panelists were excluded.
2. Our sample size was robust with 600 respondents for each treatment in the United States.
3. We were able to aggregate 12 months of actual redemption history for each of the participants. This stabilizes the actual data closer to the true long-term average.
4. We weighted our survey data by the redemption volume history of each panelist so those that redeemed more counted more.

You can also see that the redemption portfolio is very much dominated by two reward options (A&B). While all the models predicted well, using the MaxDiff utilities to predict volume shares did significantly outperform the other models which used a chip allocation question for volume estimates. The Root Mean Squared Error (RMSE) for this model was less than half of the other models (Figure 7).

Figure 7. Root Mean Squared Error by Model Predicting Actual Redemption Volume

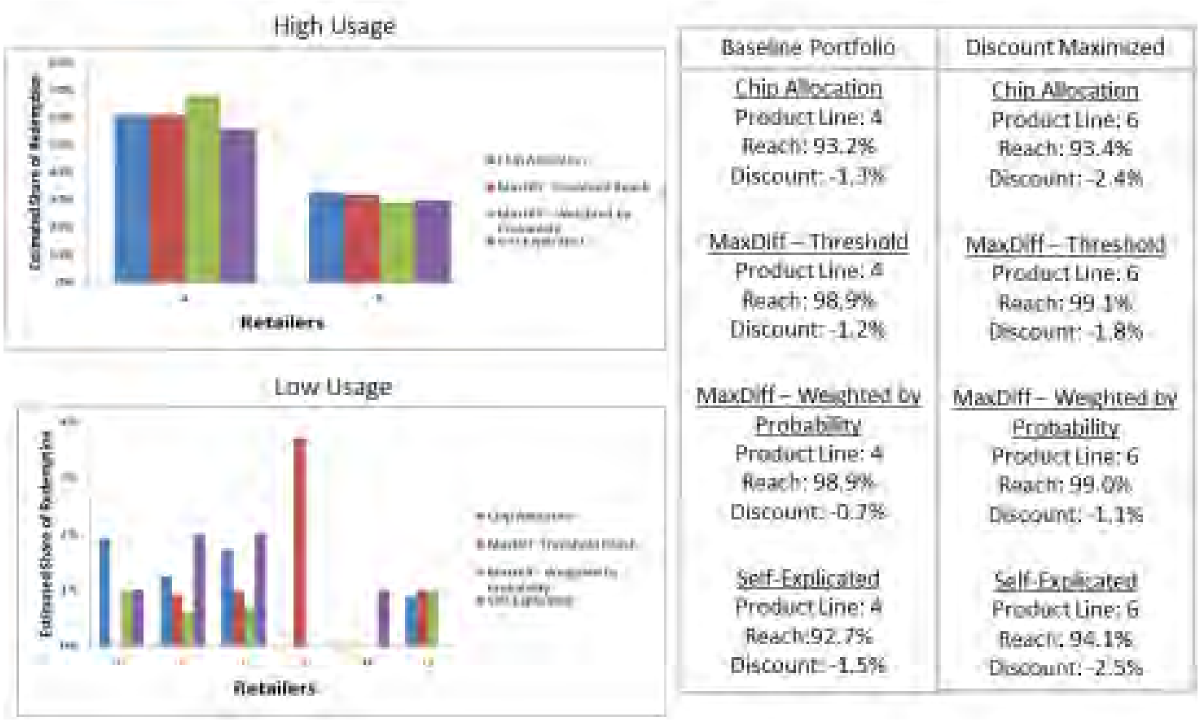


Based on these factors, our results support the theory that, in this instance, stated behavior isn't very different from actual behavior. Moreover, we were able to compare the Root Mean Square Error of each treatment to define a clear winner for accuracy in predicting actual behaviors. The MaxDiff–Weighted by Probability treatment is the clear winner for our purposes. Based on these results, SSI will use this treatment to make business decisions moving forward.

Optimizing Cost Structure in United States:

The primary goal for the United States is to maximize our panelist satisfaction level (reach) while also minimizing costs for managing the portfolio. In particular retailer A (the most popular one) charges us a premium for using them. We would like to shift work away from retailer A to other retailers that give us more of a discount in operating the redemption portfolio. Our research was able to identify the ideal portfolio to do this. Despite identifying the MaxDiff–Weighted by Probability as the best model to follow, we will include the results for all four treatments to better understand the variation provided by each model. The results are summarized in Figure 8 below.

Figure 8. Example of TURF Simulation for Cost Reduction

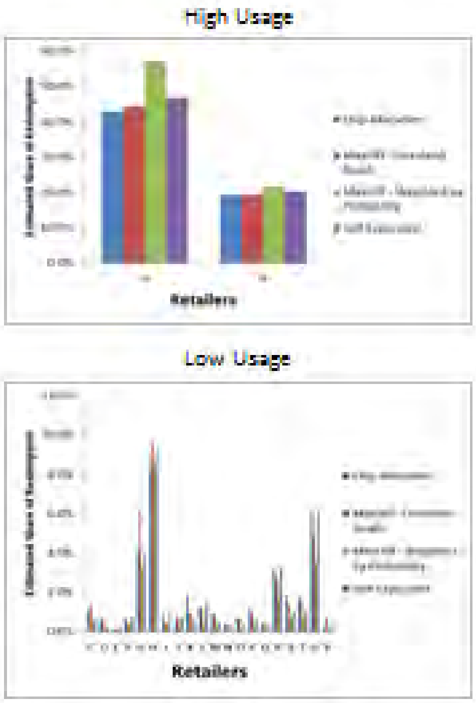


Baseline Portfolio	Discount Maximized
Chip Allocation Product Line: 4 Reach: 93.2% Discount: -1.3%	Chip Allocation Product Line: 6 Reach: 93.4% Discount: -2.4%
MaxDiff – Threshold Product Line: 4 Reach: 98.9% Discount: -1.2%	MaxDiff – Threshold Product Line: 6 Reach: 99.1% Discount: -1.8%
MaxDiff – Weighted by Probability Product Line: 4 Reach: 98.9% Discount: -0.7%	MaxDiff – Weighted by Probability Product Line: 6 Reach: 99.0% Discount: -1.1%
Self-Explicated Product Line: 4 Reach:92.7% Discount: -1.5%	Self-Explicated Product Line: 6 Reach: 94.1% Discount: -2.5%

Unsurprisingly, the four methods agreed that retaining retailers A & B would continue to be important for the success of the panel. Interestingly, the models provided different results concerning which other reward options would be best to include in order to maximize the discount. This first modelling iteration estimates that we will able to achieve .5% to 1% in cost reductions to managing our portfolio if we expanded the portfolio to add two additional options. Our MaxDiff–Weighted by Probability treatment helped us to identify two retailers (I and Q) that will be the best addition to our portfolio in the United States. The addition of these two retailers will not have a substantive impact on the overall reach of the rewards portfolio, but they will allow SSI to maximize our potential discount by an additional 0.4%.

This product line of 6 items was much more effective than blindly implementing all reward options possible (Figure 9). The issue became that when we added these other retailers they actually took volume away from retailer B instead of retailer A. Because retailer B has a significant volume discount, adding in these retailers would increase costs rather than reduce costs of overall portfolio. Implementing all possible retailers would result in at least a 0.5% decrease in discounts achieved from the rewards portfolio. That doesn’t take into account all the IT and legal resources needed to implement even one new retailer.

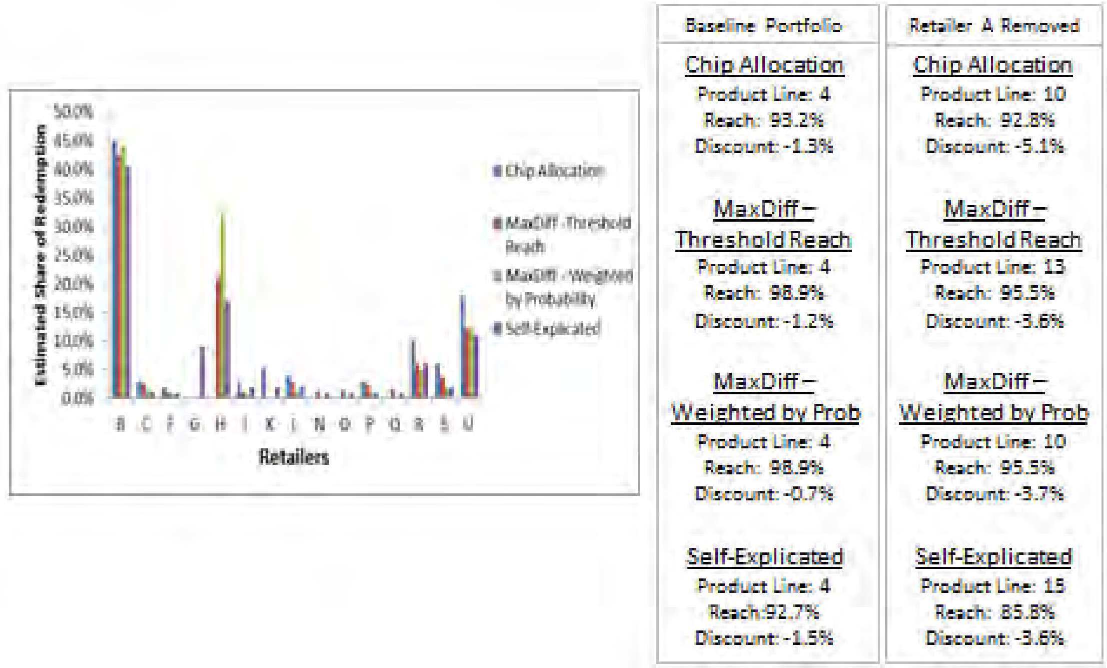
Figure 9. Example of TURF Simulation for Complete Portfolio Implementation



Discount Maximized	Total Implementation
<u>Chip Allocation</u> Product Line: 6 Reach: 93.4% Discount: -2.4%	<u>Chip Allocation</u> Product Line: 22 Reach: 98.7% Discount: -1.2%
<u>MaxDiff – Threshold Reach</u> Product Line: 6 Reach: 99.1% Discount: -1.8%	<u>MaxDiff – Threshold Reach</u> Product Line: 22 Reach: 99.4% Discount: -1.1%
<u>MaxDiff – Weighted by Probability</u> Product Line: 6 Reach: 99.0% Discount: -1.1%	<u>MaxDiff – Weighted by Probability</u> Product Line: 22 Reach: 99.4% Discount: -0.6%
<u>Self-Explicated</u> Product Line: 6 Reach: 94.1% Discount: -2.5%	<u>Self-Explicated</u> Product Line: 22 Reach: 99.8% Discount: -1.4%

Lastly, we looked at a model to replace the expensive retailer A from the portfolio while still maintaining satisfaction in the panel. Retailer A is extremely popular and averages over 65% of yearly redemptions volume, so there is a lot of ground to make up with new rewards providers if we removed it as an option. The model showed that there was a massive decrease in our reach as a result of removing this option. Decreased reach can have serious financial implications for our business; therefore, we modelled how many potential retailers we would need to add to come close to previous reach estimates. The data provided was very interesting in that it shows that removal of retailer A would probably not be a very good idea (Figure 10). While removal of this option does increase our discount, we would need to triple our product line in order to obtain even close to the same reach.

Figure 10. Example of TURF Simulation for Replacing Retailer A



Baseline Portfolio	Retailer A Removed
Chip Allocation Product Line: 4 Reach: 93.2% Discount: -1.3%	Chip Allocation Product Line: 10 Reach: 92.8% Discount: -5.1%
MaxDiff – Threshold Reach Product Line: 4 Reach: 98.9% Discount: -1.2%	MaxDiff – Threshold Reach Product Line: 13 Reach: 95.5% Discount: -3.6%
MaxDiff – Weighted by Prob Product Line: 4 Reach: 98.9% Discount: -0.7%	MaxDiff – Weighted by Prob Product Line: 10 Reach: 95.5% Discount: -3.7%
Self-Explicated Product Line: 4 Reach:92.7% Discount: -1.5%	Self-Explicated Product Line: 15 Reach: 85.8% Discount: -3.6%

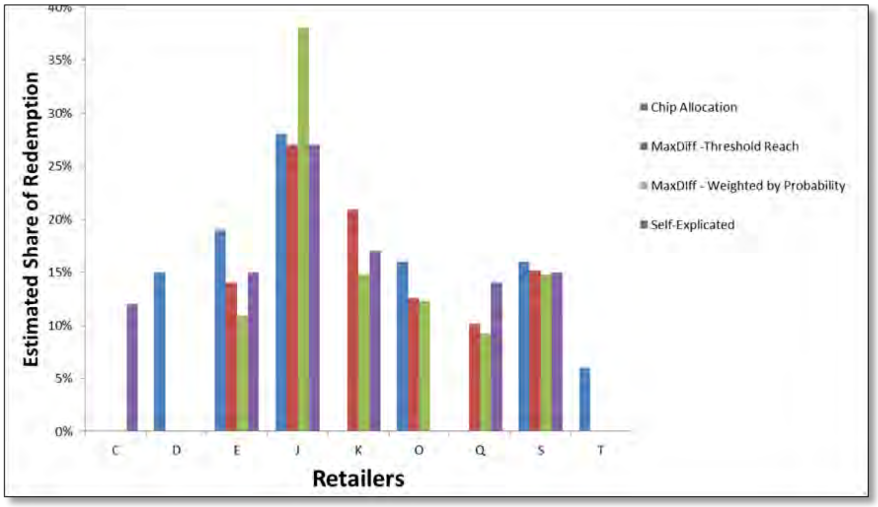
In the end, the TURF model was very useful in helping us make intelligent business decisions around what we could do to optimize costs while retaining satisfaction of our panelists.

Optimizing Portfolio Reach in Argentina:

The goal in Argentina was completely different. The cost structure didn't matter as much as reaching people in the first place and making them happy with the reward options provided. We were able to identify the optimal portfolio to launch in order to satisfy the needs of about 95% of our panelists. We found that 6 options, instead of all 22, was ideal for maximizing reach (Figure 11). In this model, there was more agreement in which retailers should be used to obtain the desired reach. Implementation of the six retailers identified by the MaxDiff–Weighted by Probability treatment is currently in progress. Once these retailers are activated inside the portfolio, SSI will be able to have another real-life data set to confirm the accuracy of our TURF modelling.

We also ran an additional model variation to test the potential impact of all the potential retailers to the new Argentina portfolio (Figure 12). This variation provided some very interesting information for us to base our business decisions on. We can clearly see that, by adding in the additional sixteen retailers, the expected impact on our reach is only 1.7%. The potential increase in our reach does not outweigh the cost of adding in sixteen additional retailers. This additional data confirms the decision to keep the new rewards portfolio in Argentina close to six retailers.

Figure 11. Example of TURF Simulator for Optimal Reach with Limited Products



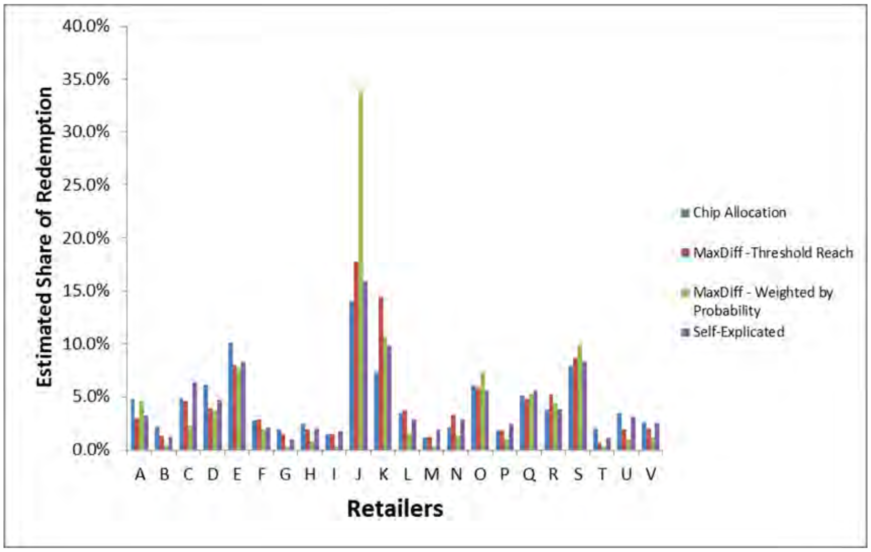
Chip Allocation
Product Line: 6
Reach: 96.0%

MaxDiff – Threshold Reach
Product Line: 6
Reach: 95.7%

MaxDiff – Weighted by Probability
Product Line: 6
Reach: 95.7%

Self-Explicated
Product Line: 6
Reach:94.5%

Figure 12. Example of TURF Simulation for Complete Portfolio Implementation



Chip Allocation
Product Line: 22
Reach: 98.2%

MaxDiff - Threshold Reach
Product Line: 22
Reach: 97.4%

MaxDiff - Weighted by Probability
Product Line: 22
Reach: 97.4%

Self-Explicated
Product Line: 22
Reach: 96.4%

CONCLUSION

Application of TURF analysis has been very successful for SSI. It has helped us to understand the rewards requirements in drastically different markets. Our research shows that MaxDiff–Weighted by Probability is, in our situation, the best model for understanding predictive behavior. Using this model, SSI has been able to quantitatively prove that the retailer portfolio of our potential new supplier in Argentina appeals to our panelists and is worth pursuing. The data was also able to provide excellent insight into the top six retailers that would maximize the utility for our panelists in Argentina.

TURF modelling has also provided direction for resource planning and allocation in the United States. Our modelling was able to identify the top two retailers, out of a pool of eighteen, which would maximize the management discounts achieved in the portfolio while also improving our reach. Increasing discounts on the reward portfolio by roughly one percent can result in upwards of $80,000 in savings for the company each year. Moreover, improving reach by roughly 1–2% will also have a direct financial implication on the business as this will directly impact the churn rates experienced on the panel.

Paul Johnson

Kyle Griffin

WORKS CITED

Callegaro M. & DiSogra, C. (2008). Computing Response Metrics for Online Panels. *Public Opinion Quarterly*, 1008–1032.

Fawson, B. & Johnson, E. (2009). Collaborative Panel Management: The Stated and Actual Preference of Incentive Structure. *Proceedings of 2009 Sawtooth Software Conference*, 113–122.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly*, 299–308.

Howell, J. (2012). A Simple Introduction to TURF Analysis. *Sawtooth Software Research Paper Series*.

Lattery, K. (2011). Anchoring Maximum Difference Scaling Against a Threshold—Dual Response and Direct Binary. *Sawtooth Software Research Paper Series.*

Miaoulis, G., Parsons, H., & Free, V. (1990). Turf: A New Planning Approach for Product Line Extensions. *Marketing Research*, 2(1).

Orme, B. (2010). Anchored Scaling in MaxDiff Using Dual Response. *Sawtooth Software Research Paper Series.*

Zufryden, F. (1975). On the dual optimization of media reach and frequency. *Journal of Business.* Vol 48. (4) pp.558–570.

BANDIT ADAPTIVE MAXDIFF DESIGNS FOR HUGE NUMBER OF ITEMS

KENNETH FAIRCHILD
BRYAN ORME
SAWTOOTH SOFTWARE, INC.
ERIC SCHWARTZ
UNIVERSITY OF MICHIGAN

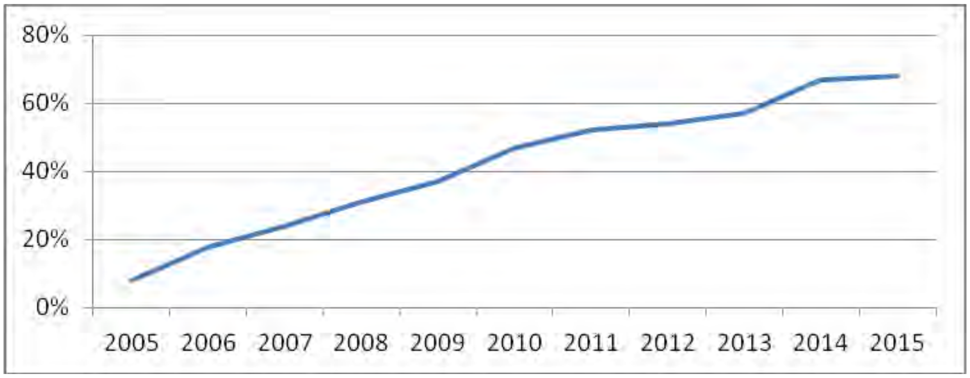
EXECUTIVE SUMMARY

For large MaxDiff studies whose main purpose is identifying the top few items for the sample, a new adaptive approach called Bandit MaxDiff may increase efficiency fourfold over standard non-adaptive MaxDiff. Bandit MaxDiff leverages information from previous respondents via aggregate logit and Thompson Sampling so later respondents receive designs that oversample the topmost items that are most likely to turn out to be the overall winners.

BACKGROUND

MaxDiff (Maximum Difference Scaling) is now a popular general item scaling method in our industry. MaxDiff (also known as best-worst scaling) was developed by Jordan Louviere in the late 1980s and first released as a software system in 2004 by Sawtooth Software. Sawtooth Software has tracked MaxDiff usage among its users since then, with penetration of the technique now reaching 68% (Figure 1).

Figure 1
Use of MaxDiff Technique among Sawtooth Software User Firms



MaxDiff provides much more discrimination among items and between respondents on the items than traditional rating scales (Cohen and Orme 2004). Besides enhanced discrimination, it avoids the scale use bias so problematic with traditional ratings scales. Intuitively, MaxDiff may be thought of as a one-attribute CBC (Choice-Based Conjoint) study with many levels. MaxDiff is not only excellent for quantifying importance or preference for an array of items, but also for conducting market segmentation via latent class or cluster algorithms.

THE DRIVE TO STUDY MORE ITEMS

MaxDiff has proven so useful that increasingly it is being relied upon for studying very large numbers of items. How many is a large number of items?

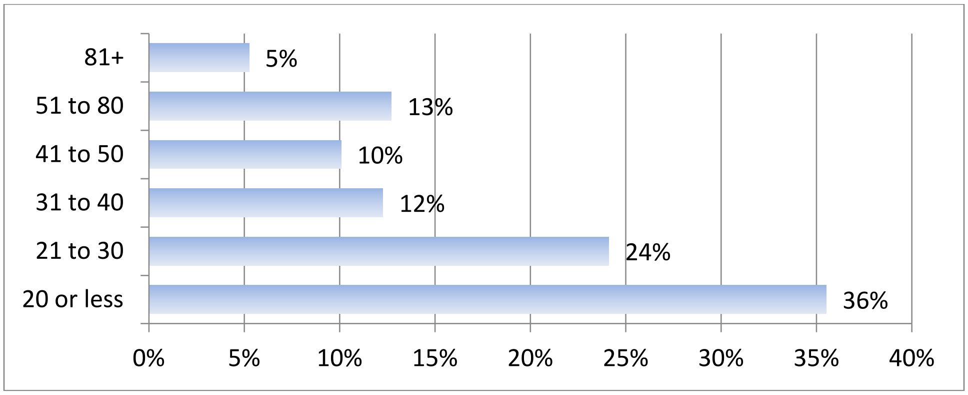
- In their 2007 paper, Hendrix and Drucker described "large sets" as about 40 to 60 items, proposing variants to MaxDiff called Augmented and Tailored MaxDiff to handle such large problems (Hendrix and Drucker, 2007).
- In their 2012 paper, Wirth and Wolfrath also investigated variants to MaxDiff called Express and Sparse MaxDiff for handling what they described as "very large sets" of items (Wirth and Wolfrath, 2012). Very large to these authors meant potentially more than 100 items. To support their findings, they conducted a study among synthetic robotic respondents with 120 items and a real study among humans with 60 items.

For Hendrix and Drucker 40 to 60 items was *large*, for Wirth and Wolfrath 120 items was *very large*. For this current paper, we're referring to *huge numbers of items* as potentially 300 or more. Indeed, it seems like an arms race to devise better MaxDiff methodologies for studying the largest number of items! More than just for academic curiosity, client demand justifies these investigations, as we're regularly asked to push MaxDiff further than it was perhaps ever intended.

The problem is that current MaxDiff approaches don't scale well to increasing the number of items. More items requires commensurately longer questionnaires, larger sample sizes, and commensurately larger data collection costs with more tired respondents. If the researcher is concerned about obtaining robust *individual-level* estimates for *all the items*, then the current methodologies especially don't scale well to large lists of items. Respondents just tire out! In contrast, our approach employs an adaptive divide-and-conquer aggregate approach that leverages prior learning to create more efficient questionnaires and more precise aggregate score estimates.

In Sawtooth Software's 2015 Customer Feedback Survey, we asked respondents to tell us the largest number of items they had included in a MaxDiff study during the last 12 months (Figure 2). Nearly one-fifth of respondents indicated their firms had conducted a study with 51 or more items. The maximum number of items studied was 400!

Figure 2
Maximum Number of Items Studied via MaxDiff over Last 12 Months

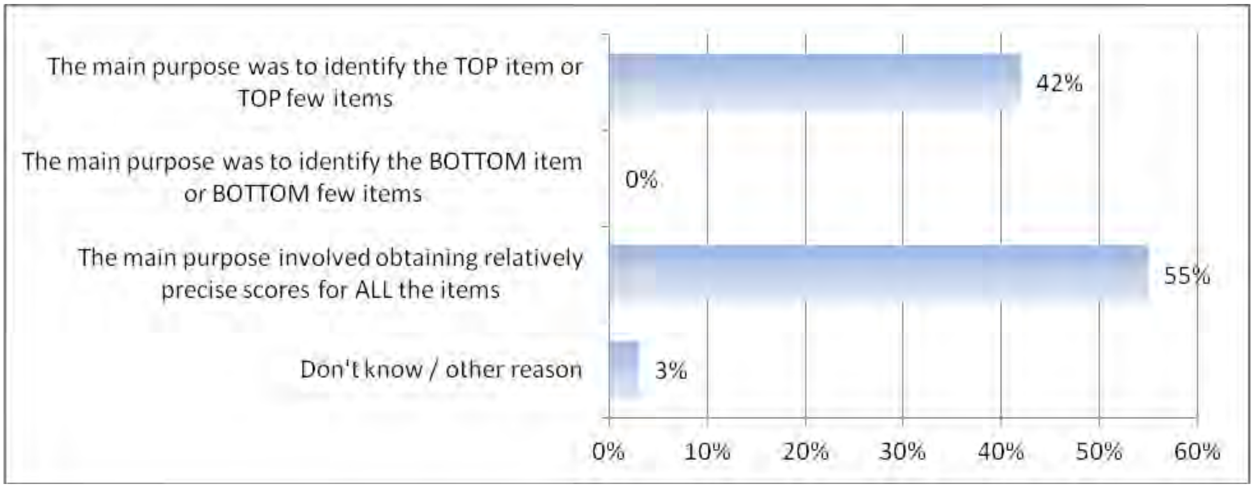


Mean= 40, Median=30, Maximum=400

To some it may seem bizarre and overwhelming that some researchers are conducting MaxDiff studies with 81+ or even 400 items! However, when we consider that individual MaxDiff items may actually represent conjoined elements that constitute a profile (say, a combination of packaging style, color, claims, and highlighted ingredients), then it can make much more sense to do 400-item studies. If the profiles involve multiple highly interactive attributes that pose challenges for CBC, then MaxDiff with huge numbers of items could be a viable alternative (given the new approach we demonstrate further below).

We also asked Sawtooth Software customers what the *main purpose* was for that study with the reported maximum number of items. For studies involving 41 or more items, the main reasons are displayed in Figure 3.

Figure 3
Main Purpose for MaxDiff Study with 41+ Items



For 42% of these large MaxDiff studies, the main purpose was to identify the TOP item or TOP few items. Our research shows that if this is the main goal, then *traditional design strategies are very wasteful*. An adaptive approach using Thompson Sampling can be about 4x more efficient. Without the Thompson Sampling approach, you are potentially wasting 75 cents of every dollar you are spending on MaxDiff data collection.

Multi-Armed Bandit Problems and Thompson Sampling

Thompson Sampling has been proposed as an efficient solution for solving the *Multi-Armed Bandit Problem*[3], wherein the "player" seeks to maximize the cumulative rewards of investments in different gambling "games" that have uncertain outcomes. Thompson Sampling involves allocating resources to an action in proportion to the probability that it is the best action (Thompson 1933). Any bandit method must find an appropriate balance between exploring to gain information and exploiting that knowledge. On the one hand, we want to learn about the relative scores of a large number of items within a MaxDiff problem. On the other hand, we want to utilize what we have learned so far to focus our efforts on a targeted set of actions that will likely yield greater precision regarding the items of most interest to the researcher. While there are many methods to accomplish this, Thompson Sampling has proven very useful for these types of problems.

Conveniently, for MaxDiff, the probability of an item being most preferred for a group of respondents can be estimated using aggregate logit[4]. Moreover, the standard error of each logit weight characterizes the uncertainty surrounding each estimate. For a marketing application of Thompson Sampling and a review of the literature, see Schwartz et al. (2013).

The traditional MaxDiff design approach shows each item an equal number of times across all respondents x tasks. However, if the main goal is to identify the top few items for the sample, after the first 50 or so respondents it seems reasonable to start paying attention to the already-collected MaxDiff responses and oversampling the items that are already viewed as most preferred (the *stars*). We can use aggregate logit to estimate (usually in a few seconds) both preference scores and standard errors at any point during data collection (say, after the 50th, 60th, 70th, etc. respondent has completed the survey).

Thompson Sampling makes a new draw from the vector of item preferences using the estimated population preferences (aggregate logit scores) plus normally distributed error, with standard deviations equal to the standard errors of the logit weights. As the sample size increases, the standard errors of course tighten.

Imagine after 50 respondents we decide to summarize their preferences (for each of 100+ items) with aggregate logit. Then, to generate a MaxDiff task for the 51st respondent, we could generate a draw from the population preferences leveraging the population means and normal

[3] In the USA, a common slang term for a slot machine for gambling is the "one-armed bandit." The machine has an arm—actually a lever—you pull and it tends to steal your money like a bandit. If you faced the analytic problem of investing your money across different slot machines each involving different uncertain outcomes, then this becomes a "Multi-Armed Bandit Problem."

[4] Due to the sparse nature of MaxDiff for huge numbers of items plus the desire for rapid real time updates, we decided to use aggregate MNL rather than a Bayesian approach.

errors with standard deviations equal to the empirically estimated standard errors. We then can sort that newly sampled vector of preference scores from the most preferred item to the least preferred item. The five most preferred items might be taken into the first task to show to the 51st respondent. The process (with or without updating the logit weights after recording the first task's answer) could be repeated to choose the five items to show in the second task for the 51st respondent, etc. To reduce the load on the server managing the data collection, perhaps only after every 10th respondents has completed the survey, the logit weights and standard errors would be updated.

A practical issue to overcome with Thompson Sampling is as the sample size grows, items that are most preferred by the population achieve high preference scores with smaller standard errors. Without any additional restrictions, the same few items will eventually tend to be drawn into adjacent MaxDiff tasks for the same respondent, causing much annoyance due to the severe degree of item repetition. Although this is statistically most efficient, it would drive human respondents mad. To avoid this, we use Thompson Sampling to draw a fixed number of items (we've experimented with 20 or 30) to show each respondent. Those draws of, say, 30 items are shown to each respondent in a balanced, near-orthogonal design, leading to a palatably low degree of repetition of items across adjacent sets. The attentive reader will notice that our approach is quite similar to Wirth's Express MaxDiff approach, except that the logic for selecting the 30 items for each respondent is adaptive, using Thompson Sampling, leveraging information from the previous respondents—focusing the most recent respondent's efforts on discriminating among items that already have been judged likely to be the stars.

Simulation Results

Using the R programming language, we compared non-adaptive MaxDiff design strategies to the Thompson Sampling approach using robotic respondents created to mimic human behavior as closely as we were able. To begin with, we used actual MaxDiff data from human respondents donated by our friends at Procter & Gamble (the subject matter and item text was hidden for confidentiality purposes). The study involved 984 respondents and 120 items (from a sparse MaxDiff study that asked a lot of MaxDiff tasks of each respondent). Only the HB utilities were shared with us. Those HB scores offered realistic patterns of preferences across the items and respondents for use in our robotic respondent simulations. Our robotic respondents simply mimicked the humans' preferences to answer each new MaxDiff task, according to an assigned human respondent's true utilities perturbed by Gumbel error. For each sample of robotic respondents, we ran aggregate logit and compared the rank order of the estimated pooled item scores to the unchanging rank order for the known true utilities. To stabilize the hit rate results (since there was a random component to the responses), we ran the simulations each 100s of times.

We used our robotic respondents to test how accurately we could recover the top few items as observed in the true preferences for the Procter & Gamble dataset. We used two measures of success:

- **Top 3 hit rate:** what percent of the top 3 true items the estimated scores using robotic respondents identified. Example: if the estimated scores identified 2 of the true top 3 items (irrespective of order), the hit rate was 66.67%.

- **Top 10 hit rate:** what percent of the top 10 true items the estimated scores using robotic respondents identified. Example: if the estimated scores identified 7 of the true top 10 items (irrespective of order), the hit rate was 70%.

With 120 items in the dataset, it shouldn't surprise us if the true preferences for the top 15 or so items were very close in terms of utility. We certainly observed that with this data set. Due to how tightly the top items in preferences clustered (there were no runaway winners), the hit rate measures we employed were quite discriminating between competing methods.

Using bootstrap sampling (sampling with replacement), we simulated the process of collecting respondent data up to sample sizes of n=1020. Each robotic respondent completed 18 choice sets where each set included 5 items, which was viewed as fairly typical of larger MaxDiff studies in practice.

The MaxDiff approaches we tested were:

- **Sparse MaxDiff:** we showed each item to each respondent an equal number of times (if possible). With 120 items, 18 sets, and 5 items per set each item appeared on average 18*5/120 = 0.75 times per respondent.

- **Express MaxDiff:** we randomly drew 30 of the 120 items to show to each respondent. Each item appeared 18*5/30 = 3 times per respondent. Across respondents, each item appeared the same number of times.

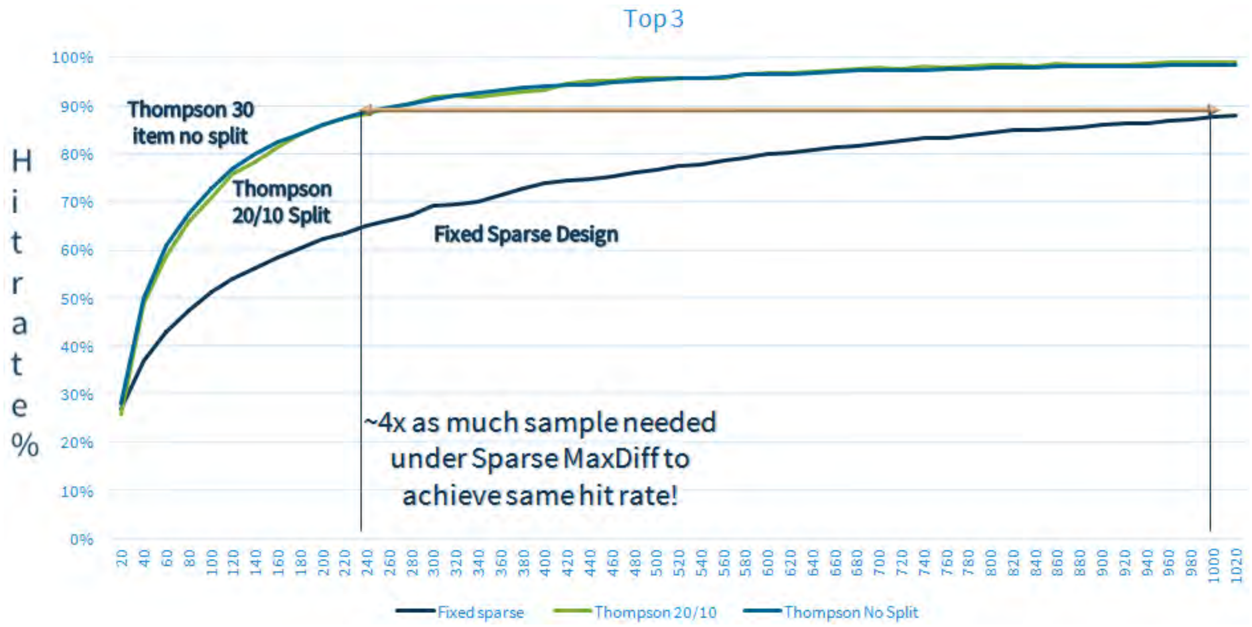
- **Bandit MaxDiff:** we used Thompson Sampling to draw 30 of the 120 items for each respondent (tending to oversample the "stars" based on aggregate logit estimates from previous respondents). We tested two different Bandit Maxdiff approaches:

    - 30 items drawn via standard Thompson Sampling.

    - 20/10 split: 20 of the items drawn using standard Thompson Sampling; 10 of the items drawn using Thompson Sampling with a much more diffuse prior (standard errors multiplied by 10).

Figure 4 shows the results for the 120-item data set. The X-axis indicates the number of cumulative respondents interviewed and the Y-axis reports the hit rate obtained at each cumulative sample size. For example, after the first 140 respondents, the Fixed Sparse Design obtains a hit rate of less than 60% whereas both Bandit MaxDiff approaches achieve a hit rate of about 80%.

Figure 4

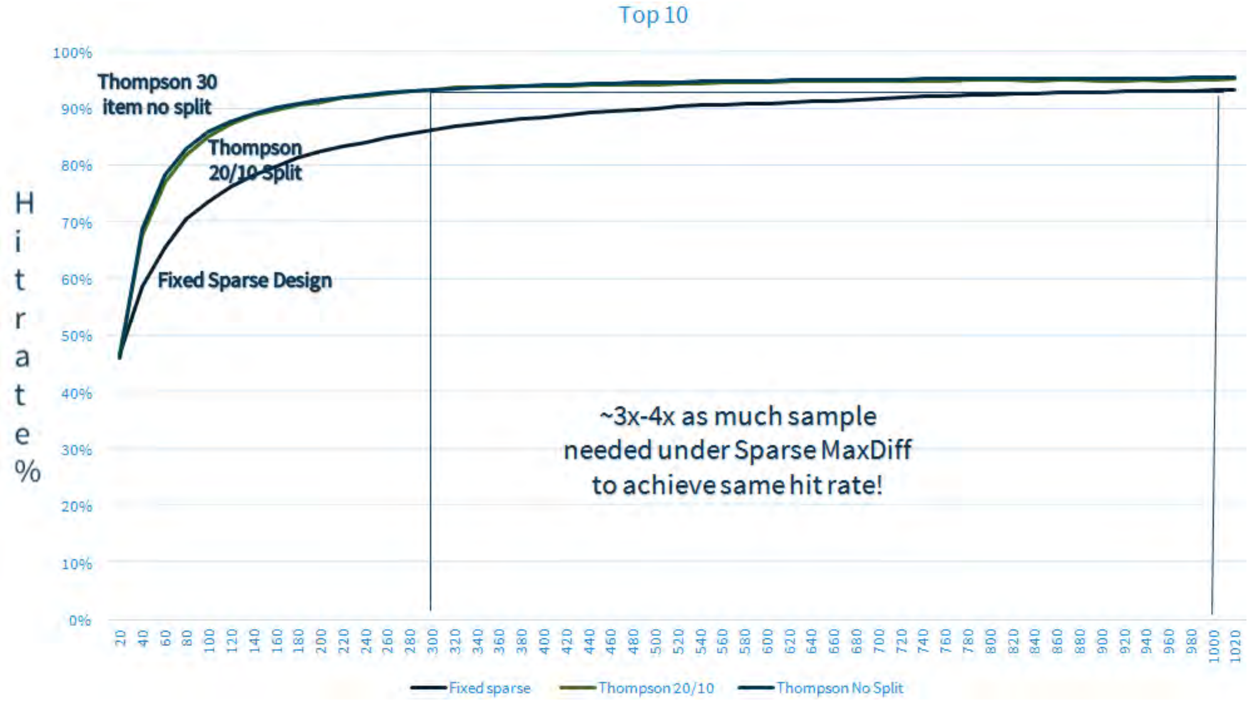


The key takeaways from Figure 4 are as follows:

1. The two Thompson Sampling approaches achieve nearly identical results (but we'll show this isn't the case if we use a misinformed start—more on that later).
2. Thompson Sampling is about 4x more efficient than the standard Sparse MaxDiff approach. After about 140 respondents, we've obtained an 80% hit rate; which we wouldn't achieve with traditional Sparse MaxDiff until about the 600th respondent. As shown, we can accomplish with about 240 respondents what it takes 1000 respondents to do with Sparse MaxDiff. Either comparison shows that you can obtain equally good hit rates using the adaptive Bandit MaxDiff methodology with ¼ the sample size.

Figure 5 shows the results for Top-10 hit rate for our 120-item MaxDiff, which is a broader measure of success that requires obtaining a high degree of precision for an even broader reach of items than the Top-3 hit rate. The conclusions are fairly similar as with Figure 4. It takes about 300 respondents to accomplish with Bandit MaxDiff what we accomplish with 1000 respondents under the non-adaptive sparse MaxDiff approach.

Figure 5



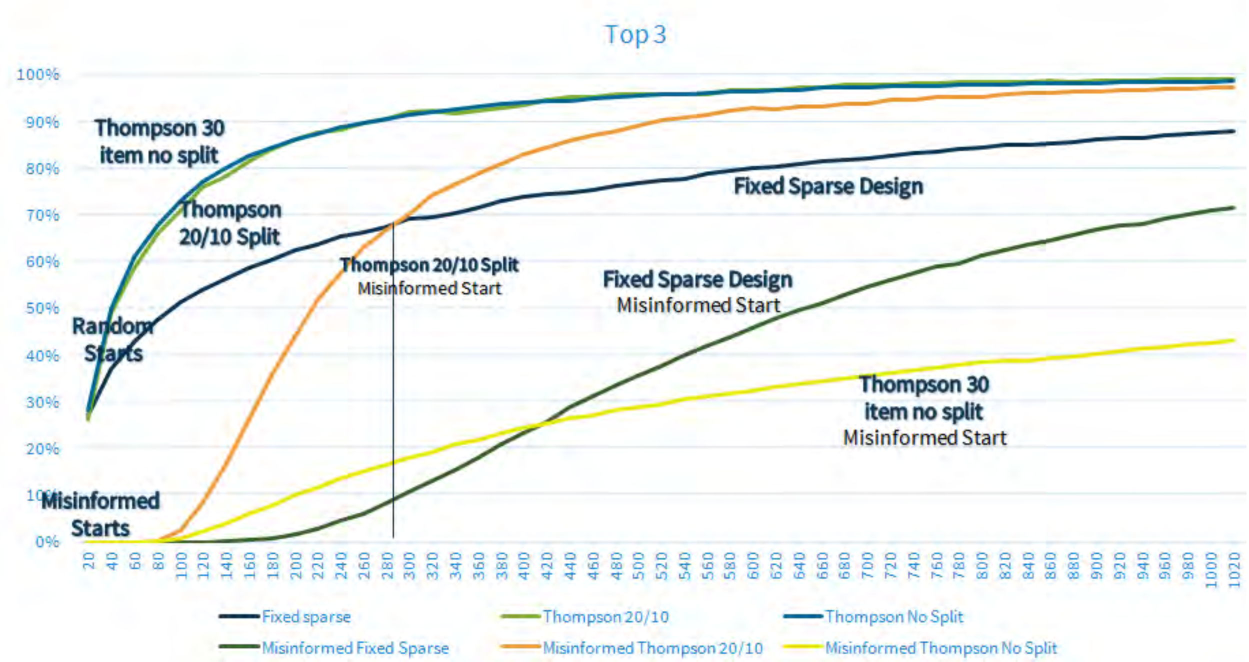
Misinformed Starts (When Early Responders Are Horribly Non-Representative)

To this point, it seems like the adaptive Bandit MaxDiff approach using Thompson Sampling is the clear winner. However, what would happen if the first 50 respondents we interviewed were actually not very representative of the average preferences for the sample? What if we tried to throw Bandit MaxDiff off the scent? In fact, let's consider a worst-case scenario: the first responders actually believe *nearly the opposite* from the rest of the sample!

For the simulations reported in Figure 6 below, the first 50 robotic respondents mimicked randomly drawn human vectors of utilities as before but were diabolically manipulated to behave as if the top 3 true items were actually nearly the worst in preference (we set the utilities for the top 3 true items for the population equal to the 25th percentile utility item for each respondent). After this misinformed start, the remaining respondents represented well-behaved respondents drawn using bootstrap sampling as before, with true individual-level preferences as given in the original dataset donated by Procter & Gamble.

We did such a diabolical thing as create 50 misinforming early responders because in the real world you are never guaranteed that the first responders represent a fair and representative draw from the population. In fact, depending on how rapidly you invite a panel of respondents to take the survey, the first 50 respondents may share some atypical characteristics (e.g., anxious and available to take the survey at your 1PM launch time). It would be a bad thing if the Thompson Sampling approach performed well in simulations with well-behaved respondents, but fell apart under conditions that were more realistic to the human world. In our opinion, our diabolical simulation is worse than anything you would realistically see in practice, so it is a good test of the robustness of the Bandit MaxDiff approach.

Figure 6



The key takeaways from Figure 6 are as follows:

1. The 20/10 split Bandit MaxDiff approach is much better in the face of misinformed starts than the standard Bandit MaxDiff approach. The more diffuse prior on the 10 items within the split allow us to continue investigating the value of some lesser chosen items with enough frequency among later respondents, even if the prior respondents seem to have generally rejected them.
2. Even in the face of a misinformed start, the 20/10 Bandit MaxDiff achieves equally good results as the standard Sparse MaxDiff without the misinformed start after 290 respondents.

The conclusions were very similar when examining Top-10 hit rate, so we save space by not displaying the results.

What about 300 Items?

Would the benefits of Bandit MaxDiff we observed with 120 items continue for 300 items? While we didn't have a data set of utilities from human respondents on 300 items, we did our best to generate such a data set by leveraging the 120-item data set Procter & Gamble shared with us. To generate preferences across an additional set of 180 items, we randomly combined pairs of existing items according to a randomly distributed weighting scheme, with additional random variation added. The result was a 300-item MaxDiff data set based on the original preferences of the 984 respondents.

Our results for both well-informed and misinformed starts were nearly identical to the 120-item results. The Bandit MaxDiff approach was again 4x more efficient than the standard Sparse MaxDiff approach on the Top-3 hit rate criterion.

WHAT ABOUT A SMALLER SET OF 40 ITEMS?

Our Bandit model has a great advantage over fixed designs for very large numbers of items, but what happens if we have a more traditional "large" MaxDiff list of 40 items. Using a random 40-item subset from our original set of 120 items, we reran our simulations. For this smaller subset, we also reduced the number of tasks per respondent to 12 and we drew 20 items per respondent using Thompson sampling rather than 30 items and used a 15/5 split between Thompson sampling and diffuse Thompson sampling.

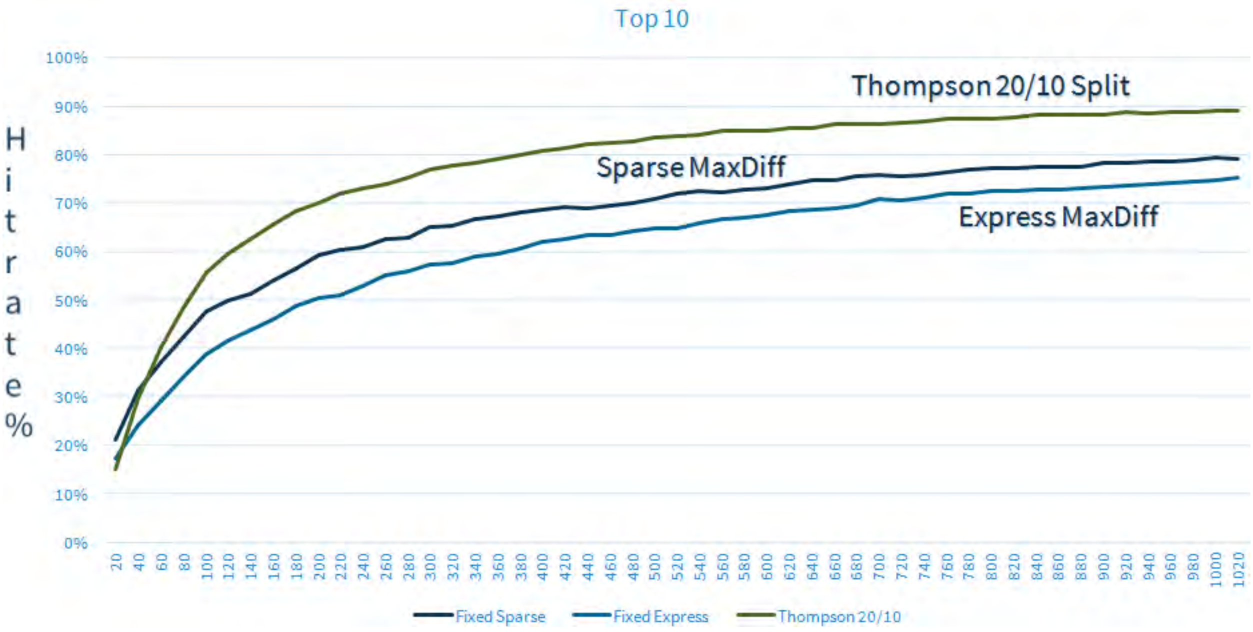
By reducing the number of items to 40 and changing the number of tasks to 12, the fixed sparse design can now show each item an average of 1.5 times per respondent, which is much less sparse than in larger item cases.

Our results for Bandit MaxDiff are still much better than traditional MaxDiff. We see about a 2x efficiency gain compared to the Sparse MaxDiff. We also still see a tremendous advantage in the case of a misinformed start.

WHAT ABOUT SPARSE MAXDIFF VS. EXPRESS MAXDIFF?

Wirth and Wolfrath compared non-adaptive Sparse MaxDiff and Express MaxDiff in their 2012 paper at the Sawtooth Software Conference (Wirth and Wolfrath 2012). We compared the results using our simulation and found a modest edge in performance for Sparse MaxDiff (Figure 7, Top-10 hit rate for 300 items).

Figure 7



This evidence in favor of Sparse MaxDiff is echoed in independent findings by Chrzan (Chrzan 2015).

What about Bests Only?

Because a key assumption for using the Bandit MaxDiff approach is that the researcher is mainly interested in identifying the top few items, we wondered about the value of spending time asking respondents to identify the worst item within each MaxDiff set. What would happen if we asked our robotic respondents only to select the best item within each set? The results somewhat surprised us. The value of asking respondents to indicate both best and worst within each set more than compensated for the 40% additional effort we suppose these "worst" questions add to the total interview time when interviewing human respondents.

In a five item set (A,B,C,D and E) there are 10 possible 2-way comparisons. If we assume A is preferred to B and B is preferred to C and so on, then asking about only the best item will let us know A>B, A>C, A>D and A>E (4/10 comparisons). By asking about worsts as well, for only one additional question we also add B>E, C>E, and D>E (7/10 comparisons), leaving only the order relationship between B, C, and D unknown.

What about Double Adaptivity?

In 2006, one of the authors presented a paper on Adaptive MaxDiff that featured within-respondent adaptation (Orme 2006) rather than what we have shown here in Bandit MaxDiff based on Thompson Sampling, which is an across-respondent adaptive approach. For the within-respondent adaptive procedure, items that a respondent indicates are worst are dropped from further consideration by that same respondent through a round-robin tournament until eventually that respondent's best item is identified. We thought adding this additional layer of within-respondent adaptivity on top of the Bandit MaxDiff approach could additionally lift its performance. To our surprise, this double-adaptive approach actually performed *worse* than Bandit MaxDiff alone in terms of hit rates for the globally best 3 or 10 items for the sample. After some head scratching (and much code checking), we determined that the lack of improvement was due to degree of heterogeneity across the robotic respondents. For example, if we are interviewing a respondent who doesn't agree much with the overall population regarding which are the top items, it is detrimental to allow that respondent to drop from further consideration (due to judging them worst) what actually are among the globally most preferred items. It serves the greater good for each respondent to spend increased effort judging among the items that previous respondents on average have judged as potentially best.

Conclusions and Future Research

Our results suggest that if your main purpose in using large item lists in MaxDiff is to identify the top items for the population (*not* individual-level estimates), then adaptive Bandit MaxDiff approaches can be 4x more efficient than standard Sparse MaxDiff designs. You are potentially wasting 75 cents of each dollar spent on data collection by not using Bandit MaxDiff.

Bandit MaxDiff leverages information from prior respondents to show more effective tradeoffs to later respondents (tending to oversample the stars, based on the Thompson Sampling mechanism). Even in the face of diabolically imposed misinformed starts (horribly unrepresentative first responders), the Bandit MaxDiff approach with our 20/10 split is extremely robust and self-correcting.

Although our simulations involve 120-item and 300-item tests, we expect that even greater efficiency gains than 4x (compared to standard Sparse MaxDiff designs) may occur with 500-item (or more) MaxDiff studies. For studies using 40 respondents, our simulation showed a 2x advantage in efficiency over fixed MaxDiff designs. Though not as dramatic, this is still a sizable boost.

Future research should test our findings using human respondents. Using an adaptive process that focuses on comparing best items may result in a more cognitively difficult task than a standard level-balanced, near-orthogonal approach. The greater expected within-set utility balance may lead to higher response error, which may counteract some of the benefits of the Bandit adaptive approach. However, based on previous research (Orme 2006) that employed within-respondent adaptivity, the additional degree of difficulty that the Bandit adaptive approach could impose upon individual respondents (owing to utility balance) would probably not counteract the lion share of the benefits we've demonstrated using simulated respondents.

We should note that as of this article's publication date, Sawtooth Software does not offer Bandit MaxDiff as a commercial tool. Sawtooth Software may perhaps one day soon offer Bandit MaxDiff as an option within its commercially available MaxDiff software. As for the authors, we look forward to this possibility as we've been especially impressed by the potential cost savings and increased accuracy.

Kenneth Fairchild

Bryan Orme

Eric Schwartz

References:

Chrzan, Keith (2015), "A Parameter Recovery Experiment for Two Methods of MaxDiff for Many Items," Sawtooth Software Research Paper, available at: http://www.sawtoothsoftware.com/support/technical-papers.

Cohen, Steve and Bryan Orme (2004), "What's Your Preference?" Marketing Research, 16 (Summer 2004), 32–37.

Hendrix, Phil and Stuart Drucker (2007), "Alternative Approaches to MaxDiff with Large Sets of Disparate Items—Augmented and Tailored MaxDiff," 2007 Sawtooth Software Conference Proceedings, PP. 169–188.

Orme, Bryan (2006), "Adaptive Maximum Difference Scaling," Sawtooth Software Research Paper, available at www.sawtoothsoftware.com/support/technical-papers.

Schwartz, Eric M., Eric T. Bradlow, and Peter S. Fader (2013), "Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments," Ross School of Business Paper No. 1217, available at: http://ssrn.com/abstract=2368523.

Thompson, Walter R. (1933), “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” *Biometrika*, 25(3) 285–294.

Wirth, Ralph and Anette Wolfrath (2012), “Using MaxDiff for Evaluating Very Large Sets of Items,” 2012 Sawtooth Software Conference Proceedings, pp. 59–78.

WHAT IS THE RIGHT SIZE FOR MY MAXDIFF STUDY?

STAN LIPOVETSKY[1]
DIMITRI LIAKHOVITSKI
MICHAEL CONKLIN
GFK NORTH AMERICA

ABSTRACT

MaxDiff studies are ubiquitous in Market Research, and the clients are frequently pushing for more items, fewer tasks, smaller sample sizes *N* and more data cuts. But how should a practitioner judge what *N* for a MaxDiff study would ensure that preference estimates from the sample are *representative of the preferences in the population*? MaxDiff utilities are usually estimated using HB, and the statistical properties of the distribution of MaxDiff utilities are unknown so it is hard to derive an a priori analytical rule of thumb. "Each respondent sees each item 3 times" rule of thumb is independent of *N*; "Simulate the MaxDiff responses and assess the standard errors from aggregate logit" is not currently offered by Sawtooth Software for MaxDiff studies, so it requires several manual steps. The aim of this study consists in helping marketing scientists to determine a desired sample size *N* based on: MOE—desired Margin of Error, *α*—level of significance (1 - Confidence Probability), *n*—total number of items in MaxDiff, *t*—total number of tasks in the design, *m*—number of items per task for each respondent (so $k = tm/n$ is the number of times each respondent sees each item). We propose analytical derivation, check it with massive computer simulations on the estimation of the needed sample size depending on the given parameters, and obtain "rules of thumb" convenient for reliable estimations.

I. INTRODUCTION

MaxDiff (Best-Worst Scaling) method is widely used in Marketing Research in general and at GfK in particular. Every year GfK fields over 100 studies that contain at least one MaxDiff each. As for any other statistical technique, an important issue for MaxDiff is estimation of the sample size needed. Market researchers in charge of selling research studies and fielding them turn to marketing scientists again and again with the same question: What is the minimum sample size acceptable for my MaxDiff? However, there is no simple answer to this question. MaxDiff responses are most typically analyzed using Hierarchical Bayesian (HB) estimation. The output of this estimation is the "utilities" that represent each respondent's preference for a given item. Each utility is just a point estimate, an average across several thousand posterior draws (the exact number of the draws is defined by the analyst). The statistical properties of the distribution of MaxDiff utilities are unknown and therefore, it is impossible to assess the precision of MaxDiff point estimates for any given sample size the way it is possible with such traditional frequentist statistics as a mean or proportion.

Sample size needed for a MaxDiff study is a question that even the provider of the most popular HB software in Market Research field (Sawtooth Software) cannot answer unambiguously. Sawtooth Software's "rule of thumb" of defining sample size for traditional Conjoint studies based on the standard errors of the aggregate logit analysis of simulated conjoint

[1] stan.lipovetsky@gfk.com, dimitri.liakhovitski@gfk.com, mike.conklin@gfk.com

responses ("Advanced Test" in SSI Web, aspired standard errors for main effects are <0.05) does not apply to MaxDiff. As a result, researchers turn to approximations which are not related HB analyses per se. For example, for a MaxDiff study with 15 items one could argue that the proportion of times each item would be preferred could be considered equal to ~1/15 = 0.067. Given this proportion and the statistical rules for defining sample size for traditional proportions one could determine a minimal sample size for such a situation, assuming a desired precision level and a given confidence interval. Another simplistic approach to sample size for a MaxDiff study would be to pretend that items in a MaxDiff are just attributes of a regular DCM with 2 levels each (present vs. absent), run an advanced test in Sawtooth Software's SSI Web for a given sample size, and assess the standard errors of the aggregate logit analysis. Unfortunately, none of the aforementioned rules of thumb is based on solid research. The current study addresses this deficit.

We conducted a computer simulation study that would estimate the impact of the sample size on the quality of HB estimates for a MaxDiff exercise. After considering the results, we propose a new heuristic method ("rule of thumb") or maybe validate an existing one that would help marketing scientists estimate the appropriate sample size for a given MaxDiff study.

More specifically, we perform the study as follows:

1. Generate "true" utilities for a large population of synthetic respondents. Number of items and underlying item preferences in the population will be varied systematically from one experimental condition of the study to another.
2. Draw samples of consecutively increasing sizes from the population; for each sample, individual responses to a MaxDiff exercise will be generated based on the true utilities, the MaxDiff design and Gumbel error.
3. Estimate individual level utilities for the sample drawn using the R package ChoiceModelR. Certain parameters (e.g., prior variance) of MaxDiff estimation will be systematically varied from one experimental condition of the study to another.
4. Assess the quality of the estimated MaxDiff utilities at the (a) individual level, and (b) aggregate level.
5. Compare the observed quality of the utility estimates with those expected based on several approximation methods (rules of thumb) mentioned above.

We expect the results of this study to be of great interest and considerable practical use for all Marketing and Data Sciences colleagues.

II. Assumptions and Definitions

MaxDiff utilities are usually estimated using HB, and the statistical properties of the distribution of MaxDiff utilities are unknown so it is hard to derive an a priori analytical rule of thumb. "Each respondent sees each item 3 times" rule of thumb is independent of N; "Simulate the MaxDiff responses and assess the standard errors from aggregate logit" is not currently offered by Sawtooth Software for MaxDiff studies, so it requires several manual steps. The aim of this study is to help marketing scientists determine a desired sample size N for a MaxDiff study based on:

- MOE—desired Margin of Error,
- α—level of significance, or (1 - Confidence Probability),

- *n*—total number of items in MaxDiff,
- *t*—total number of tasks in the design,
- *m*—number of items per task for each respondent,

thus, $k = tm/n$ is the number of times each respondent sees each item.

The analytical approach to estimating errors for MaxDiff utilities and probabilities is presented in Appendix. Here we repeat the main needed formulae.

Based on well-known logit and multinomial-logit (MNL) modeling assumptions, the absolute probability p of choosing an item is:

$$p_j = \frac{\exp\left(util_j\right)}{1+\exp\left(util_j\right)}, \quad j = 1,2,...,n\,. \tag{1}$$

Relative probability P of choosing an item (MNL share) is:

$$P_j = \frac{\exp\left(util_j\right)}{1+\sum_{j=1}^{n}\exp\left(util_j\right)} \tag{2}$$

The relative error rate is:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k p_k(1-p_k)}} + \sum_{j=1}^{n} P_j \sqrt{\frac{1}{2N_j p_j(1-p_j)}}\,. \tag{3}$$

where ΔP_k is the standard error of P_k. We want to simplify the last expression to get rid of the parameters we do not have before the study.

Relative error rate term in *Simplification* 1: let us assume all p_j are at some constant level, p_j=p_k=p=const. Then (3) can be reduced to the expression (see details in Appendix, [40]):

$$\frac{\Delta P_k}{P_k} = \sqrt{\frac{2n}{mtNp(1-p)}}\,. \tag{4}$$

For a given level of significance α, the margin error (MOE) δ is defined via (4) as:

$$\delta = z_{\alpha/2}\frac{\Delta P_k}{P_k} = z_{\alpha/2}\sqrt{\frac{2n}{mtNp(1-p)}}\,, \tag{5}$$

so for a given MOE the total sample size is estimated from (5) as follows:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{2n}{mtp(1-p)}\,. \tag{6}$$

Relative error rate term in *Simplification* 2: let us assume approximately equal values of P_j in (3), and N_j=N_k as the design is balanced. Then we simplify the formula (3) (see in Appendix, [47]) to:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right)} = \sqrt{\frac{n}{2mtN}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right)}, \tag{7}$$

and similarly to (5) we can write the MOE:

$$\delta = z_{\alpha/2}\frac{\Delta P_k}{P_k} = z_{\alpha/2}\sqrt{\frac{n}{2mtN}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right)}. \tag{8}$$

From (8), for a given MOE, the total sample size corresponds to:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{n}{2mt}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right)^2. \tag{9}$$

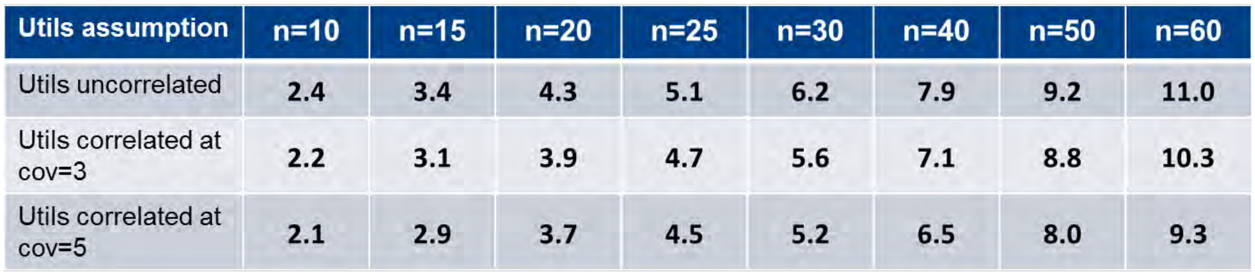
Which of these two formulas (6) or (9) should we use? If we ignore *π/n* for larger *n*, then assuming the same *n*, *m*, *t*, and *p*: the required *N* in the formula (6) yields about 4 times bigger value than the formula (9), that is a very big difference in a sample size estimation. We need to know which of the two formulas gives more adequate results.

III. Computer Simulation

For simulating multiple populations with "true" MaxDiff utilities, we performed the following:

1. Created 100 populations with "true" utilities for a given *n*,
2. Calculated absolute probability *p* and relative probability (share) *P* for each item,
3. Calculated Standard Deviations (SD) of both *p* and *P*—across *n* items,
4. Calculated SD(*p*)/SD(*P*) ratio,
5. Calculated the average of this ratio across 100 populations.

Ratios of SD(*p*)/SD(*P*) for different population parameters are as follows:



Utils assumption	n=10	n=15	n=20	n=25	n=30	n=40	n=50	n=60
Utils uncorrelated	2.4	3.4	4.3	5.1	6.2	7.9	9.2	11.0
Utils correlated at cov=3	2.2	3.1	3.9	4.7	5.6	7.1	8.8	10.3
Utils correlated at cov=5	2.1	2.9	3.7	4.5	5.2	6.5	8.0	9.3

Clearly, absolute probabilities *p* vary much more widely than the shares *P*. Thus, we should better use the *Simplification* 2 assumption: P_j = P_k= *P* = const, and we can focus on the formula (9).

The next step consists in the main simulation, with the objective:

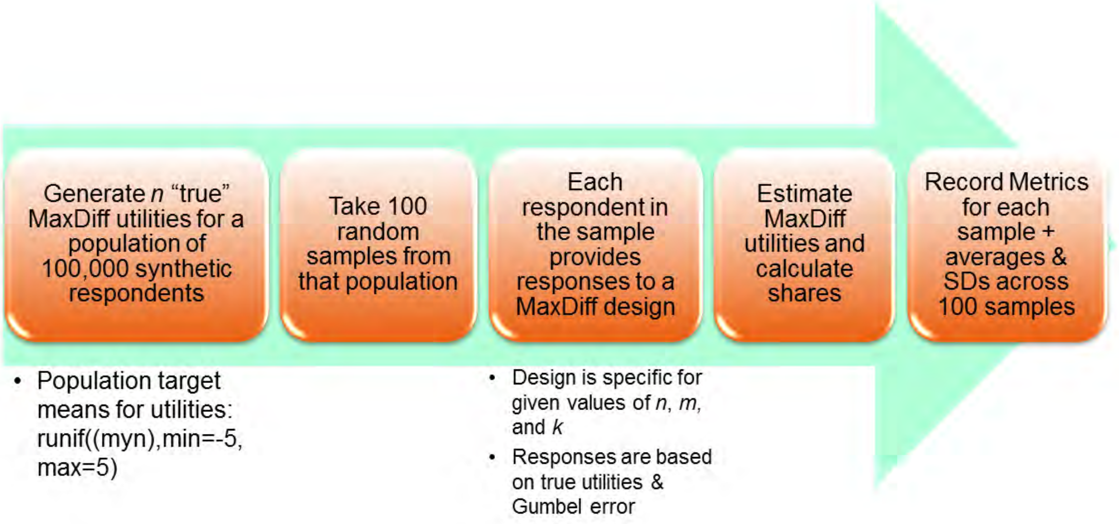
1. Simulate “true” MaxDiff utilities for different scenarios,
2. Simulate sample responses to a MaxDiff exercise,
3. Compare observed vs. analytically derived (formula-based) Relative Error *ΔP*/*P*.

Experimental factors varied in the simulation study are shown in the table below:

Symbol	Explanation	# of Levels Tested	Levels Tested
n	# of items in MaxDiff	8	10, 15, 20, 25, 30, 40, 50, and 60
8 levels of *n* determined 8 populations we sampled from; Each population had 100,000 potential “respondents”.			
N	Sample size	12	80,100,120,140,160,180,200,250,300,400,500, and 600
m	# of items per task	5	3,4,5,6, and 7
k	# of times each item was seen by respondents	2	Fewer (~1) and More (~2)

The total number of combinations tried was 8 * 12 * 5 * 2 = 960. For every combination of factors 100 random samples were drawn. With *k* = *nm/t,* we created different MaxDiff designs in SSI Web to ensure the right value of *k.*

Simulation Flow can be presented as follows:



Generate *n* “true” MaxDiff utilities for a population of 100,000 synthetic respondents

Take 100 random samples from that population

Each respondent in the sample provides responses to a MaxDiff design

Estimate MaxDiff utilities and calculate shares

Record Metrics for each sample + averages & SDs across 100 samples

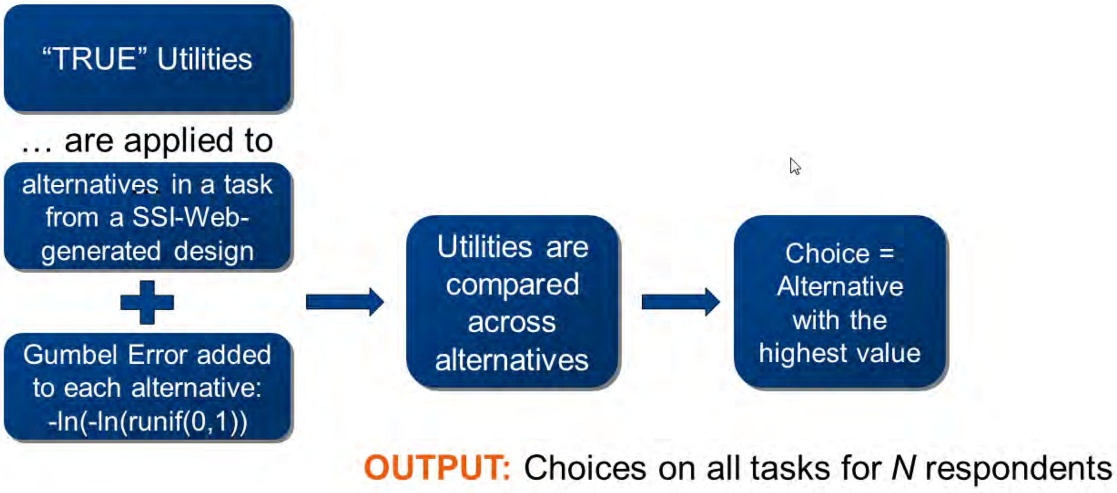
- Population target means for utilities: runif((myn),min=-5, max=5)

- Design is specific for given values of *n*, *m*, and *k*
- Responses are based on true utilities & Gumbel error

The original (raw) utilities ranged between ~ -5 and +5. The table below shows the percentiles for the relationship of the absolute Gumbel error to the absolute original raw utility for one condition (*n*=60, *N*=200). For 40% of individual item utilities the size of the absolute amount of Gumbel error added was 40% of the original utility or larger. For 30% of cases it was 58% of the utility or larger.

Percentile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Value	0.043	0.087	0.137	0.197	0.280	0.395	0.579	0.921	1.897

Generation of MaxDiff Choices Based on “True” Utilities can be seen on the following graph:


“TRUE” Utilities
… are applied to
alternatives in a task from a SSI-Web-generated design
+
Gumbel Error added to each alternative: -ln(-ln(runif(0,1))
Utilities are compared across alternatives
Choice = Alternative with the highest value
OUTPUT: Choices on all tasks for N respondents

IV. Closed Form Solution and Comparisons to HB

We had to estimate MaxDiff utilities for 96,000 samples. It would have taken an unbearable burden of computer time to run an HB for each sample.

Instead, we relied on the analytical closed-form solution from Lipovetsky & Conklin (2014a). According to this method, the sample level utility u for any MaxDiff item equals to:

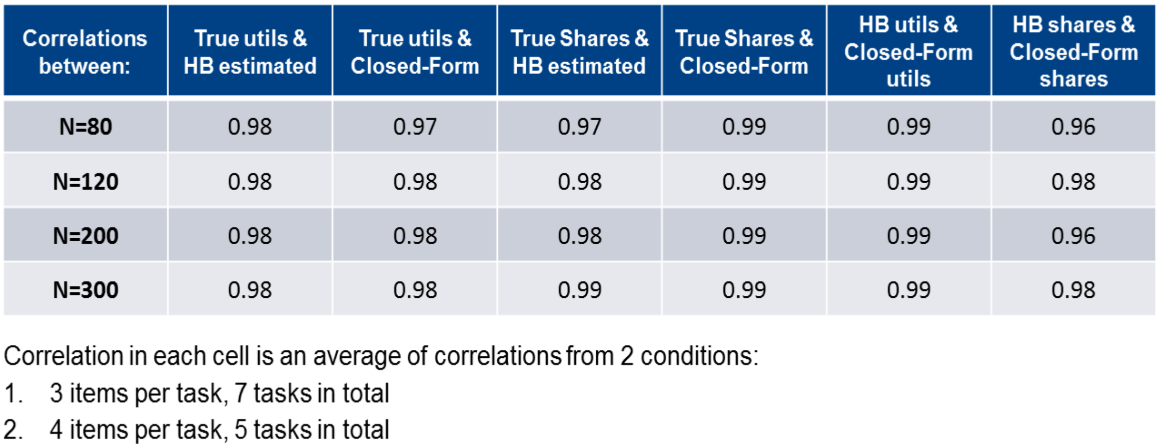
$$u = \ln \frac{1 + \Delta / n_{shown}}{1 - \Delta / n_{shown}} , \tag{10}$$

where $\Delta = n_{best}$ - n_{worst} is the difference between the number of times n_{best} and n_{worst} this item was selected as the “best” and the “worst,” respectively. It is useful to note the following: the total number of times some item was shown equals $n_{shown} = n_{best} + n_{worst} + n_{neutr}$, where n_{neutr} is the number of times the item was not indicated as the “best” or “worst.” So in the case of strongly polarized respondents, when all the answers on an item are only “best” or “worst,” no neutral, we have the equality $n_{shown} = n_{best} + n_{worst}$. Then the general expression for utility (10) reduces to:

$$u = \ln \frac{1 + (n_{best} - n_{worst}) / (n_{best} + n_{worst})}{1 - (n_{best} - n_{worst}) / (n_{best} + n_{worst})} = \ln \frac{n_{best}}{n_{worst}} . \tag{11}$$

The last expression coincides with Jordan Louviere’s recommendation (1993) to use it for simple estimation of utility. But if n_{neutr} is not zero, it is better to use the general formula (10). Returning to computations, at first we compare the results of the estimation via HB and via closed-form solution for several conditions.

Estimation results for two methods for n=10 items are as follows:

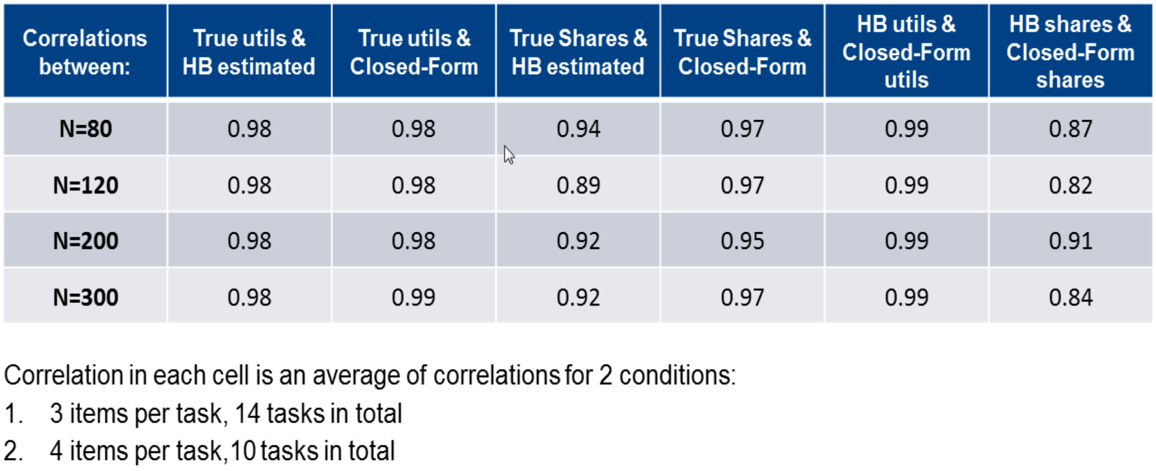


Correlations between:	True utils & HB estimated	True utils & Closed-Form	True Shares & HB estimated	True Shares & Closed-Form	HB utils & Closed-Form utils	HB shares & Closed-Form shares
N=80	0.98	0.97	0.97	0.99	0.99	0.96
N=120	0.98	0.98	0.98	0.99	0.99	0.98
N=200	0.98	0.98	0.98	0.99	0.99	0.96
N=300	0.98	0.98	0.99	0.99	0.99	0.98

Correlation in each cell is an average of correlations from 2 conditions:

1. 3 items per task, 7 tasks in total
2. 4 items per task, 5 tasks in total

Estimation results for two methods for *n*=20 items are as follows:

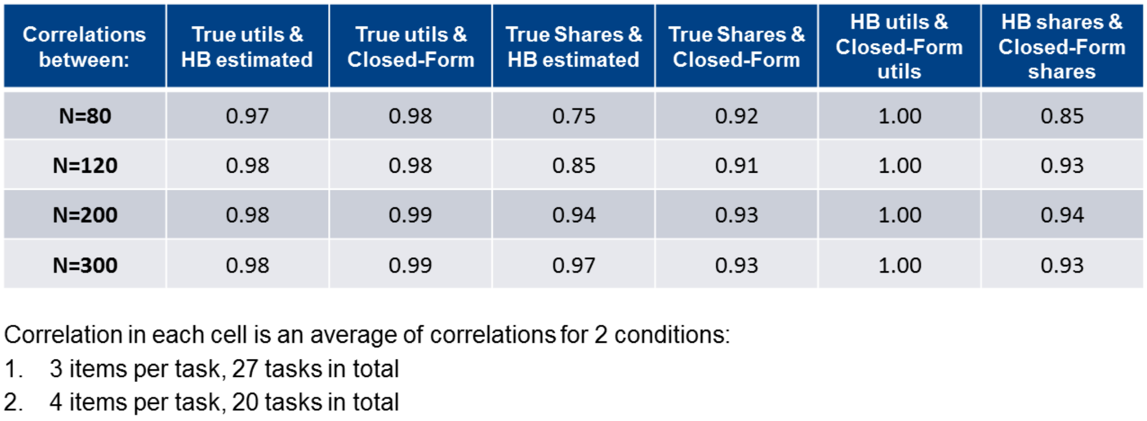


Correlations between:	True utils & HB estimated	True utils & Closed-Form	True Shares & HB estimated	True Shares & Closed-Form	HB utils & Closed-Form utils	HB shares & Closed-Form shares
N=80	0.98	0.98	0.94	0.97	0.99	0.87
N=120	0.98	0.98	0.89	0.97	0.99	0.82
N=200	0.98	0.98	0.92	0.95	0.99	0.91
N=300	0.98	0.99	0.92	0.97	0.99	0.84

Correlation in each cell is an average of correlations for 2 conditions:

1. 3 items per task, 14 tasks in total
2. 4 items per task,10 tasks in total

Estimation results for two methods for *n*=30 items are as follows:



Correlations between:	True utils & HB estimated	True utils & Closed-Form	True Shares & HB estimated	True Shares & Closed-Form	HB utils & Closed-Form utils	HB shares & Closed-Form shares
N=80	0.97	0.98	0.75	0.92	1.00	0.85
N=120	0.98	0.98	0.85	0.91	1.00	0.93
N=200	0.98	0.99	0.94	0.93	1.00	0.94
N=300	0.98	0.99	0.97	0.93	1.00	0.93

Correlation in each cell is an average of correlations for 2 conditions:

1. 3 items per task, 27 tasks in total
2. 4 items per task, 20 tasks in total

From this comparison we can conclude:

- The results of both the HB estimation and the Closed-Form solution reflect the relative standing of true utilities and true shares accurately enough;

- The Closed-Form solution can be used as a proxy for the HB estimation in the main simulation study that helps save an inordinate amount of computer time.

V. Findings in Simulations

Comparing observed relative errors with the analytically derived one by the formula (9), we calculated the left-hand side (LHS) values *ΔP/P* for each item, based on the observed population share *P,* so it is our *observed* value. On the right-hand side (RHS) of (9), we calculated it for each item based on the parameters *N*, *n*, *m*, and *t* for each condition and the *p* in the population, so it is our analytically *predicted* value.

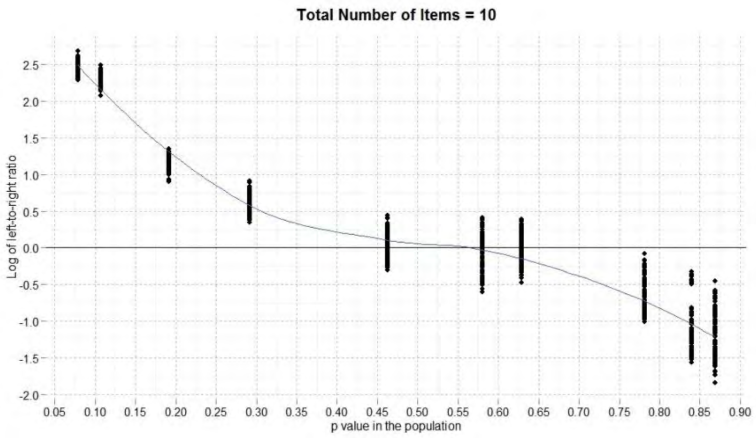
We want to determine for which value of *p* the both parts in (9) are equal, or LHS/RHS = 1. More exactly, we consider condition

$$ln(LHS/RHS) = 0, \qquad (12)$$

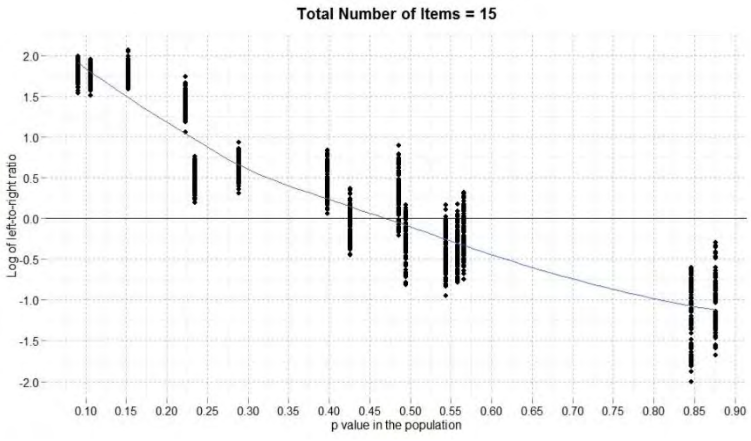
because in logarithms the differences between quotients that are >1 or <1 are equal, and it is easier to see on the graphs for which values of *p* the function has a root, or intercepts the horizontal axis.

On the next few graphs we present the log-ratio (12) in its dependence on the values of *p*, for several values of *n*. Each dot on the graphs presents one experimental condition of various parameters, and a smoother of Local Polynomial Regression Fitting (*loess* in R) is used.

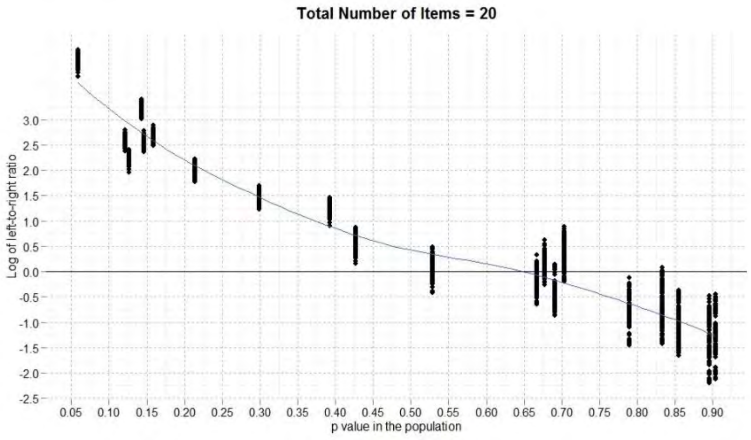
For n=10, correlation between *p* and the log-ratio is -0.94, and the fitted line crosses 0 at *p*=0.55:



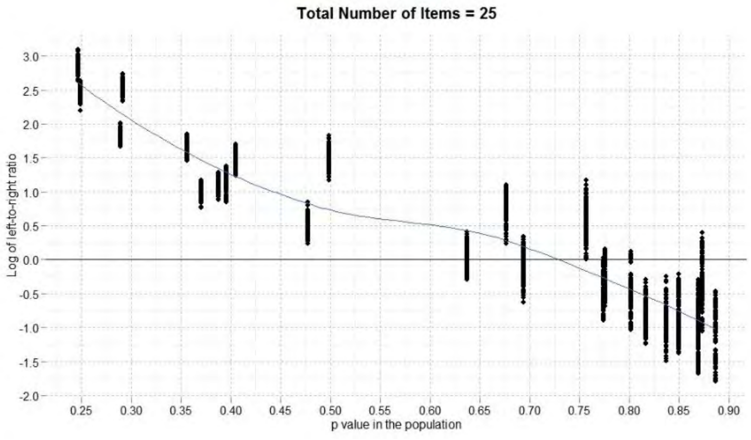
For n=15, correlation between *p* and the log-ratio is -0.93, and the fitted line crosses 0 at *p*=0.47:



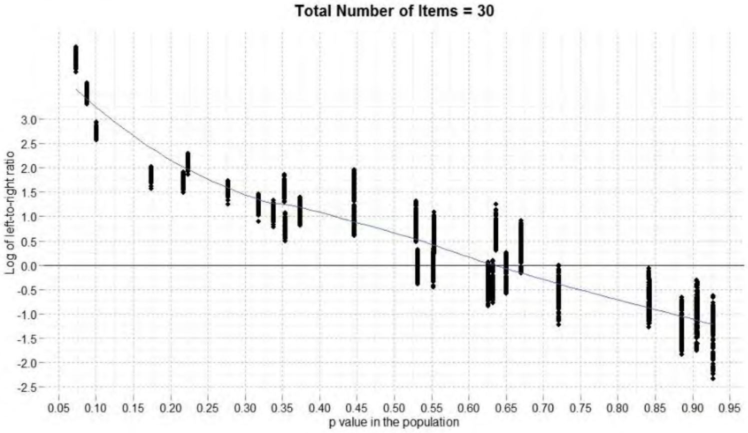
For n=20, correlation between *p* and the log-ratio is -0.96, and the fitted line crosses 0 at *p*=0.65:



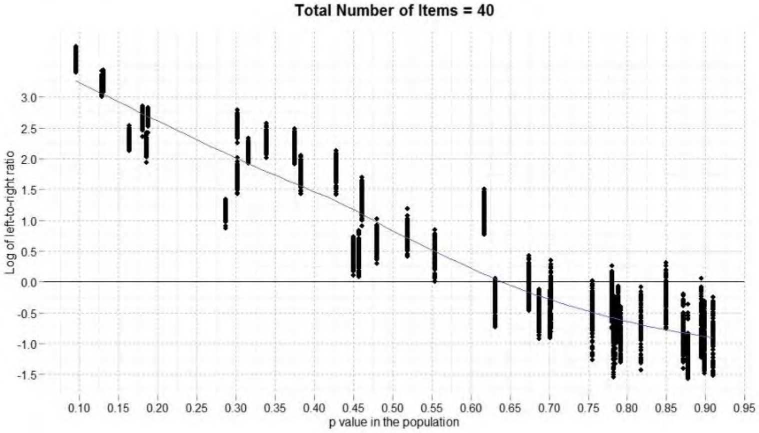
For n=25, correlation between *p* and the log-ratio is -0.93, and the fitted line crosses 0 at *p*=0.72:



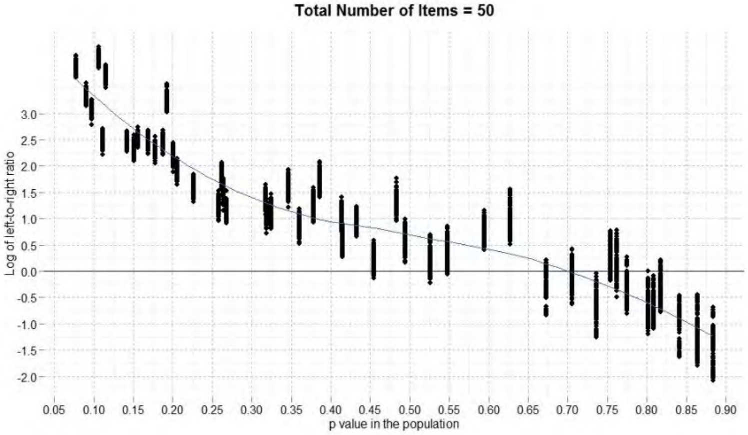
For n=30, correlation between *p* and the log-ratio is -0.93, and the fitted line crosses 0 at *p*=0.63:



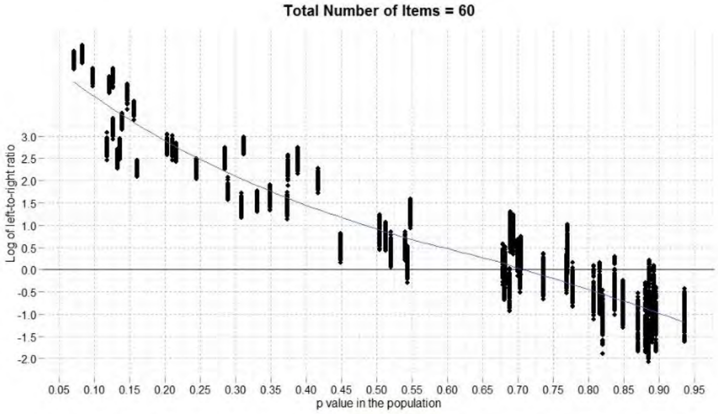
For n=40, correlation between *p* and the log-ratio is -0.95, and the fitted line crosses 0 at *p*=0.62:



For n=50, correlation between *p* and the log-ratio is -0.93, and the fitted line crosses 0 at *p*=0.70:



For n=60, correlation between *p* and the log-ratio is -0.94, and the fitted line crosses 0 at *p*=0.71:



VI. Conclusions: Formulas for *N* and Rough "Rules of Thumb"

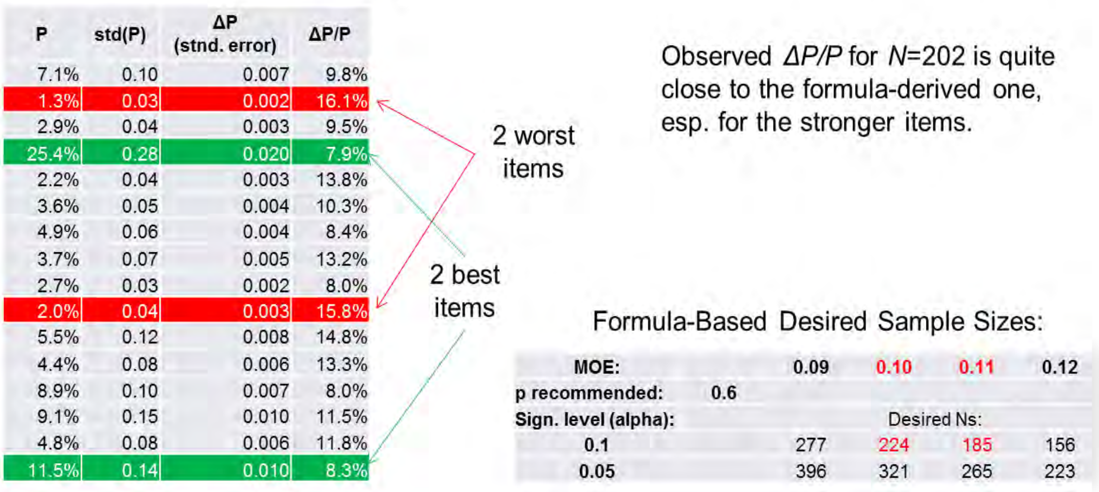
The proposed Relative Error formula (9) approximates the observed Relative Error ratio well, especially at the values $0.5 < p < 0.7$. For conditions most frequent in practice n = 10, 15, 20, 25, 30, the intercepts equal p=0.55, 0.47, 0.65, 0.72, 0.63, so the mean value is *p*=0.617. With it in the formula (7) the sample size (9) yields slightly changed coefficient:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{n}{2mt}\left(2.06+\frac{\pi}{n}\right)^2, \tag{13}$$

and for big *n*, when $2.06 >> \pi / n$, this formula simplifies to

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{2.12n}{mt}. \tag{14}$$

For example, look at a real MaxDiff Study: *n*=16 (items), *t*=12 (tasks), *m*=4 (items per task), *N*=202.



P	std(P)	ΔP (stnd. error)	ΔP/P
7.1%	0.10	0.007	9.8%
1.3%	0.03	0.002	16.1%
2.9%	0.04	0.003	9.5%
25.4%	0.28	0.020	7.9%
2.2%	0.04	0.003	13.8%
3.6%	0.05	0.004	10.3%
4.9%	0.06	0.004	8.4%
3.7%	0.07	0.005	13.2%
2.7%	0.03	0.002	8.0%
2.0%	0.04	0.003	15.8%
5.5%	0.12	0.008	14.8%
4.4%	0.08	0.006	13.3%
8.9%	0.10	0.007	8.0%
9.1%	0.15	0.010	11.5%
4.8%	0.08	0.006	11.8%
11.5%	0.14	0.010	8.3%

2 worst items

2 best items

Observed *ΔP/P* for *N*=202 is quite close to the formula-derived one, esp. for the stronger items.

Formula-Based Desired Sample Sizes:

MOE:		0.09	0.10	0.11	0.12
p recommended:	0.6				
Sign. level (alpha):		Desired Ns:			
0.1		277	224	185	156
0.05		396	321	265	223

For a simple rough "rule of thumb," we see that (14) is similar to the last formula in Appendix:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{2n}{mt} \tag{15}$$

With the parameter alpha=0.05 and MOE delta=0.1, (15) simplifies to:

$$N \geq 800\frac{n}{mt}. \tag{16}$$

For a given ***n***, we can try several combinations of ***m*** and ***t*** that satisfies a need for the sample size ***N***. We see that for a reasonable sample size of about, say, *N*=200, the product *mt* should be 4 times bigger than *n*. With a bigger error margin delta=0.15, a smaller sample size is required:

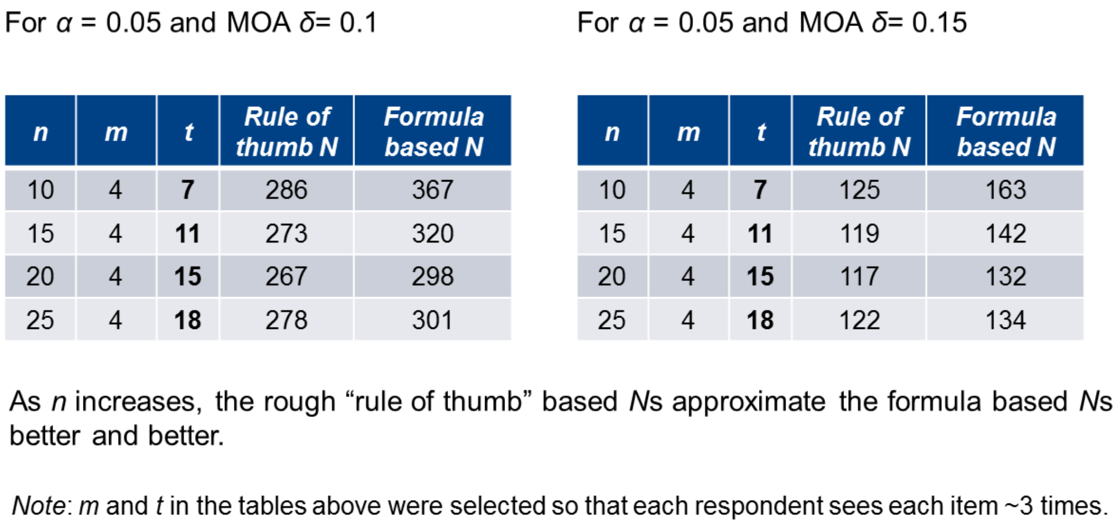
$$N \geq 350\frac{n}{mt}. \tag{17}$$

With a more rough error margin, say, delta=0.2, a "rough" rule of thumb is:

$$N \geq 200\frac{n}{mt}. \tag{18}$$

For a given ***n***, we try combinations of ***m*** and ***t*** that satisfies the inequality for sample size ***N*** (17) which shows that the product *mt* should be approximately equal *n*.

Example "rule of thumb":



For α = 0.05 and MOA δ= 0.1

n	***m***	***t***	***Rule of thumb N***	***Formula based N***
10	4	**7**	286	367
15	4	**11**	273	320
20	4	**15**	267	298
25	4	**18**	278	301

For α = 0.05 and MOA δ= 0.15

n	***m***	***t***	***Rule of thumb N***	***Formula based N***
10	4	**7**	125	163
15	4	**11**	119	142
20	4	**15**	117	132
25	4	**18**	122	134

As *n* increases, the rough "rule of thumb" based *N*s approximate the formula based *N*s better and better.

Note: *m* and *t* in the tables above were selected so that each respondent sees each item ~3 times.

In future work, it could be useful to try Bayesian approach in MaxDiff sample size estimation. In assumption of normal distribution for the individual proportions with the conjugate prior (Gelman et al., 2004, p. 49 and p. 79; Wonnacott and Wonnacott, 1977, pp.556–562), the expression for the posterior mean (which is the proportion in our case) is

$$\widetilde{p}_i = \frac{n_i}{n_i + S} p_i + \frac{S}{n_i + S} p , \qquad (19)$$

where the individual p_i and prior p frequencies are combined into the posterior frequency $\widetilde{p}_i$ with the number of observations n_i and S for individual and prior data. The value S for priors makes sense of a reciprocal variance which is not always easy to find, so more work is needed on this issue.

Summary

In this research we have performed the following:

- Conducted simulations and demonstrated the usefulness of a closed-form analytical solution for MaxDiff simulation studies;
- Derived analytical formulas for determining the needed sample size for MaxDiff studies;
- Tested the assumptions of the formulas in multiple simulations;
- Demonstrated the soundness of the formula to be used for determining sample size for MaxDiff studies;
- Proposed rough "rules of thumb" for determining the sample size for MaxDiff, convenient for practical applications.

Acknowledgements

We are grateful to Keith Chrzan for the comments improving the work and correspondence on the utility relations (10)–(11) in Section IV, and to Anthony Marley for discussion on the formulae (17)–(22) in the Appendix.

Stan Lipovetsky

Dimitri Liakhovitski

Michael Conklin

APPENDIX: ANALYTICAL CONSIDERATION

The problem of sample size estimation needed for a MaxDiff project is very important in practical marketing research. It can be considered as estimation of a proportion with a required precision, or margin of error, with a needed confidence probability, considered in (Louviere et al., 2000, ch.9.3, with a small evident typo; Hensher et al., 2005, ch. 6.7; Orme, 2010, ch. 7).

Let us briefly describe it. For a sample size *N*, a proportion *p* has its standard error defined as

$$\Delta p = \sqrt{p(1-p)/N}\,. \tag{1}$$

In normal approximation to binomial distribution, the confidence intervals for proportion are:

$$p_{\pm} = p \pm z_{\alpha/2}\Delta p\,, \tag{2}$$

where for the significance level alpha (or confidence probability 1-alpha) of two-tail test, z-value of the normal distribution is

$$z_{\alpha/2} = F^{-1}(1-\alpha/2)\,. \tag{3}$$

The commonly used levels of significance alpha and corresponding *z*-values are as follows: alpha=0.1, z=1.64; alpha=0.05, z=1.96; alpha=0.01, z=2.58.

The absolute margin of error to a proportion in (2) is

$$\varepsilon = z_{\alpha/2}\Delta p = z_{\alpha/2}\sqrt{p(1-p)/N}\,, \tag{4}$$

but it is more convenient to work with a relative error requiring a desired percent of the error to the proportion:

$$\delta = \varepsilon / p = z_{\alpha/2}\Delta p / p = z_{\alpha/2}\sqrt{\frac{1-p}{pN}}\,. \tag{5}$$

Then (2) can be represented via the relative margin error as

$$p_{\pm} = p(1 \pm z_{\alpha/2}\delta)\,. \tag{6}$$

For a required % of margin error δ and level of significance α, we solve (5) for the needed minimum sample size and obtain the estimation commonly used in literature.

$$N > \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{1-p}{p}\,. \tag{7}$$

Let us consider sample size estimation with the specific features of MaxDiff modeling. As it was shown in (Lipovetsky and Conklin, 2014a), the MaxDiff results for utilities and choice probabilities can be presented in analytical closed-form solution for multinomial-logit (MNL) models. The analytical formulae permit the inference of the characteristics of the model's quality, including standard errors of the utilities and choice probabilities, the residual deviance and pseudo-R^2. To review some properties of MaxDiff, each respondent is presented with several (two or three dozen) subsets with a few items in each one, by way of a balanced design where each of the items is shown an approximately equal number of times. The respondents answer which of several presented items is the best and the worst of them in each task. The estimation of utilities is usually performed using MNL.

For explicit presentation of the approach, consider an example from a marketing research project for prioritizing seventeen items. Each of 3,062 respondents saw four out of all seventeen items in each of the ten tasks from which the best and worst of the four were chosen. So the data matrix with 3,062x10=30,620 rows can be constructed.

As it is used in discrete choice modeling (DCM), each row of such a matrix can be split into several rows, by the number of the items shown in each task, so in this example it will be split to four rows with the shown items indicated by the value 1, otherwise 0. In the last column of the outcome one of these four rows contains the value 1 indicating that this item was chosen as the best one. Thus, the total number of rows in the DCM matrix equals 30,620x4=122,480. The DCM corresponds to a choice among several outcomes and can be described by a multinomial-logit model with the probability of a choice presented as:

$$P_k = \frac{\exp(a_k x_k)}{\sum_{j=1}^{n} \exp(a_j x_j)} , \tag{8}$$

where x_j are the binary variables. The parameters a_k are the utilities which define the probability of each k-th choice among all n of them (n=17 in our example). For the sake of identification, one of the parameters (8) is taken as a reference, so put to zero. The possibility of using a binary logit model for the estimation of parameters for MNL is discussed in many studies which show that the ratio of k-th and j-th multinomial shares (8) does not depend on the other alternatives, so it is identical to comparing two choices and ignoring the other ones (the feature of Independence of Irrelevant Alternatives, or IIA). This means that a multinomial model can be estimated via the set of pairwise logistic models providing consistent estimates of the parameters.

The MaxDiff design has the orthogonal structure of the compared items, so there is no loss in efficiency and no impact on the standard errors or t-statistics. Thus, instead of the MNL model (8) it is possible to construct the logistic regression model:

$$p_k = \frac{\exp\left(a_1 x_1 + ... + a_n x_n\right)}{1 + \exp\left(a_1 x_1 + ... + a_n x_n\right)} , \tag{9}$$

where p_k is the binary outcome of the best choice. The logit model (9) defines the probabilities of choice for any k-th alternative because in any observation (a row in the binary data matrix) only one variable x_k equals one and others equal zero. So actually (9) presents in a unified form all pairwise logit models. When the parameters in (9) are estimated, the probabilities of different choices are found using MNL (8). Estimation of the parameters via the model (9) is theoretically equivalent to MNL model (8) (Hausman and McFadden, 1984, pp. 1222–1223; Alvarez and Nagler, 1998, p. 61).

The DCM problem of the “Worst” item choice can be considered in the same approach as (8)–(9). For a simultaneous estimation by all the best and worst choices in one combined dataset we use the following property: if we change the signs of all variables then the probability estimated by the logistic model (9) equals 1-*p* which defines the absence of a binary event. Indeed, consider the transformation of sign change:

$$\frac{\exp\left(-(a_1x_1+...+a_nx_n)\right)}{1+\exp\left(-(a_1x_1+...+a_nx_n)\right)}=1-\frac{\exp\left(a_1x_1+...+a_nx_n\right)}{1+\exp\left(a_1x_1+...+a_nx_n\right)}=1-p. \qquad (10)$$

The design matrix for the "worst" part of data is given by the tasks split to four rows each with the shown items indicated by -1, and the binary outcome of the choice of the "worst" in the last column indicated by 1. The total number of rows in this DCM design matrix also equals 30,620x4=122,480, and it can be used for modeling the worst item choice by (8)–(9).

In practical MaxDiff modeling both DCM design matrices are combined into one total matrix of choices. With such a design, the positive and negative values of the binary predictors would push the outcomes with the values 1 to the sides of maximum (9) and minimum (10) probability, respectively, while zero values would tend to belong to the middle segment of the logit curve. The combined data for all respondents is of doubled size, 122,480x2=244,960 rows, because from the original 3,062 respondents we now have 80 times more rows (ten sets of tasks by four shown items doubled due to the best and the worst choices). The individual parameters of the logit model can be estimated in hierarchical and empirical Bayesian approach. With enough data (such as 80 rows per person in our example) it is possible to construct logit models (9) and find the MaxDiff coefficients of utility for each respondent separately.

One of the very important features of MaxDiff data is the orthogonality of its binary predictors. Indeed, the scalar product of any two variables x_j and x_k equals zero because each item in a row enters only one column, so $x'_j x_k = 0$ for $j \neq k$. This means that in a model without an intercept (9) all *n* items (all 17 variables in the example) can be used. The orthogonality of predictors leads to the separability of the total logit (9) into logistic models by each variable. Indeed, similarly to what was discussed in relation to formula (9), this logit model defines probabilities of choice for any *k*-th alternative because in any row only one variable x_k equals one and others equal zero. So the formula (9) presents a unified form of all the pairwise logit models for MaxDiff.

Speaking more formally, in constructing a model (9) by the orthogonal binary variables x_j, it is easy to notice that the Newton-Raphson procedure for parameter estimation of the non-linear logistic regression reduces to the separate equations by each variable. This happens because the predictors are orthogonal and binary (where an x_j variable has ones, all the other x_k variables have zeros) so the Hessian matrix of the second derivatives (the analogue of the covariance matrix for a linear model) reduces to the diagonal matrix, thus, iterations go independently by each x_j variable. Thus, estimation of the parameters (9) is separable by each of them, so it can be performed by the set of pairwise logit regressions:

$$p_j=\frac{\exp\left(b_jx_j\right)}{1+\exp\left(b_jx_j\right)}, \quad j=1,2,...,n\,. \qquad (11)$$

The coefficients of utility coincide with estimation by the logit models with all the predictors (9) and with each one (11) so $a_j = b_j$.

The next important step is that in the utility estimation for any paired model (11) with only one *j*-th predictor, we use only the rows where x_j is presented as $x_j =1$ or $x_j = -1$. All the other rows are redundant as containing only zeros by the predictors and the outcome—because if x_j is not presented it cannot be chosen either as the best or the worst item. The model (11) for this small dataset can be represented as:

$$p_j = \frac{\exp\left(c_j x_j [x_j = \pm 1]\right)}{1 + \exp\left(c_j x_j [x_j = \pm 1]\right)}. \tag{12}$$

Its coefficients coincide with the coefficients of models (11) and (9), so the equalities hold:

$$a_j = b_j = c_j, \quad j = 1, 2, ..., n\,. \tag{13}$$

The standard deviations and *t*-statistics of the coefficients are the same as if found by the total (9), or partial (11), or small data subset (12) models.

Actually, logistic regression modeling is not needed at all for estimation of the coefficients in the model (12). For a binary predictor x_j and binary event p_j the distribution of the outcomes can be presented in a 2x2 contingency table as follows:

Choices in x_j		Binary outcome		How many times the j-th item was shown total
		not chosen	chosen as the best or worst	
		p=0	p=1	
Subset for the worst	x = -1	$N_j\,{}^{-}N_j^{worst}$	N_j^{worst}	N_j
Subset for the best	x = 1	N_j- N_j^{best}	N_j^{best}	N_j
Total		$2N_j$ – (N_j^{best} +N_j^{worst})	N_j^{best} +N_j^{worst}	$2N_j$

The hitrate defined as a quotient of the diagonal elements' sum divided by base size equals

$$p_j = \frac{N_j - N_j^{worst} + N_j^{best}}{2N_j} = \frac{1}{2} + \frac{f_j^{best} - f_j^{worst}}{2} = 0.5(1 + \Delta f_j)\,, \tag{14}$$

with the proportions defined as

$$f_j^{best} = N_j^{best} / N_j, \quad f_j^{worst} = N_j^{worst} / N_j\ , \tag{15}$$

and their difference denoted as

$$\Delta f_j = f_j^{best} - f_j^{worst}. \tag{16}$$

The value (14) equals the sample probability in the model with one parameter (12). Indeed, in (12) for any j, we have the following cases (our thanks to Prof. Anthony Marley for discussions on this issue):

i. x= -1, chosen. Theoretical value (12) and estimated value by (15) or by Table 2 are:

$$p_j^{worst} = \frac{\exp\left(-c_j\right)}{1+\exp\left(-c_j\right)}, \quad \tilde{p}_j^{worst} = \frac{N_j^{worst}}{N_j} = f_j^{worst} \quad . \tag{17}$$

ii. x= -1, non-chosen. Theoretical value (12) and estimated value by (15) or Table 2 are:

$$1 - p_j^{worst} = \frac{1}{1+\exp\left(-c_j\right)}, \quad 1 - \tilde{p}_j^{worst} = \frac{N_j - N_j^{worst}}{N_j} = 1 - f_j^{worst} \quad . \tag{18}$$

iii. x= 1, chosen. Theoretical value (12) and estimated value by (15) or by Table 2 are:

$$p_j^{best} = \frac{\exp\left(c_j\right)}{1+\exp\left(c_j\right)}, \quad \tilde{p}_j^{best} = \frac{N_j^{best}}{N_j} = f_j^{best} \quad . \tag{19}$$

iv. x= 1, non-chosen. Theoretical value (12) and estimated value by (15) or Table 2 are:

$$1 - p_j^{best} = \frac{1}{1+\exp\left(c_j\right)}, \quad 1 - \tilde{p}_j^{best} = \frac{N_j - N_j^{best}}{N_j} = 1 - f_j^{best} \quad . \tag{20}$$

From these cases, the hitrate (14) is the mean of the terms in ii and iii, so theoretical value is

$$\frac{1}{2}\left[\frac{1}{1+\exp\left(-c_j\right)} + \frac{\exp(c_j)}{1+\exp\left(c_j\right)}\right] = \frac{\exp(c_j)}{1+\exp\left(c_j\right)} = p_j , \tag{21}$$

and the estimated value is:

$$\frac{1}{2}\left[\frac{N_j - N_j^{worst}}{N_j} + \frac{N_j^{best}}{N_j}\right] = \frac{1}{2} + \frac{f_j^{best} - f_j^{worst}}{2} , \tag{22}$$

which coincide with the results in (14).

If in place of the difference (16) between the proportions in (15) we consider their quotient, then it can be seen as a ratio for the odds between two groups, like the treatment and control (in the contingency table above). In the analogue with the well-known log-linear models, the considered approach can be called the logit-linear model. The described formulae can be extended further to derive the characteristics of quality for logistic regression. The sample proportions (15) can be found simply by counting the totals for the number of times N_j each item was presented to respondents and how many times it was chosen as the best or the worst one. The relations between choice experiments and cross-tables have been known since the works (Manski and McFadden, 1981; Louviere and Woodworth, 1983; Louviere, 1993). The idea of a simple estimation based on the difference of proportions or counts of the best and worst choices has been discussed in the literature (Finn and Louviere, 1992; Louviere et al., 2008, 2013; Orme, 2009; Flynn, 2010) but rather in a general qualitative description, without the exact formulation (8)–(22). In our approach, we can find all the parameters in the exact analytical formulae which are very convenient for numerical and analytical consideration of the choice probabilities. More results on the analytical formulae of the standard errors and t-statistics for utilities, and various other characteristics of the model quality and numerical comparisons are given in (Lipovetsky and Conklin, 2014a,b; Lipovetsky, 2014).

The standard errors for proportions (14) can be presented similarly to (1) as:

$$\Delta p_j = \sqrt{\frac{p_j(1-p_j)}{2N_j}}\ , \tag{23}$$

where the sample size is $2N_j$ as in (14). Let us express the number of times N_j which a j-th item is presented to the respondents via the other parameters of MaxDiff modeling:

$$N_j = \frac{m_{items\ per\,task}}{n_{items\ total}} t_{number of\ tasks} N_{respondents\,total} \equiv \frac{m}{n} tN\ , \tag{24}$$

where meaning of parameters and their brief notations are given. Using (24) in (23) we get:

$$\Delta p_j = \sqrt{\frac{p_j(1-p_j)}{2mtN/n}}\ . \tag{25}$$

Substituting (25) into (5) yields:

$$\delta = z_{\alpha/2}\frac{\Delta p_j}{p_j} = z_{\alpha/2}\sqrt{\frac{1-p_j}{p_j 2mtN/n}}\ . \tag{26}$$

For the given % of margin error δ and level of significance α, also with the taken number of items in one task m within the total number of items n, and with the number of tasks for each respondent t, we solve (26) for the needed total sample size:

$$N \ge \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{1-p_j}{p_j}\frac{n}{2mt}\ . \tag{27}$$

This formula can be simplified taking into account that the confidence probability is usually taken at the level 95%, and probability can be taken at the mean level, so

$$z_{\alpha/2} \approx 2, \quad p_j = 0.5\ . \tag{28}$$

Then with these values (27) reduces to:

$$N \ge \frac{2n}{\delta^2 mt}\ . \tag{29}$$

The precision δ is usually taken about 0.05, or 0.1. For instance, $\delta = 0.1$ simplifies (29) to the estimate convenient for practical needs:

$$N \ge 200\frac{n}{mt}\ . \tag{30}$$

Using there any assigned numbers of n, m, and t (for instance, n=20, m=5, t=10) we obtain an output. Sometimes we need to consider an opposite problem of estimating the number of tasks by the given sample size. In this case we have:

$$t \geq \left(\frac{z_{\alpha/2}}{\delta} \right)^2 \cdot \frac{n}{2mN} . \tag{31}$$

Till now we worked with the absolute choice probability of each one item, but let us consider how to estimate sample size for the conditional choice probability of one item among the others given by the MNL model (8). Finding differential of (8) in presence of all items (x=1) we get:

$$dP_k = \frac{\exp(a_k)}{\sum_{j=1}^{n} \exp(a_j)} da_k - \sum_{j=1}^{n} \frac{\exp(a_k)}{\sum_{j=1}^{n} \exp(a_j)} \frac{\exp(a_k)}{\sum_{j=1}^{n} \exp(a_j)} da_j = P_k da_k - P_k \sum_{j=1}^{n} P_j da_j . \tag{32}$$

Substituting differentials with finite changes yields the standard error for MNL probability as total of absolute values of all items:

$$\Delta P_k = P_k (\Delta a_k + \sum_{j=1}^{n} P_j \Delta a_j) , \tag{33}$$

and the corresponding relative error is

$$\Delta P_k / P_k = \Delta a_k + \sum_{j=1}^{n} P_j \Delta a_j . \tag{34}$$

By estimated probability p_j taking into account (13) we find utility from the logit (11) in log(odds):

$$a_j = \ln \frac{p_j}{1 - p_j} , \tag{35}$$

so its standard error can be found by differential:

$$da_j = d \left(\ln \frac{p_j}{1 - p_j} \right) = \left(\frac{1}{p_j} + \frac{1}{1 - p_j} \right) dp_j = \frac{dp_j}{p_j (1 - p_j)} . \tag{36}$$

Then using (23) (neglecting subtracting one) for the proportion standard error in (36) yields the expression (see also in Lipovetsky and Conklin, 2014a, formula [13]) for the standard error of the utility coefficient as follows:

$$\Delta a_j = \sqrt{\frac{1}{2 N_j p_j (1 - p_j)}} , \tag{37}$$

Then substituting it into (34) yields:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k p_k(1-p_k)}} + \sum_{j=1}^{n} P_j \sqrt{\frac{1}{2N_j p_j(1-p_j)}} \,. \tag{38}$$

For a balanced design with almost equal number of times each item shown we have an equality N_j=N_k=const. And for simplification of the formula (38) we can take all p_j at some constant (mean) level, p_j=p_k=p=const, when it reduces to the following:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k p(1-p)}} + \sqrt{\frac{1}{2N_k p(1-p)}} \sum_{j=1}^{n} P_j = \sqrt{\frac{2}{N_k p(1-p)}} \,, \tag{39}$$

where we take into account that the total of all MNL shares (8) equals one.

Using the same relation (24) in (39) produces:

$$\frac{\Delta P_k}{P_k} = \sqrt{\frac{2n}{mtNp(1-p)}} \,. \tag{40}$$

Then similarly to (26), for a given level of significance α , the percent of margin error δ is defined via (40) as:

$$\delta = z_{\alpha/2} \frac{\Delta P_k}{P_k} = z_{\alpha/2} \sqrt{\frac{2n}{mtNp(1-p)}} \,. \tag{41}$$

For the given % of margin error δ , the total sample size is estimated from (41) as follows:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta} \right)^2 \frac{2n}{mtp(1-p)} \,. \tag{42}$$

As in transformation (27) into (29), taking the mean level p=0.5 simplifies the expression (42) giving the lower border for the sample size estimate:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta} \right)^2 \cdot \frac{8n}{mt} \,. \tag{43}$$

Similarly to (31), we can estimate the number of tasks by the given sample size from (43) as

$$t \geq \left(\frac{z_{\alpha/2}}{\delta} \right)^2 \frac{8n}{mN} \,. \tag{44}$$

Another way of simplification of expression (38) is in assuming approximately equal values of P_j in the sum at the right-hand side, and also N_j=N_k as it is in a balanced design, then we get:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k p_k(1-p_k)}} + \sqrt{\frac{1}{2N_k}} \frac{1}{n} \sum_{j=1}^{n} \sqrt{\frac{1}{p_j(1-p_j)}} \,. \tag{45}$$

In assumption of evenly distributed p_j, the summing in (45) is approximated by the integral:

$$\sum_{j=1}^{n}\sqrt{\frac{1}{p_j(1-p_j)}} \approx \int_{0+}^{1-}\frac{dp}{\sqrt{p(1-p)}} = \arcsin(2p-1)\Big|_0^1 = \pi\,. \tag{46}$$

Then using (24) and (46), the expression (45) is reduced to:

$$\Delta P_k / P_k = \sqrt{\frac{1}{2N_k}}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right) = \sqrt{\frac{n}{2mtN}}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right), \tag{47}$$

and similarly to (41) we can write:

$$\delta = z_{\alpha/2}\frac{\Delta P_k}{P_k} = z_{\alpha/2}\sqrt{\frac{n}{2mtN}}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right). \tag{48}$$

From (48), for the given % of margin error δ, the total sample size is estimated as follows:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{n}{2mt}\left(\frac{1}{\sqrt{p_k(1-p_k)}}+\frac{\pi}{n}\right)^2 . \tag{49}$$

For the mean level of the absolute probabilities p_k=0.5, it is reducing to the expression:

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{n}{2mt}\left(2+\frac{\pi}{n}\right)^2 , \tag{50}$$

and for big n, when $2 >> \pi / n$, the formula (50) simplifies to

$$N \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \frac{2n}{mt}. \tag{51}$$

Numerical simulations show that the last formulae (47)–(51) are preferable.

REFERENCES

Alvarez R.M. and Nagler J. (1998) When politics and models collide: Estimating models of multiparty elections, American Journal of Political Science 42, No.1, 55–96.

Finn A. and Louviere J.J. (1992) Determining the appropriate response to evidence of public concern: the case of food safety. J. Public Policy Marketing 11(1), 12–25.

Flynn T. N. (2010) Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling, Expert Review of Pharmacoeconomics and Outcomes Research, 10(3), 259–267.

Gelman A., Carlin J.B., Stern H.S., and Rubin D.B. (2004) Bayesian Data Analysis, Chapman & Hall/CRC, Suffolk.

Hausman J. and McFadden D. (1984) Specification tests for the multinomial logit model, Econometrica, 52, No.5, 1219–1240.

Hensher D.A., Rose J.M., and Greene W.H. (2005) Applied Choice Analysis: A Primer, Cambridge University Press, Cambridge, England.

Lipovetsky S. and Conklin M. (2014a) Best-Worst Scaling in Analytical Closed-Form Solution, The Journal of Choice Modelling, 10, 60–68.

Lipovetsky S. and Conklin M. (2014b) Finding Items Cannibalization and Synergy by BWS Data, The Journal of Choice Modelling, 2014, v. 12, DOI: 10.1016/j.jocm.2014.08.001.

Lipovetsky S. (2014) Analytical closed-form solution for binary logit regression by categorical predictors, J. of Applied Statistics, DOI: 10.1080/02664763.2014.932760

Louviere J.J and Woodworth G.G. (1983) Design and Analysis of Simulated Consumer Choice of Allocation Experiments: An Approach Based on Aggregate Data, Journal of Marketing Research, 20, 350–367.

Louviere J.J. (1993) The Best-Worst or Maximum Difference Measurement Model: applications to behavioral research in marketing. Proceedings of the American Marketing Association’s Behavioral Research Conference, Phoenix, Arizona.

Louviere J.J., Hensher D.A. and Swait J. (2000) Stated Choice Methods: Analysis and Applications, Cambridge University Press, Cambridge.

Louviere J.J.,_Street D., Burgess L., Wasi N., Islam T., and 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), 128–164.

Louviere J.J., Lings I., Islam T., Gudergan S., and Flynn T. (2013) An introduction to the application of (case 1) best-worst scaling in marketing research, International Journal of Research in Marketing, 30, Issue 3, 292–303.

Manski F.C. and McFadden D. (1981) Alternative Estimators and Sample Designs for Discrete Choice, pp. 2–50, in: Manski F.C. and McFadden D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications, Cambridge, MA, The MIT Press.

Orme B. (2009) MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB, Sawtooth Software Research Paper Series, Orlando, FL.

Orme B. (2010) Getting Started with Conjoint Analysis, 2nd edition, Research Publishers LLC, Madison, WI, USA.

Wonnacott T.H. and Wonnacott R.J. (1977) Introductory Statistics, 3rd edition, Wiley, New York.

“Performance, Motivation and Ability”— Testing a Pay-for-Performance Incentive Mechanism For Conjoint Analysis

Philip Sipos
Markus Voeth
University Of Hohenheim

Abstract

To gain the value of incentive-aligned conjoint analysis conjoint studies required the availability of at least one version of the object of research offered as prize. However, there are important fields of application of conjoint analysis in which offering the object of research is not possible. We test a new approach to provide a solution to the dependency of the research object and other current challenges of incentive-aligned conjoint analysis in which respondents are motivated to provide consistent responses. As key results we find that pay-for-performance enhances predictive accuracy, increases cognitive effort and consistency compared to hypothetical study settings. Our findings demonstrate that the pay-for-performance mechanism works well within the context of both low and high ability and yields better results in case of higher knowledge of the product class. Generally, higher knowledge of the product class, or in more general terms, ability, has a substantial impact on consistency, reliability and predictability. For this reason, we conclude, that it is not all about (extrinsic) motivation so as to increase predictive accuracy and further measures of the quality of conjoint study results but also about ability. Thus, with the equation “*Performance = Motivation + Ability”* in mind both main factors should be taken into account by marketing researchers and practitioners when preparing and conducting a conjoint study.

1. Introduction

Incentive-alignment of conjoint studies has become popular in marketing research and practice (e.g., Toubia et al. 2012; Wlömert and Eggers 2014): Since its introduction in marketing literature by Ding et al. (2005) incentive-aligned conjoint analysis has already been established as state of the art (Ding et al. 2011). However, it has to be admitted that “for some contexts incentive-alignment is not easy to accomplish” (Rao 2014, 152), for instance “when it is not cost effective to offer real product to each participant or to generate all product variations” (Agarwal et al. 2015, 28). In brief, the restricted applicability of incentive-aligned conjoint analysis (Ding and Huber 2009) primarily results from two major challenges extant incentive-aligned mechanisms are vulnerable to. First, all mechanisms require the availability of at least one version of the product to be used as incentive denoted as *research object dependency*. Research object dependency due to a certain range of products required is a severe problem in case of preference elicitation in course of the new product development process, which constitutes a key application of conjoint analysis (see also Voeth and Sipos 2015): the object of research may either not yet be available at all or using it as incentive may be impossible due to legal prohibitions. Second, even if real versions of the product are available the application of one of the extant mechanisms may turn out too costly as one cannot reward every respondent with, for example, a yacht or a car describing the *cost issue* challenge. In this case, the costs of conducting market research would exceed the benefits from gaining insights into preferences structures by far. One may argue now that it is reasonable to let respondents participate in a lottery for an expensive product like an automobile using prize indemnity insurances (Ding et al. 2011). However, others argue that a very low probability of being compensated may work against the incentive.

Against the background of current challenges of incentive-aligned conjoint analysis there is a need for an alternative approach. Sipos and Voeth (2014) developed and proposed the idea of a *pay-for-performance* mechanism as a solution to current challenges of incentive-aligned conjoint analysis that has been tested in course of a large-scale empirical study. The mechanism rewards every respondent depending on predictive accuracy in holdout sets. It is intended to increase the motivation to make the cognitive effort required during the conjoint tasks. In doing so, the overall goal of the pay-for-performance mechanism is to extend the fruitful idea of incentive-aligned conjoint analysis to new fields of application and offer an alternative to extant mechanisms in case that they could be applied but it may be too costly to do so. Sipos and Voeth (2014) as well as Voeth and Sipos (2015) have already provided evidence that the pay-for-performance is effective in terms of increasing effort, predictability and performance (i.e., testing whether pay-for-performance works at all). We test the relative influence of motivation generated by incentive-compatibility against ability measured by respondents' self-evaluations concerning their general knowledge of the product class under investigation.

We find that (i) there are no significant performance differences between purely hypothetical conjoint and groups who receive a low fixed reimbursement, (ii) performance-dependent reimbursement seems to be capable of enhancing predictive accuracy compared to conducting conjoint analysis in a hypothetical setting constituting the key finding of our analysis, (iii) the proportion of fixed to variable compensation has little effect on predictive accuracy performance unless one is zero, and (iv) performance-dependent payment proves itself as a satisfactory alternative to the state-of-the-art benchmark we compared our proposed approach.

In line with our purpose, the remainder of this paper is structured as follows: Section 2 provides a short review of research on incentive-aligned choice-based conjoint. Section 3 discusses the basic idea of the pay-for-performance mechanism in some more detail and develops the ability issue. In section 4 we outline the background of our empirical study and present the study results. Section 5 concludes with a short summary of the key findings and provides some implications for marketing research and practice.

2. Review of Research on Incentive Alignment

Due to promising first evidence of its capability of reducing hypothetical bias problem and enhancing predictive accuracy as well as offering "additional motivation to respondents to provide truthful input" (Toubia et al. 2012, 138), a huge body of research on incentive-aligned conjoint analysis has been conducted recently. On the one hand there is substantial research which reaffirms the advantages and adds further evidence in favor of making use of incentive-alignment (e.g., Ding et al. 2009; Miller et al. 2011). On the other hand, alternative mechanisms have been developed in addition to the initial so-called *Direct mechanism* proposed by Ding et al. (2005) such as the *Willingness-To-Pay* mechanism by Ding (2007), the *Rank Order* mechanism by Dong et al. (2010) or the *Progressive Direct* mechanism by Eggers (2012). The mechanisms currently available to incentive-align a conjoint study differ with respect to (i) winning probability, (ii) the range of products required and (iii) whether willingness-to-pay can be estimated. Dong et al. (2010) provide a recommendation framework which mechanism can be applied in which context guided along the two dimensions (ii) and (iii). When applying the Direct mechanism every respondent has a chance to win any item chosen, which may turn out very costly or even impossible. Hence, the Direct mechanism is restricted in its applicability. The Willingness-To-Pay mechanism, however, only requires the availability of one version of the product but requires an estimate of the respondent's willingness-to-pay for it. Modelled after the BDM-mechanism (Becker et al. 1964) a random lottery (i.e., winning probability < 100%) draws one or more respondents whose willingness-to-pay for the product available is derived from the data provided during the conjoint tasks. The derived willingness-to-pay is then compared to a randomly determined price. If derived willingness-to-pay exceeds the randomly determined price the respondent has to buy the product; otherwise he cannot buy the product. The Rank Order mechanism does not depend on the estimation of the willingness-to-pay. However, it requires that at least two versions of the product are available. Respondents participate in a random lottery (i.e., winning probability < 100%) and a rank order of the feasible products is derived from their preference data provided. If a respondent is drawn as winner, he receives the product first in rank according to his preference data as prize. The Progressive Direct mechanism works similar to the Direct mechanism by Ding et al. (2005). Contrary to the Direct mechanism, the winning probability is below 100%. However, respondents can increase their probability of winning by voluntarily answering more choice sets: All answered choice sets of all respondents are pooled. Out of this pool one or more choice sets are randomly drawn and the winner(s) receive(s) the choice made in the respective choice set as prize. In doing so, the Progressive Direct mechanism increases respondents' motivation to provide more preference data, which in turn may result in better parameter estimates.

3. Pay-for-Performance, Motivation and Ability

The improvements in predictive accuracy by incentive-aligning conjoint studies justified increasing use of incentive mechanisms. But until now it only worked when offering the prizes was feasible. Due to this restriction we propose the pay-for-performance incentive concept as an attempt to increase the accuracy of conjoint by rewarding every respondent with money depending on his individual performance in a certain task. In general, applying a pay-for-performance approach within a conjoint study context requires a performance indicator to assess a respondent's individual performance. For researchers as well as practitioners the overall goal is to collect high quality data. The quality of conjoint data of course depends on the study participant's response behavior and can be assessed by a variety of validity and reliability measures (see e.g., Green and Srinivasan 1990). As far as predictive accuracy evaluation is concerned measuring first choice hit rates in holdout tasks is considered to be state of the art in conjoint studies (Ding et al. 2005). Therefore, in the context of our study Pay-for-Performance involves reimbursing respondents proportionately to the number of first choice hits in holdout tasks. This allows us to characterize the performance-dependent reimbursement scheme with $m_{\text{pay-for-perfromance}}$ as reimbursement paid to the respective respondent, τ as individual performance parameter counting first choice hits, $p_{\text{pay-for-performance}}$ as monetary reward per first choice hit and $\beta_{\text{pay-for-performance}}$ as variable reimbursement parameter: $m_{\text{pay-for-performance}}(\tau) = \beta_{\text{pay-for-performance}} * \tau * p_{\text{pay-for-performance}}$.

In general, applying the pay-for-performance concept as a tool to cause extrinsic motivation is expected to result in enhanced consistency, reliability and predictive accuracy by increasing cognitive effort and decision time per choice set. One may argue now that it is not all about motivation defined as willingness to make the required cognitive effort during a conjoint study to increase performance but also about ability as a general requirement to be capable of making that cognitive effort. Embedded into information integration theory, cognitive algebra models show that judgments on performance are influenced by information given regarding degrees of motivation and ability suggesting a multiplicative relation such as "*Performance = Motivation * Ability*" (Anderson and Butzin 1974). Therefore, we decided to take ability as a further performance impact factor into account by having respondents evaluate their knowledge of the product class under investigation (low vs. high). Combining low and high motivation depending on whether the nature of the task is hypothetical or non-hypothetical and low and high ability results in a 2 x 2 ability-motivation-performance-matrix with different performance expectations illustrated in Figure 1. To analyze the effect of our supposed impact factors on performance we particularly focused on the main effects in our study i.e., in formal terms "*Performance = Motivation + Ability*."

Figure 1: Ability-Motivation-Performance Matrix

Extrinsic Motivation **Ability**	**low**	**high**
low	low performance	medium performance
high	medium performance	high performance

4. Empirical Study

Study Background

To test the pay-for-performance mechanism in the context of varying degrees of ability we referred to study data used to assess general suitability to use pay-for-performance mechanisms in conjoint studies (Voeth and Sipos 2015). Data was collected in a large-scale lab-based empirical study distinguishing between five experimental treatments in which respondents were randomly assigned to different reimbursement conditions. First, there is a control group denoted as "hypothetical" in which respondents did not receive any payment at all. For the control group the motivation to participate is largely of intrinsic nature such as feeling good when supporting research projects. Second, we compare the pay-for-performance mechanism with the Rank Order mechanism by Dong et al. (2010). In the third treatment respondents received a fixed payment regardless of their performance in the holdout sets. We further distinguished between lower and higher fixed reimbursement with two subgroups of fixed reimbursement with a guaranteed payment of either 10 EUR or 16 EUR on average respectively. Fourth, we had a purely variable reimbursement group in which respondents received between 0 EUR and 16 EUR depending on the number of first choice hits in four holdout tasks serving as performance indicators. Thus, a first choice hit was rewarded with 4 EUR each. Fifth, in a mixed group respondents received some fixed amount of money (8 EUR) and in addition to that they had the chance to win up to 8

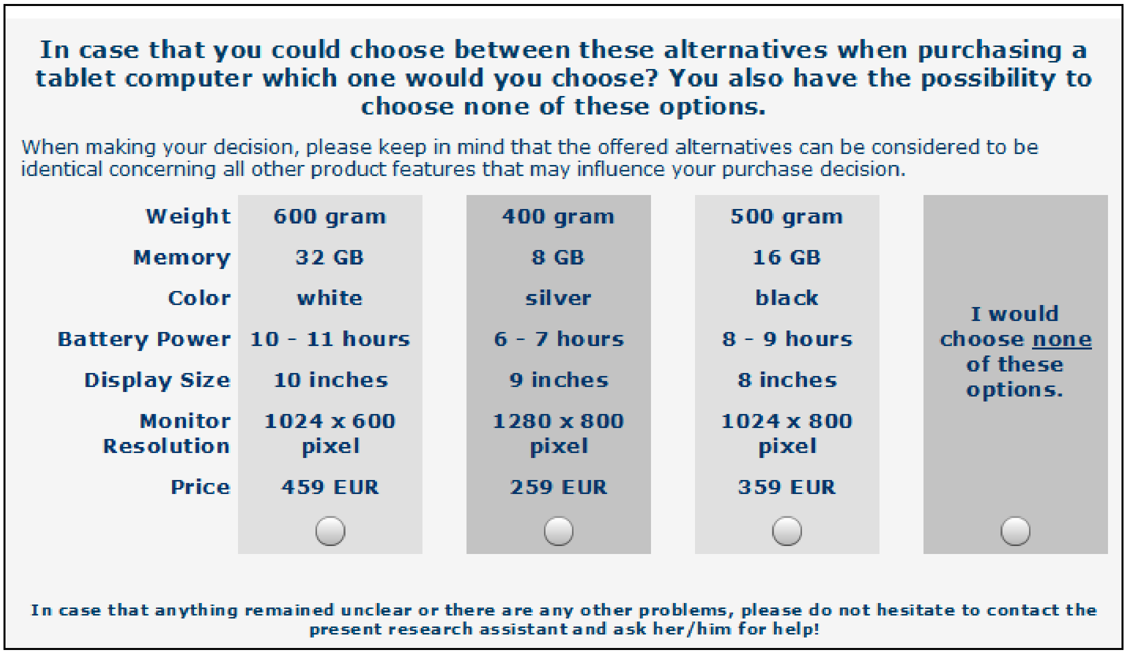
EUR depending on the number of first choice hits which means that each first choice hit was rewarded with 2 EUR each.

Differing between groups with fixed and variable (i.e., performance-dependent) reimbursement components is supposed to take two constraints into account that, according to economic literature (e.g., Spremann 1987), needed to be met to make a pay-for-performance mechanism work. First, the so-called "participation constraint" has to be met in order to increase a respondent's willingness to participate in the study at all which has mainly to do with reservation wage considerations. In our case, the fixed payment is intended to satisfy this constraint. However, meeting the participation constraint does not guarantee that the respondent is willing to make a cognitive effort during the conjoint task. Therefore, a second constraint denoted as "incentive compatibility constraint" has to be met by connecting payment to some observable performance indicator which, e.g., can done through the variable reimbursement component depending on individual predictive accuracy in holdout tasks. These two constraints in mind allow us to extend the purely performance-dependent reimbursement scheme to a more general payment model (see also Voeth and Sipos 2015): $m_{\text{pay-for-performance}}(\tau) = \alpha_{\text{pay-for-performance}} + \beta_{\text{pay-for-performance}} * \tau * p_{\text{pay-for-performance}}$ with $\alpha_{\text{pay-for-performance}}$ as an added fixed reimbursement parameter.

The experimental design, considering different theoretical requirements to make a pay-for-performance approach work, enables us to answer four research questions.

i. Are there overall and ability-specific performance differences in predictive accuracy between a purely hypothetical and a purely fixed reimbursement setting?
ii. Are there overall and ability-specific performance differences in predictive accuracy between performance-dependent groups (variable and mixed) and purely hypothetical/fixed groups?
iii. Are there overall and ability-specific performance differences in predictive accuracy differences between the performance-dependent reimbursement groups themselves?
iv. Are there overall and ability-specific performance differences in predictive accuracy between the performance-dependent reimbursement groups and the Rank Order group?

Figure 2: Running the Study: Random Choice Set Example (Translated from German)



In case that you could choose between these alternatives when purchasing a tablet computer which one would you choose? You also have the possibility to choose none of these options.

When making your decision, please keep in mind that the offered alternatives can be considered to be identical concerning all other product features that may influence your purchase decision.

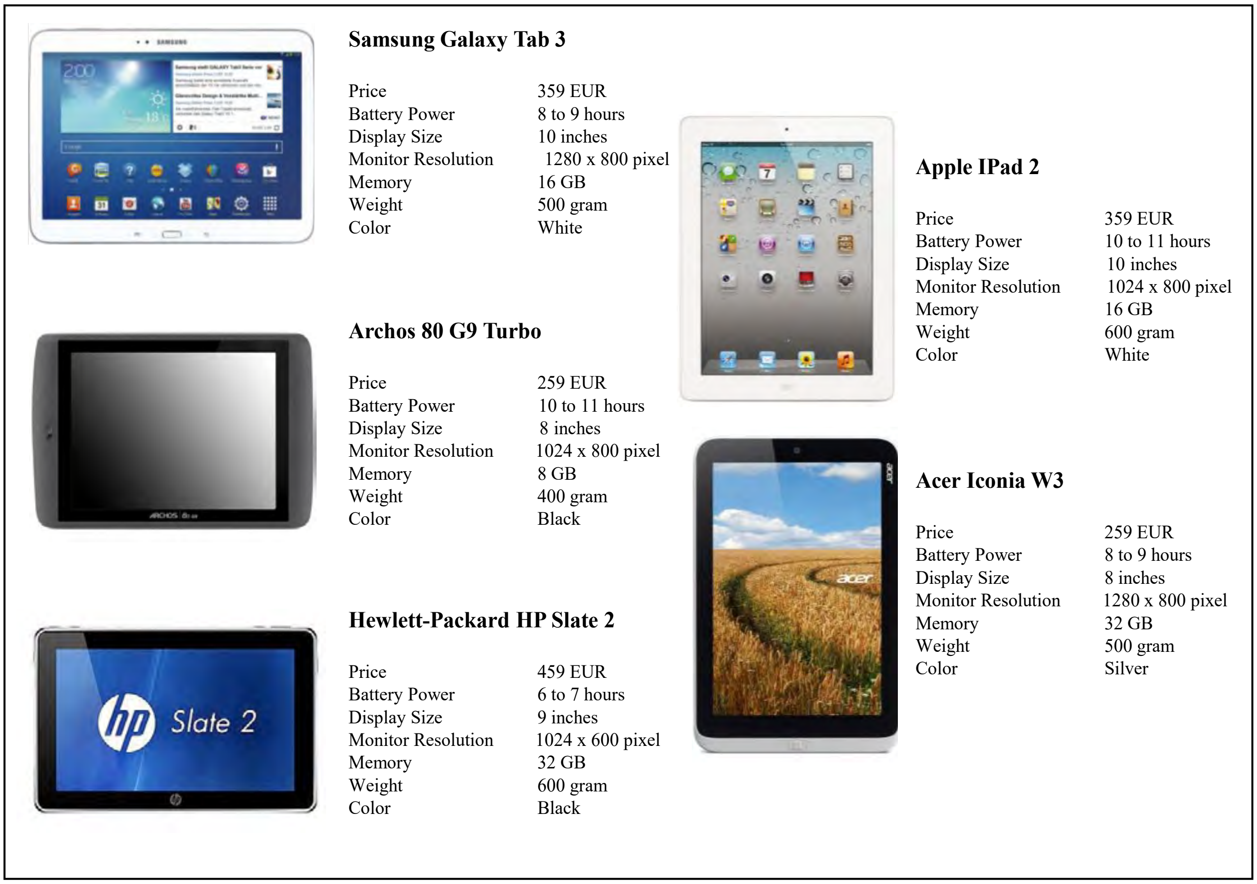
Weight	600 gram	400 gram	500 gram	I would choose none of these options.
Memory	32 GB	8 GB	16 GB	
Color	white	silver	black	
Battery Power	10 - 11 hours	6 - 7 hours	8 - 9 hours	
Display Size	10 inches	9 inches	8 inches	
Monitor Resolution	1024 x 600 pixel	1280 x 800 pixel	1024 x 800 pixel	
Price	459 EUR	259 EUR	359 EUR	
	○	○	○	○

In case that anything remained unclear or there are any other problems, please do not hesitate to contact the present research assistant and ask her/him for help!

Tablet computers served as a research object, described by seven features with three levels each (see also Figure 2) identified in course of a short pre-study with 50 students. All respondents were provided with the following information regardless of group membership: We told them that they were about to buy a tablet computer in an electronic market store and that they could choose between the three alternatives shown or indicate that they would not purchase a tablet in the respective choice round. In case their purchase decision was influenced by a feature not shown in the choice tasks respondents should consider all alternatives to be identical with respect to this/these feature(s). Furthermore, respondents received additional group-specific information regarding their reimbursement (except for the hypothetical group): In the Rank Order group respondents were told that on top of 6 EUR as fixed reimbursement they had a 1/25 chance to win a tablet computer. After the conjoint task we revealed five different tablet computers (see Figure 3). To encourage respondents to pay the same attention to price that they would in a real purchase situation, a coin flip would decide whether the randomly drawn winner received 465 EUR to buy a choice made in one of the holdout sets (respondents did not know which of the choice sets were random and which were fixed as all choice sets, regardless of type, consisted of three alternatives plus the none) or whether they received the top ranked product according to their data provided. The winning respondent(s) would receive the tablet first in rank according to a preference ranking of the five tablets derived from his/her preferences data provided in the conjoint tasks. In the fixed group respondents were told that the respective (lower or higher) fixed amount of money was paid to them. In the performance-dependent groups respondents were informed that the amount of reimbursement increased in linear terms with higher quality of their response behavior, evaluated by means of a statistical measure on five levels ranging from 0 = very poor, 1 = poor, 2 = satisfactory, 3 = good to 4 = very good (which can be equated with the number of holdout first choice hits). We further tested whether respondents understood the way of being reimbursed by means of a simple manipulation check having respondents select the reimbursement description that in their opinion corresponds to the reimbursement offered to them. In case a respondent failed the manipulation check he/she was eliminated from successive data analysis. In order to account for the ability factor, respondents were asked to self-evaluate their general knowledge of the product class with the following question: “Compared to your friends how do you evaluate your general knowledge concerning tablet computers?” That question generated low and high product knowledge categories.

SSI Web Lab generated randomized full-profile designs according to the complete enumeration procedure, meeting requirements for efficient design such as level balance, minimal overlap and orthogonality (Huber and Zwerina 1996). The choice task consisted of 16 random choice sets for each respondent and five fixed choice sets (holdouts), four of which were used to assess predictive accuracy and determine the amount of payment in performance-dependent groups. The remaining one (identical to one of the four holdouts) served as test-retest reliability check to assess consistency of the response behavior. In addition, we analyzed group-specific average decision times per choice set as a further measure of cognitive effort besides predictive accuracy.

Figure 3: Applying the Rank Order Mechanism: Versions of the Product Available



Samsung Galaxy Tab 3

Price	359 EUR
Battery Power	8 to 9 hours
Display Size	10 inches
Monitor Resolution	1280 x 800 pixel
Memory	16 GB
Weight	500 gram
Color	White

Apple IPad 2

Price	359 EUR
Battery Power	10 to 11 hours
Display Size	10 inches
Monitor Resolution	1024 x 800 pixel
Memory	16 GB
Weight	600 gram
Color	White

Archos 80 G9 Turbo

Price	259 EUR
Battery Power	10 to 11 hours
Display Size	8 inches
Monitor Resolution	1024 x 800 pixel
Memory	8 GB
Weight	400 gram
Color	Black

Acer Iconia W3

Price	259 EUR
Battery Power	8 to 9 hours
Display Size	8 inches
Monitor Resolution	1280 x 800 pixel
Memory	32 GB
Weight	500 gram
Color	Silver

Hewlett-Packard HP Slate 2

Price	459 EUR
Battery Power	6 to 7 hours
Display Size	9 inches
Monitor Resolution	1024 x 600 pixel
Memory	32 GB
Weight	600 gram
Color	Black

The entire student sample amounts to 541 respondents that are nearly distributed among the different reimbursement groups (see also Table 1). Part-worth utilities were estimated by means of the Hierarchical Bayes procedure (e.g., Lenk et al. 1996).

Table 1: Sample Distribution

Motivation (reimbursement group)	**Ability** (self-evaluation of knowledge of the product class)		**Total**
	Low	**High**	
Hypothetical	19	80	99
Rank Order	18	76	94
Fixed_low	18	77	95
Fixed_high	8	40	48
Variable	12	92	104
Mixed	16	85	101
Total	91	450	541

Study Results

We tested the effectiveness of the pay-for-performance mechanism referring to performance measures indicating consistency (i.e., reliability and predictive accuracy) as well as cognitive effort (i.e., average decision times and predictability). With reference to the ability-motivation-performance matrix depending on the hypothetical nature of the task (yes or no), the hypothetical and fixed_i (i = low, high) groups are considered to be low extrinsic motivation groups whereas the remaining groups in which decisions are connected to real consequences constitute high extrinsic motivation groups. Thus, in total we differ between group-specific results of twelve cells (6 motivation categories/groups * 2 ability categories). Descriptive results are reported in Table 2 as well as Figures 4 to 6.

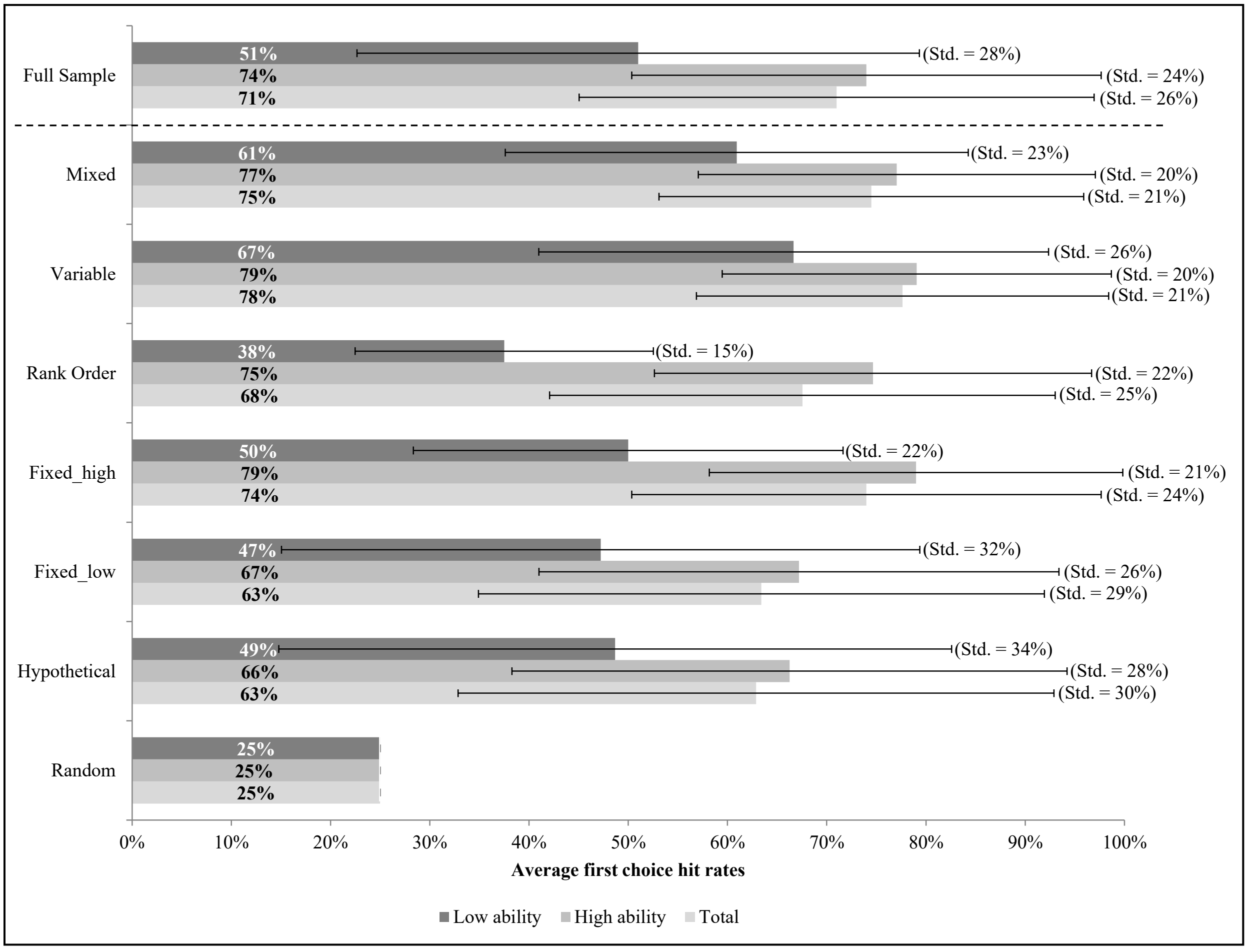
Table 2: Group-Specific Average Decision Times Per Choice Set (in Seconds) and Standard Deviations

Motivation	**Reimbursement group**	**Ability** (self-evaluation of knowledge of the product class)				**Total**	
		Low		**High**			
		Ø	Std.	Ø	Std.	Ø	Std.
Low	**Hypothetical**	18.8	18.6	18.8	20.2	18.8	18.9
	Fixed_low	17.4	13.6	17.6	15.7	17.6	15.3
	Fixed_high	16.9	13.0	18.3	17.2	18.0	16.6
High	**Rank Order**	23.2	22.9	19.8	19.8	20.4	20.4
	Variable	20.0	15.5	22.8	21.4	22.5	20.8
	Mixed	22.2	16.7	21.0	18.4	21.2	18.1
Total		20.0	18.0	20.1	19.0	20.1	18.8

The group-specific total average decision times indicate higher decision times in groups with a non-hypothetical setting (i.e., Rank Order, Variable and Mixed), whereas in the hypothetical and purely fixed groups in which the nature of the task remains hypothetical overall average decision times per choice set are lower. A Kruskal-Wallis test confirms overall significant differences between group-specific average decisions times (p =.000). Successive multigroup comparisons conducted by SPSS indicate that significant differences can be found between Hypothetical vs. Variable (p =.002), Hypothetical vs. Mixed (p =.050), Fixed_low vs. Variable (p =.006), Fixed_low vs. Mixed (p =.000), Fixed_high vs. Variable (p =.008) as well as Fixed_high vs. Mixed (p =.089). Thus, cognitive effort tends to be higher in non-hypothetical settings. Moreover, ability appears to have no effect on average decision times spent on a choice set, as some groups show higher decision times in the low ability case than in the high ability category (Rank Order and Mixed), whereas others show lower decision times (Fixed_high and Variable) when knowledge of the product class is low, or (almost) equal decision times (Hypothetical and Fixed_low) independent from product knowledge.

As already shown by Voeth and Sipos (2015) predictability in performance-dependent groups outperforms hypothetical settings which allows us to conclude that the pay-for-performance mechanism is effective in terms of predictive accuracy enhancement. The Rank Order group shows an overall average performance which is slightly better than the hypothetical and lower fixed groups but a little bit worse than the performance-dependent groups. Furthermore, Voeth and Sipos (2015) have discussed the unexpected overall (i.e., not differing between low and high ability) results of the high fixed reimbursement group which reveals some cost reduction potential for marketing researchers and practitioners by applying a pay-for-performance approach such that the pay-for-performance mechanism not only turns out as effective but also as cost efficient. Performance in the high fixed reimbursement and the performance-dependent groups is similar with respect to predictive accuracy of about 75%, which means three out of four holdouts are predicted correctly per respondent on average. Replacing at least some amount of the high fixed reimbursement by a variable reimbursement component decreases the payoff from 16 EUR in the high fixed reimbursement group to an average (at least) partially performance-dependent payoff between 12 EUR in the purely variable group or 14 EUR in the mixed group while maintaining high predictive accuracy at the same time.

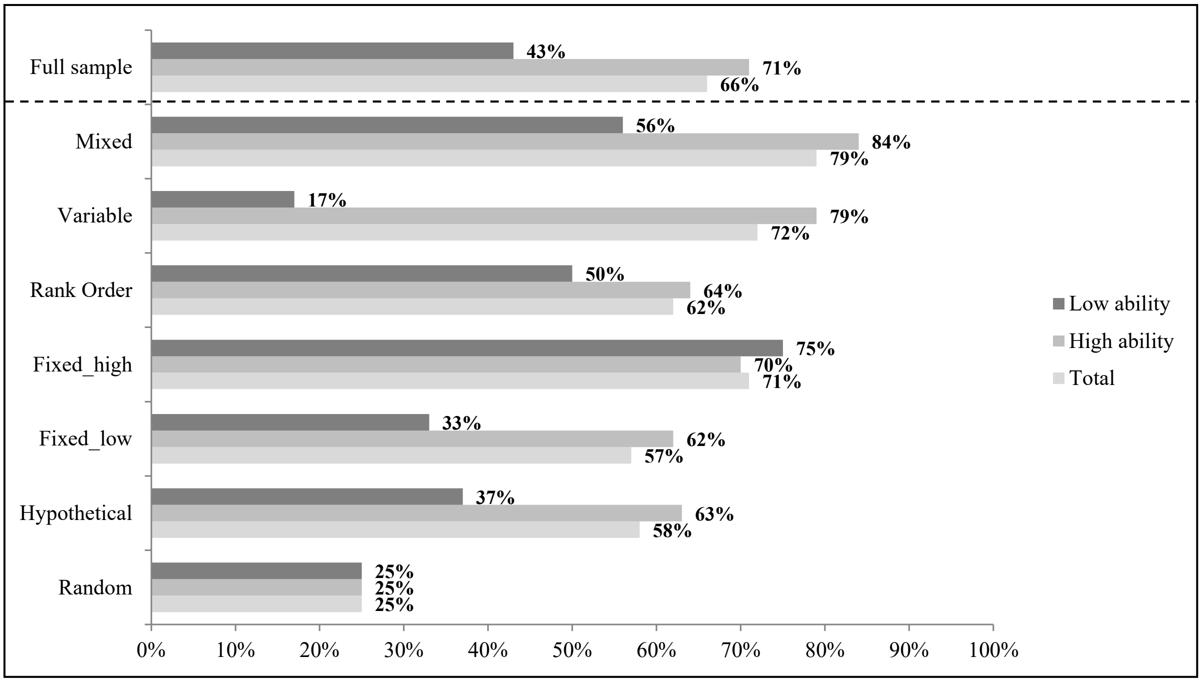
**Figure 4: Predictive Accuracy—
Group-Specific Average First Choice Hit Rates (Four Holdout Tasks)**



In a second step, we now take ability into account by making an "ability split," referring to the respondents' self-evaluation of their knowledge of the product class. In particular, that is the case for the Rank Order group, which performs as well as the performance-dependent groups when product knowledge is high but worse than the hypothetical and lower fixed groups when product knowledge is low. Low product knowledge may go along with a low interest in the product category resulting in a low desire to win a tablet computer or a nonexistent need to own one at all. For this reason, the willingness to make a cognitive effort may not be enhanced by having the chance of winning a tablet computer as long as the respondent does not have a current or future need to own one. Even though first choice hit rates are consistently lower in the low than in the high ability category, the pay-for-performance mechanism turns out to be effective in both ability categories. In the low and high ability cases the hypothetical/lower fixed groups are outperformed by the performance-dependent groups with respect to predictability of holdout choices. Moreover, performance differences between the performance-dependent groups are higher in the low ability category than in the high ability category. Hence, high ability seems to be helpful in making results of performance-dependent groups more stable. It is also noteworthy that the high performance of the high fixed group mainly goes back to the high ability category. In the low ability category the high fixed group shows a performance similar to the hypothetical and low fixed groups serving as further proof of the high impact of ability on performance.

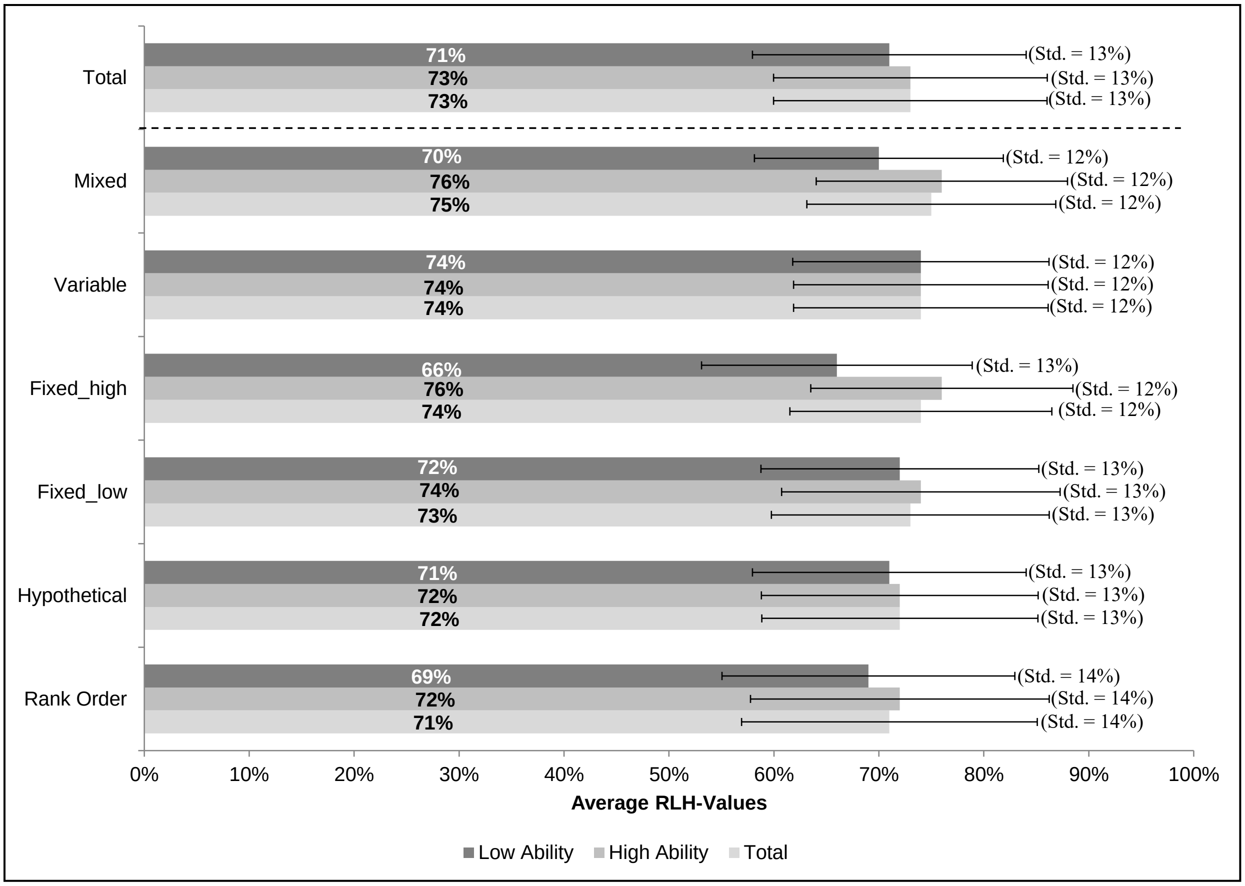
Furthermore, we evaluated test-retest reliability by means of duplicating one of the four holdout sets used to determine reimbursement in the performance-dependent groups. Reliability could be assessed by analyzing whether a respondent made the same choice in those two identical choice sets. The results reported in Figure 5 show a similar picture as the predictive accuracy results repeating the found main effects of ability and motivation on performance: performance-dependent reimbursement yields a higher consistency in response behavior than conducting a conjoint analysis in a hypothetical or lower fixed reimbursement setting. The high fixed group again performs as well as the performance-dependent groups. Regarding the Rank Order group reliability values are rather comparable to the hypothetical and lower fixed groups than to the performance-dependent groups. Average reliability is higher in the high ability category than in the low ability category except for the fixed high ability group showing surprising and unexpected results once again.

Figure 5: Test-Retest Reliability Group Specific Average Hit Rates



In addition, we reported overall and ability-specific RLH measures as an index of fit which can be found in Figure 6. As opposed to predictive accuracy and test-retest reliability we do not observe wide within-group ability-specific disparities except for the high fixed group.

Figure 6: Index of Fit: Average RLH Values



In sum, our expectations with respect to the ability-motivation-performance matrix are mostly met: low ability and low extrinsic motivation result in low performance whereas the combination of high ability and motivation yield the highest performance. The combinations of low ability/high motivation and vice versa result in a similar average performance. Regarding the four research questions the results outlined above allow the following conclusions:

i. There are only minor overall differences between the hypothetical and low fixed group, which is in line with our expectations. We did not expect higher fixed reimbursement to outperform the lower fixed and hypothetical groups since the nature of the task remains hypothetical regardless of lower or higher fixed reimbursement. However, the surprising results can only be found in the high ability case and thus may be due to statistical anomalies that need to be retested as well as replicated in future studies. If knowledge of the product class is low there are no significant differences regarding predictive accuracy between the high fixed and the lower fixed as well as the hypothetical group.

ii. Performance-dependent reimbursement seems to work in terms of predictive accuracy enhancement compared to the hypothetical and low fixed groups. It has to be admitted that against the background of the good performance of the high fixed group in the high ability case, performance-dependent reimbursement does not necessarily yield better results than fixed reimbursement. However results turn out to be more stable and thus less volatile which is true for both the low and the high ability case.

iii. The two performance-dependent groups, variable and mixed, show similar overall and ability-specific results. From this it follows that the effectiveness of the pay-for-performance mechanism does not depend on the amount of money paid for increased predictive accuracy, but it is rather important that the reimbursement scheme at least contains a partially performance-dependent reimbursement component supporting the general idea of applying pay-for-performance mechanisms in conjoint studies.

iv. Performance-dependent reimbursement turns out to be a suitable alternative to the Rank Order mechanism. When respondents are confident in their product class knowledge the Rank Order mechanism and the pay-for-performance mechanism perform equally well. If the product under investigation may be an expensive one such as a yacht or a luxury car, for which reason the application of the Rank Order mechanism may be very costly, the pay-for-performance mechanism can be used. In case the expected knowledge of the product class is low we recommend making use of the pay-for-performance mechanism as we did not observe high predictive accuracy of the Rank Order mechanism. However, it has to be acknowledged that due to small sample sizes in the low knowledge category results need further verification within future studies to find evidence that we did not simply observe a statistical anomaly that may not replicate.

5. Summary and Implications

Though incentive-aligned conjoint analysis constitutes a promising tool to increase predictive accuracy in conjoint studies, up to now it required that offering the prizes be feasible. We tested Sipos and Voeth's (2014) idea of a performance-dependent approach denoted as pay-for-performance mechanism by rewarding respondents on a monetary basis depending on statistical consistency. We show that such a mechanism increases the cognitive and predictive consistency of the conjoint choices in both cases of low and high knowledge of the product class whereby performance is consistently higher when product knowledge is high.

We also found that the effectiveness of the pay-for-performance mechanism does not depend on the amount offered for predictive accuracy, suggesting that the proportion of fixed to variable compensation has little effect unless one is zero. However, we did find that category knowledge was very important in terms of accuracy. Since product knowledge is so important for reliable conjoint choices it is important to identify and make use of strategies that are supposed to ensure higher product knowledge. That can be done by (i) choosing respondents who turn out to be more knowledgeable in general, (ii) choosing respondents who may have a current need for the product under investigation and thus would welcome information concerning that product class, (iii) framing the study context in a way that evokes a respondent's current need in the respective product class and/or (iv) giving detailed information on the product attributes used to describe the object of research. We particularly consider (i) and (ii) to be of major importance as a general requirement of incentive-aligned conjoint analysis to achieve "real" incentive-alignment, be it in case of research object dependent or performance-based monetary reimbursement. Without a current or future need for the product category, respondents may feel less forced to justify their decisions and may care less about the consequences of their choice behavior; for which reason especially a preselection of respondents depending on the existence of a real need for the product category under investigation, may be very helpful to enhance truthfulness of respondents' response behavior. Therefore, we conclude by strongly encouraging future research activities on this issue. For example, greater interest could be evoked by creating a realistic scenario in which respondents who own a tablet computer are told that their tablet computer was destroyed irretrievably. Respondents who do not own a tablet computer could be asked to think about reasons why to purchase one.

Philip Sipos

Markus Voeth

REFERENCES

Agarwal J, DeSarbo WS, Malhotra NK, Rao VR (2015) “An Interdisciplinary Review of Research in Conjoint Analysis: Recent Developments and Directions for Future Research” *Costumer Needs and Solutions* 2(1):19–40.

Anderson NH, Butzin CA (1974) “Performance = Motivation x Ability: An integration-theoretical analysis” *Journal of Personality and Social Psychology* 30(5):598–604.

Becker GM, Degroot MH, Marschak J (1964) “Measuring utility by a single-response sequential method” *Behavioral Science* 9(3):226–232.

Ding M (2007) “An Incentive-Aligned Mechanism for Conjoint Analysis” *Journal of Marketing Research* 44(2):214–223.

Ding M, Huber J (2009) “When Is Hypothetical Bias A Problem In Choice Tasks, And What Can We Do About It?” in: *Proceedings of the Sawtooth Software Conference*, Florida 2009: 263–272.

Ding M, Grewal R, Liechty J (2005) “Incentive-Aligned Conjoint Analysis” *Journal of Marketing Research* 42(1):67–82.

Ding M, Park YH, Bradlow ET (2009) “Barter Markets for Conjoint Analysis” *Management Science* 55(6):1003–1017.

Ding M, Hauser JR, Dong S, Dzyabura D, Yang Z, Su C, Gaskin SP (2011) “Unstructured Direct Elicitation of Decision Rules” *Journal of Marketing Research* 48(1):116–127.

Dong S, Ding M, Huber J (2010) “A simple mechanism to incentive-align conjoint experiments” *International Journal of Research in Marketing* 27(1):25–32.

Eggers F (2012) “Would you like some more? Voluntary choice sets and progressive incentive alignment in conjoint analysis” in: *AMA Winter Educators’ Conference Proceedings*, Chicago 2012:381–382.

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

Huber J, Zwerina K (1996) "The Importance of Utility Balance in Efficient Choice Designs" Journal of Marketing Research 33(3):307–317.

Lenk PJ, DeSarbo WS, Green PE, Young MR (1996) "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs" Marketing Science 15(2):173–191.

Miller KM, Hofstetter R, Krohmer H, Zhang ZJ (2011) "How Should Consumers' Willingness to Pay Be Measured? An Empirical Comparison of State-of-the-Art Approaches" Journal of Marketing Research 48(1):172–184.

Rao VR (2014) "Applied Conjoint Analysis" Springer, Heidelberg et al.

Sipos P, Voeth M (2014) "It's Up to You!—Pay for Performance Mechanisms in Conjoint Studies" in: Proceedings of the 43rd European Marketing Association Conference, Valencia 2014.

Spremann K (1987) "Agent and Principal" in: Agency Theory, Information, and Incentives, Berlin 1987:3–38.

Toubia O, de Jong MG, Stieger D, Fueller J (2012) "Measuring Consumer Preferences Using Conjoint Poker" Marketing Science 31(1):138–156.

Voeth M, Sipos P (2015) "A pay-for-performance mechanism as solution to research object dependency and other challenges of incentive-aligned conjoint analysis" Working Paper.

Wlömert N, Eggers F (2014) "Predicting new service adoption with conjoint analysis: external validity of BDM-based incentive-aligned and dual-response choice designs" Marketing Letters DOI 10.1007/s11002-014-9326-x.

PERCEPTUAL CHOICE EXPERIMENTS: ENHANCING CBC TO GET FROM WHICH TO WHY

BRYAN ORME
SAWTOOTH SOFTWARE, INC.

INTRODUCTION

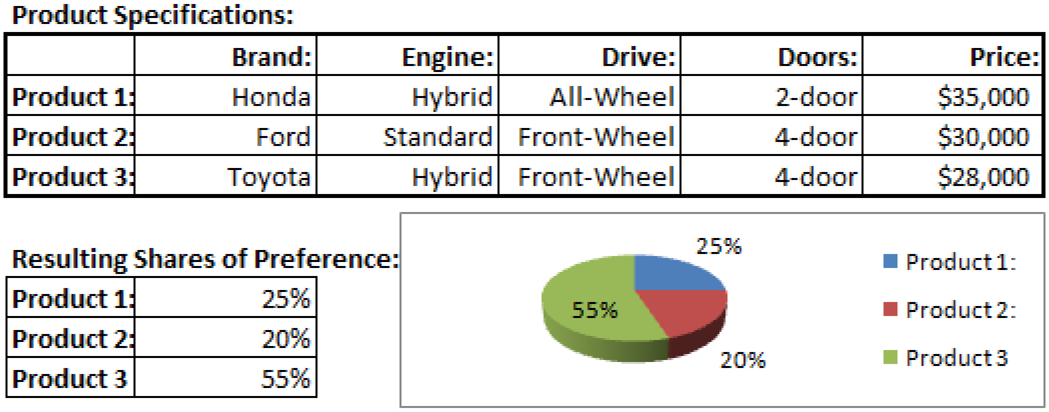
CBC (Choice-Based Conjoint) choice simulators predict share of choice for product concepts within competitive market scenarios, but they provide no insights into perceptions—the *why's* behind the choice. We introduce *Perceptual Choice Experiments* as an extension of CBC questionnaires for integrating diagnostic perceptual dimensions into CBC analysis and simulators. With Perceptual Choice Experiments pick-any agreement questions (on perceptual dimensions) are the dependent variables and traditional conjoint attributes (using standard CBC experimental designs) are the independent variables. The perceptual questions may be added beneath traditional CBC questions, or may be done as separate series of questions.

For years, researchers have investigated perceptual dimensions via additional batteries of brand by attribute questions, such as "How much do you agree or disagree that <Honda> is <A Safe Vehicle to Drive>?" Such sequences often measure two or more brands on several dimensions. The data may be analyzed via tables of means (with means often displayed graphically via line charts or heat maps) or perceptual maps (e.g., correspondence analysis, discriminant analysis, biplot). However, these approaches focus only on brand perceptions. Aspects other than brand make up a product offering. Those additional non-brand attributes certainly affect perceptions, attitudes, influence usage occasions, and resonate with motivations.

Perceptual Choice Experiments add pick-any agreement questions to the CBC questionnaire. Given the data, we estimate weights for the conjoint attribute levels (brands as well as levels of other attributes) to predict for any product concept the likelihood that respondents would agree that it is described by different perceptual/motivational dimensions. Using those weights, the researcher can specify a product concept within a choice simulator (e.g., Honda, hybrid engine, all-wheel drive, 2-doors, $35,000) and predict the likelihood that respondents would agree that this particular Honda specification is *A Safe Vehicle to Drive*, *Good for the Environment*, *A Good Value for the Money*, *A Car I Want to Be Seen in*, etc. A simulator enhanced with such data not only does the standard work of predicting *which* products respondents prefer (share of choice, see Exhibit 1), but also provides insights into *why* they prefer each one (see Exhibit 2)[1]. Given such a tool, the researcher could conduct sensitivity analysis (for a given product concept in the simulator, holding the competitive concepts constant, changing each conjoint attribute level-by-level) to see how each attribute level contributes to perceptions on the diagnostic dimensions.

[1] Fictitious data for illustration only presented in Exhibits 1 & 2.

Exhibit 1: A Standard Market Simulator Interface and Output



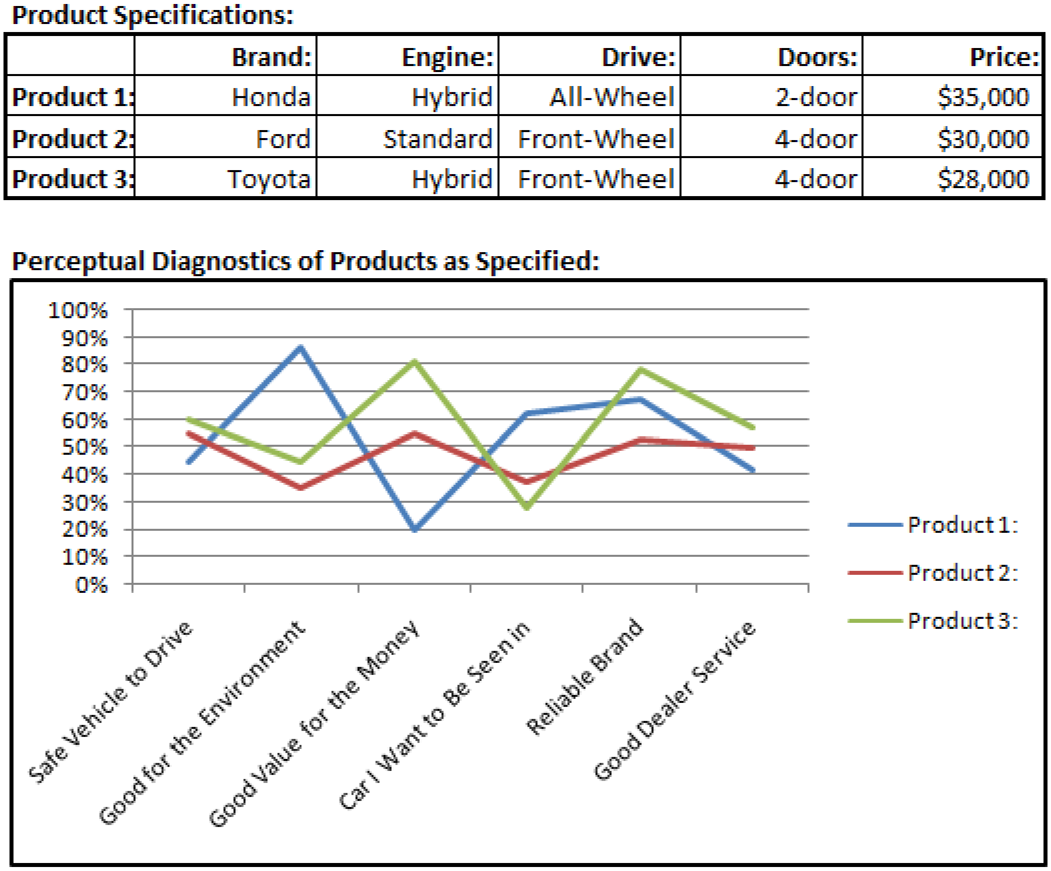
Product Specifications:

	Brand:	Engine:	Drive:	Doors:	Price:
Product 1:	Honda	Hybrid	All-Wheel	2-door	$35,000
Product 2:	Ford	Standard	Front-Wheel	4-door	$30,000
Product 3:	Toyota	Hybrid	Front-Wheel	4-door	$28,000

Resulting Shares of Preference:

Product 1:	25%
Product 2:	20%
Product 3	55%

Exhibit 2: Additional Perceptual Diagnostic Output



Product Specifications:

	Brand:	Engine:	Drive:	Doors:	Price:
Product 1:	Honda	Hybrid	All-Wheel	2-door	$35,000
Product 2:	Ford	Standard	Front-Wheel	4-door	$30,000
Product 3:	Toyota	Hybrid	Front-Wheel	4-door	$28,000

Perceptual Diagnostics of Products as Specified:

Questionnaire Appearance

To extend CBC for Perceptual Choice Experiments, we recommend asking additional follow-up perceptual questions beneath Best-Worst (BW) CBC questions (Exhibit 3).

Exhibit 3: Questionnaire Appearance

	Which of the following would you be most and least likely to purchase?		
	Product A:	Product B:	Product C:
	Brand C Red Package Style 1 Performance Level 3 Price Level 2	Brand A Green Package Style 3 Performance Level 1 Price Level 3	Brand B Yellow Package Style 2 Performance Level 2 Price Level 1
Most Likely to Buy:	○	○	○
Least Likely to Buy:	○	○	○

Which of the following descriptions describe or apply to these products above?

	Product A:	Product B:	Product C:
Statement 8:	□	□	□
Statement 2:	□	□	□
Statement 5:	□	□	□
	□ None describe Package A	□ None describe Package B	□ None describe Package C

The perceptual items shown at the bottom left of Exhibit 3 are randomized and a subset (for example, three items) are shown in each task. To facilitate this, one could use SSI Web's randomized lists capability—or better yet, an experimental plan generated by Sawtooth Software's MaxDiff designer. Showing a subset of the items avoids the burden of too many statements to consider each time (which could bias the selection rate of associations downward). It also makes it possible (with very large sample sizes) to include a long list of perceptual statements without overwhelming any one respondent.

The perceptual statements could be claims, occasions, motivations, perceptual adjectives, etc. developed based on the researcher's expertise as well as upfront qualitative research. Examples:

- A good value
- A product I'd use on the weekends
- Something my mother would buy
- A product I'd tell my friends I was using
- Modern

ANALYSIS OF PERCEPTIONS

The perceptual choice experiment as displayed in Exhibit 3 leverages the attribute list and experimental design of a standard CBC and involves a pick-any association task for a list of perceptual items. For the CBC portion of the task, we may estimate part-worth utilities via HB, leading to the usual market simulator that predicts choice probabilities for product concepts in competitive scenarios.

For the perceptual pick-any questions, the data are typically too sparse at the individual level to use HB. Thus, we suggest employing aggregate logit analysis for each of the perceptual items (a separate binary logit model per item). The conjoint attributes serve as independent variables (effects- or dummy-coded), plus an additional constant to capture the utility of the “not selected” alternative (equivalent to the “None” in CBC). For each perceptual item, the dependent variable is the choice of the item or not (binary logit setup, with two alternatives per choice task). With large enough sample size, interaction effects could be specified. As an example of an interaction effect, a vacation package might be viewed as *a good value* only if it had a low cost per person together with a longer duration in terms of number of nights.

The market simulator may be built in Excel, with shares of choice estimated as usual using the CBC part-worths and the additive, logit rule (or other variants such as RFC). For each product concept in the simulator, we may also use the logit rule (with the part-worth perceptions estimated from a series of independent binary logit models predicting the choice of diagnostic perceptual statements) to predict the percent of respondents who would check the box for each perceptual item about that specific product. Those results could be shown as a Line Chart (Exhibit 4) or as a Heat Map (Exhibit 5).

Exhibit 4: Line Chart for Diagnostics/Perceptions

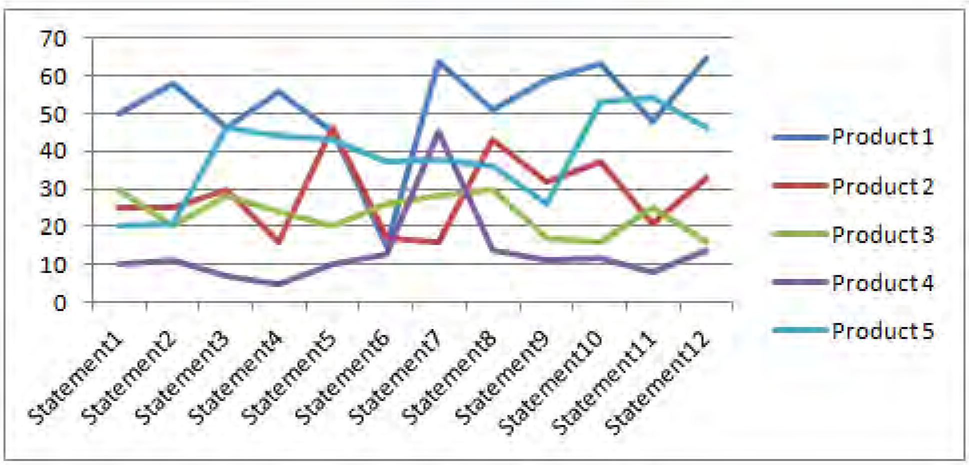
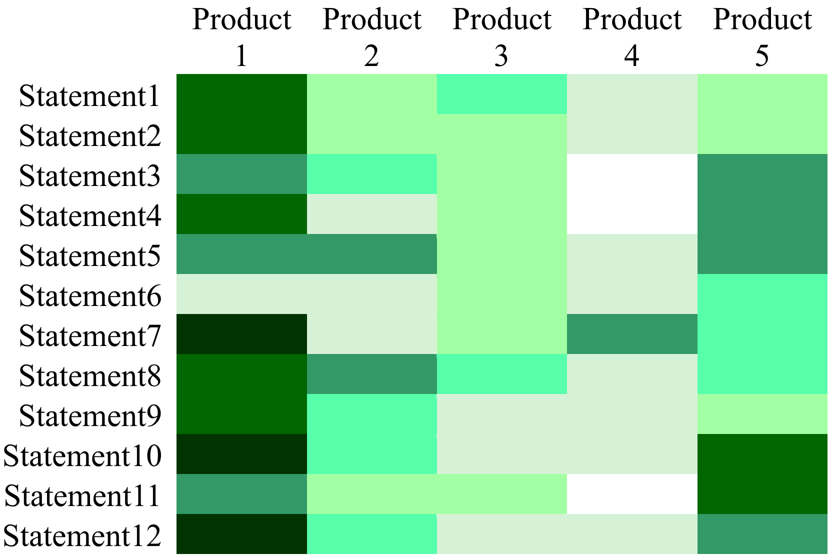


Exhibit 5: Heat Map for Diagnostics/Perceptions



Some perceptual statements may have low correlation with product choice (see explanation in the next section). These could be dropped from the charts. The significant statements can be sorted from most impact to least impact for presentation. For better visualization with the heat map, we could emphasize the importance of each statement by making its row (or column) height proportional to importance.

Aggregate analysis cannot reveal the different perceptions of heterogeneous groups, so researchers may decide to conduct the analysis by segments.

Sample Size Considerations

One may wonder about sample size requirements to stabilize models where conjoint attributes predict choice of perceptual statements. Following Rich Johnson's logic and recommendations for aggregate CBC models (Johnson and Orme, 2003) we might recommend that each level of each attribute appear with each perceptual item at minimum 500 times and preferably 1000 times across all respondents x choice tasks. Some algebra allows us to solve for the suggested sample size according to this simple rule of thumb:

C = Largest number of levels for any one conjoint attribute
D = Number of perceptual diagnostic items
A = Number of alternatives per CBC task
T = Number of CBC tasks
F = Number of perceptual items shown per CBC task

Minimum N = 500CD / ATF
Preferred N = 1000CD / ATF

Consider a study with the largest number of levels for any one conjoint attribute being 5 (C), 12 perceptual diagnostic items (D), 3 alternatives per CBC task (A), 8 CBC tasks (T), and 3 perceptual items shown per CBC task (F). Solving for N, Johnson's rule of thumb suggests a minimum sample size of 417 respondents and a preferred sample size of 833.

Determinance Scores for Perceptual Items

Exhibits 4 and 5 show predicted perceptual item scores (percent of respondents that agree) for products specified in the choice simulator, but they don't tell us which items are positively associated[2] with product choice. Alpert (1971) referred to these as *determinant attributes*. Determinant attributes are those that the buyer perceives as differing among product offerings and that positively influence preference. For example, *A safe airline to fly* is certainly important to respondents; but if buyers don't perceive any difference among airlines on safety, then it cannot be a determinant attribute since it alone will not influence choice among airlines.

One could perform a simple counting analysis to compute determinance scores for each perceptual statement (reflecting positive association with product choice). Recall that the respondent clicks which statements she associates with each product concept. Sometimes the checked perceptual statements are associated with concepts chosen as *best* and other times with

[2] We stop short of referring to these as "drivers of choice" since this would imply causality.

concepts indicated as *worst* (from the B/W CBC task directly above the perceptual grid)[3]. If we find that *best* product choices are often associated with a certain perceptual statement but *worst* choices are rarely associated with the same statement, then we might conclude that this perceptual statement is somehow related to choice. For each perceptual item, we can compute a summary determinance score by taking the %Best - %Worst; or, alternatively, %Best / %Worst, where the two scores are computed as follows:

Only considering the association data for Best concepts for item i:

$$\%\text{Best}_i = \#\text{Times\_Picked}_i \,/\, \#\text{Times\_Available\_to\_be\_Picked}_i$$

Only considering the association data for Worst concepts for item i:

$$\%\text{Worst}_i = \#\text{Times\_Picked}_i \,/\, \#\text{Times\_Available\_to\_be\_Picked}_i$$

Rather than using counting, a straightforward logit modeling approach[4] for computing determinance scores yields standard errors for performing t-tests of significance and computing confidence intervals. Aggregate logit, latent class, or disaggregate HB logit analysis could be used. The model follows the best/worst pattern suggested first by Louviere and also used within Sawtooth Software's popular MaxDiff analysis, as shown in Exhibit 6. For each concept x perceptual statement seen by the respondent, the task is coded for a binary logit model (in Exhibit 6, *IV* refers to an Independent Variable, *DV* refers to a Dependent Variable).

Exhibit 6: Coding for Estimating Attribute Determinance

	IV	DV	
Task1	1	1	(Perceptual item was selected for a "Best" concept)
	0	0	
Task2	1	0	(Perceptual item was not selected for a "Best" concept)
	0	1	
Task3	-1	1	(Perceptual item was selected for a "Worst" concept)
	0	0	
Task4	-1	0	(Perceptual item was not selected for a "Worst" concept)
	0	1	

For ease of interpretation, we may convert the estimated logit utility scores to a scale reflecting the difference between the likelihood of association with best choices less worst choices using the transform: $\text{Exp}(U_i)/[\text{Exp}(U_i)+1] - \text{Exp}(-U_i)/[\text{Exp}(-U_i)+1]$. Alternatively, one could covert to an odds ratio scale: $\text{Exp}(U_i)/[\text{Exp}(U_i)+1] \,/\, \text{Exp}(-U_i)/[\text{Exp}(-U_i)+1]$.

Notes:

1. We assume that choices of "Best" or "Worst" concepts are influenced by the same latent dimension of attribute determinance.

[3] Although our example uses B/W CBC questions, determinance analysis (by either counting or soon-to-be-described logit) can be conducted with standard best-only CBC.

[4] Many readers will recognize that the determinance modeling approach we present here is just a best/worst, univariate (single variable at a time) variation of the kind of derived importance regressions that have been used already for decades in market research. Researchers have commonly used ratings (or choices) of brands on perceptual statements to predict past brand choice, future intentions, or other brand preference ratings. A weakness of most of these approaches (including the determinance score estimation we present here) is the tendency to achieve significant parameters due to the halo effect.

2. This measure of determinance has the desirable quality of being a derived versus an overtly stated measure, meaning that it should mitigate common problems with stated measures such as social desirability and acquiescence bias.
3. If a perceptual item has at most a relatively low association with all product concepts (i.e., at most 10% agree that the product alternative is associated with a perceptual attribute, irrespective of the product composition), the researcher might decide to drop the item from the perceptual choice simulator, even if the determinance coefficient is statistically significant.

PAST RESEARCH INTEGRATING PERCEPTIONS AND CONJOINT DATA

At the 1989 Sawtooth Software Conference, Harla Hutchinson delivered a paper entitled, "Gaining a Competitive Advantage by Combining Perceptual Mapping and Conjoint Analysis." Within, she described efforts to leverage both part-worth utilities on hard conjoint attributes (for automobiles) with how respondent's perceived that the different automobile makes were positioned on softer attributes. Some of the softer attributes, while described using discrete levels of conjoint attributes, still involved a perceptual aspect, such as amount of *backseat legroom* in cars. The effort involved assigning part-worth utilities to product concepts within a choice simulator based on the conjoint attribute levels and also by assigning part-worth utilities based on perceived (rather than actual) attribute levels by respondents (even if those perceptions differed with reality, on such attributes such as *backseat legroom*, for instance).

Regarding this effort, the author of this paper (Orme) wrote in 2003:

> *"Combining perceptual information and preference part worths is not new. My colleagues Rich Johnson and Chris King developed choice simulators that used part worths mapped to each respondent's perceptions of brand performance quite a bit when at John Morton Company in the late 70s and early 80s. One of their colleagues, Harla Hutchinson, delivered a paper on this topic in the 1989 Sawtooth Software Conference entitled 'Gaining a Competitive Advantage by Combining Perceptual Mapping and Conjoint Analysis.'"*
> *"Based on conversations with Rich and Chris, combining perceptual information and preference part worths was not without problems. The perceptual information often seemed to dominate the overall sensitivity of the simulator. And, working with a model in which attributes did not necessarily have specific objective meaning, but that were mapped to subjective perceptions for each individual, made it difficult to assess how concrete changes to product specifications might affect demand." (Orme, 2003)*

At the 2003 Sawtooth Software Conference, Larry Gibson described a related approach of Eric Marder Associates called the SUMM method. Like the effort described in Hutchinson 1989, their method leveraged respondents' subjective perceptions of the alternatives on the various attributes. Preferences (a self-explicated method using an unbounded scale) were then combined with the respondents' idiosyncratic perceptions of alternatives on the various features to produce an integrated choice simulator.

At the 1997 Sawtooth Software Conference, Tom Pilon demonstrated how to create MDS perceptual maps based on perceived similarities among brands or SKUs for FMCG categories, such as beverages. Pilon's proxy for similarity was price cross-elasticity coefficients from CBC experiments involving brands and prices. A drawback of the approach was that the perceptual maps only had brand positions—no attribute information was shown. Thus, the researcher was left to interpret what the dimensions meant on the map (i.e., the y-axis might separate the beverages that are fruity from the colas; the x-axis might separate premium brands from the store brands).

At the 1999 Sawtooth Software Conference, Rich Johnson presented an extension of perceptual mapping called Composite Product Mapping (Johnson 1999). The technique combined the standard brand x attribute perceptual ratings with preference information on the brands (from either chip-allocation or conjoint part-worths on the brands). The perceptual space was developed to emphasize attributes that not only discriminated on brand perceptions, but also on brand preferences. Johnson overlaid contours of preference on the perceptual map, showing how areas of the map were associated with higher relative preference.

Ray Poynter presented a paper at the 1999 Sawtooth Software Conference that inspired the title of this current work. Ray described a qualitative approach for playing back on the computer screen the conjoint survey that a respondent had just completed while having an in-person human interviewer ask respondents open-ended questions to probe why they chose the concepts they did.

At the 2003 Sawtooth Software Conference, Marco Vriens and Curtis Frazier modeled the brand part-worth as a function of perceptual dimensions. Their choice simulator allowed managers not only to specify products within scenarios on the hard conjoint attributes and predict shares of preference, but to see how changes to perceptions of the brand name could also affect product choice. Frazier and co-authors updated the approach in a follow-up paper (Frazier et al. 2006).

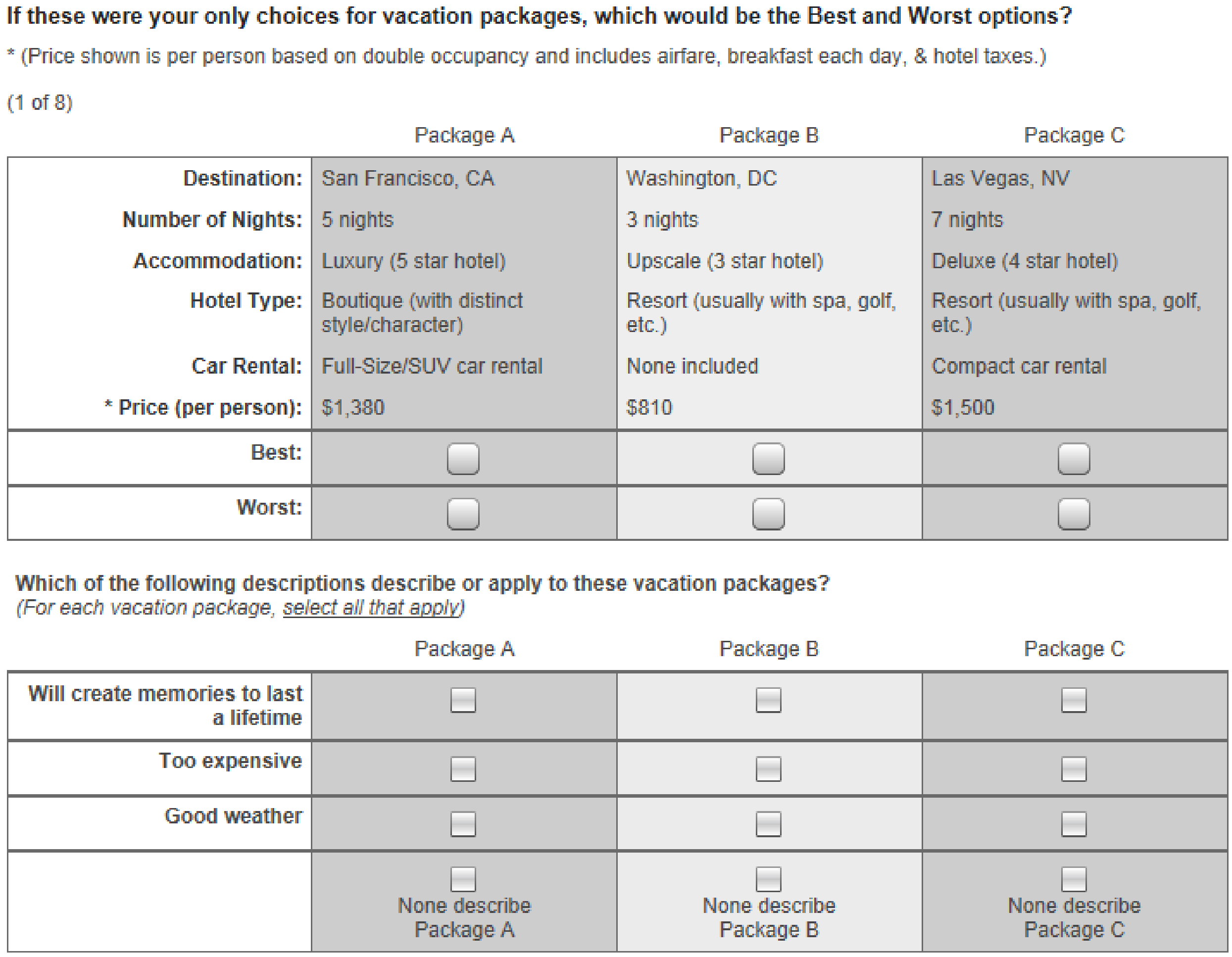
Yet another effort by Glerum et al. (2014) employed semi open-end questions, where respondents were asked to supply several adjectives to describe, for example, transportation modes. These open-end responses were coded by human researchers (evaluators) into a manageable number of pre-coded categories and then used as explanatory variables of revealed choice (the modes of transportation actually used by respondents). Specifically, the authors examined the impact of *perception of comfort of public transportation* on choice. A major challenge to overcome was reconciling how different evaluators related the adjectives to the latent construct.

To summarize the reviewed works, the Pilon, Vriens & Frazier, Johnson, and Glerum et al. approaches created associations between soft attribute perceptions and brand, SKU, or transportation mode preference. The Hutchinson and Gibson efforts elicited respondents' perceptions for brands (or vehicle makes) on each of multiple hard or soft attributes. However, what distinguishes our approach is the following, 1) we do not employ self-explicated ratings of brands on the product attributes individually; we use pick-any association data related to experimentally designed full-profile product concepts, 2) our approach predicts how respondents would perceive a given full-profile product concept given its brand and other conjoint attribute levels, not how people's product choices would be influenced based on changes to their perceptions of characteristics associated with the products. Of all the works cited here, the Poynter effort seems most similar to ours in spirit; though our implementation is quite different, since we perform a quantitative analysis on a pre-specified list of perceptual attributes to uncover the why's behind product choice as opposed to the open-end qualitative approach he did.

Pilot and Empirical Tests: Vacation Package Choices

We initially conducted a pilot test in June 2014 among a convenience sample of n=51 using *single-concept presentation* (described in Appendix A). While it appeared to work and the data had good face validity, we quickly thought of yet another approach that we call the *grid-style presentation* (Exhibit 7). In September 2014 we conducted a rigorous split-sample methodological test to compare these approaches using 627 respondents from SSI's online panel (many thanks to Survey Sampling International for supporting this research!). We report details of that experiment in Appendix A, the conclusion being that the grid-style approach worked better.

Exhibit 7: The Grid-Style Approach for Perceptual Choice Experiments



If these were your only choices for vacation packages, which would be the Best and Worst options?

* (Price shown is per person based on double occupancy and includes airfare, breakfast each day, & hotel taxes.)

(1 of 8)

	Package A	Package B	Package C
Destination:	San Francisco, CA	Washington, DC	Las Vegas, NV
Number of Nights:	5 nights	3 nights	7 nights
Accommodation:	Luxury (5 star hotel)	Upscale (3 star hotel)	Deluxe (4 star hotel)
Hotel Type:	Boutique (with distinct style/character)	Resort (usually with spa, golf, etc.)	Resort (usually with spa, golf, etc.)
Car Rental:	Full-Size/SUV car rental	None included	Compact car rental
*** Price (per person):**	$1,380	$810	$1,500
Best:	○	○	○
Worst:	○	○	○

Which of the following descriptions describe or apply to these vacation packages?
(For each vacation package, select all that apply)

	Package A	Package B	Package C
Will create memories to last a lifetime	□	□	□
Too expensive	□	□	□
Good weather	□	□	□
	□ None describe Package A	□ None describe Package B	□ None describe Package C

Our September 2014 methodological experiment involved the following conjoint attribute characteristics of different vacation packages for domestic travel within the United States (Exhibit 8).

Exhibit 8: Conjoint Attribute List for Vacation Package Choice

1) Destination:
   - Las Vegas, NV
   - Orlando, FL
   - Anaheim, CA
   - San Francisco, CA
   - Chicago, IL
   - New York, NY
   - Washington, DC
2) Number of Nights:
   - 3 nights
   - 5 nights
   - 7 nights
3) Accommodation:
   - Moderate (2 star hotel)
   - Upscale (3 star hotel)
   - Deluxe (4 star hotel)
   - Luxury (5 star hotel)
4) Hotel Type:
   - Business (with meeting/business services)
   - Resort (usually with spa, golf, etc.)
   - Boutique (with distinct style/character)
5) Car Rental:
   - None included
   - Compact car rental
   - Full-Size/SUV car rental
6) Price (per person):
   - $650 to $1,800 depending on number of nights.

	Low Price	Medium Price	High Price
3 nights	$650	$810	$970
5 nights	$920	$1,150	$1,380
7 nights	$1,190	$1,500	$1,800

For the perceptual choice experiment design, we used the following list of perceptual diagnostic statements (Exhibit 9).

Exhibit 9: Perceptual Items:

A trip I'd like to take kids on
A great summer vacation
A great winter vacation
I'd feel very safe in this city
Good weather
Too expensive
Fun
I'd feel pampered
Will create memories to last a lifetime
Relaxing time
Educates and expands horizons
A romantic vacation

We analyzed the upper (CBC) portion of the choice task (see Exhibit 7) using the standard CBC/HB approach as supported by Sawtooth Software's SSI Web software system. In addition, we estimated determinance coefficients for each of the 12 perceptual statements as described earlier (Exhibit 6) using 12 separate univariate, binary aggregate logit models[5].

Exhibit 10 displays the raw determinance coefficients for the statements (sorted in terms of absolute magnitude) for the third of the sample (n=218) that completed the grid-style version of the perceptual choice experiment questionnaire.

Exhibit 10: Determinance Coefficients

	Beta	Std Err.	T-Ratio
Fun	0.82	0.075	10.9
Creates memories	0.73	0.073	9.9
Great summer	0.69	0.073	9.4
Relaxing	0.62	0.072	8.6
I'd feel pampered	0.53	0.071	7.5
Too expensive	-0.49	0.071	-7.0
Good weather	0.47	0.071	6.7
Educates	0.40	0.070	5.8
Romantic	0.40	0.070	5.8
Take kids on	0.36	0.070	5.1
Feel safe	0.34	0.070	4.9
Great winter vacation	0.26	0.069	3.8

The perceptual dimension *Fun* is the most determinant item, meaning that it was most highly related to product choice. Using the formula we earlier introduced, we can compute the

[5] It should be noted that the separate models could be formulated as a single multivariate logit model. This could be a useful approach if using latent class analysis to develop market segments of respondents who share similar motivations and perceptions as related to choice. Extending this idea, the CBC data could also be integrated within this same choice model for an integrated latent class model of choice involving conjoint utilities and diagnostic betas (but not without the problems of mixing two choice contexts with different response error rates). An alternative that avoids this problem is to use the segment membership assignments from separate latent class runs on determinance scores and CBC utilities within Cluster Ensemble analysis (e.g., CCEA software).

difference in choice likelihood between product concepts that respondents perceived as *Fun* versus those not perceived as *Fun*:

Exp(0.82)/[Exp(0.82)+1] - Exp(-0.82)/[Exp(-0.82)+1] = 0.3885

When a concept was viewed as *Fun*, its choice probability for the sample was 38.85% higher (in absolute magnitude) than a concept not viewed as *Fun*. Expressed as an odds ratio . . .

Exp(0.82)/[Exp(0.82)+1] / Exp(-0.82)/[Exp(-0.82)+1] = 2.2705

. . . concepts associated with *Fun* are 2.27 times more likely to be chosen than those not viewed as *Fun*.

The least determinant item was *Great Winter Vacation*. Following the same formulas, respondents chose concepts marked as a *Great Winter Vacation* with a 12.93% higher absolute probability (or as an odds ratio, selected 1.30 times as often) as those not perceived as a *Great Winter Vacation*.

Perceptual Choice Experiment Modeling

The key trick with perceptual choice experiments is building aggregate logit models[6] relating the conjoint attribute levels to choice of the perceptual items (one binary logit model per perceptual item). In Exhibit 11, we show results for just three of the perceptual statements (with significant attributes bolded). These models are in reality each discrete choice conjoint experiments, except that the dependent variable is perceptual association rather than product choice. As with standard conjoint output, the part-worths may only be compared *within* attributes.

In terms of the item, *A Romantic Vacation*, Orlando, Anaheim, and Chicago score relatively lower than Las Vegas, San Francisco or New York. Respondents find a resort-style or boutique hotel more romantic than a business style hotel. Regarding the statement *I'd Feel Pampered*, spending either 5 or 7 nights at a 3-star or higher quality hotel—especially a resort type hotel—is more associated with that perception. However, longer trips (5 and 7 nights) are also positively associated with the perception of *Too Expensive*. But (holding price constant) the perception of being too expensive can be lowered by the vacation package including an upscale hotel or a Full-Size/SUV car rental.

[6] Each model potentially could use all conjoint attribute levels as predictors of perceptual item choice. With certain perceptual items, however, some conjoint attributes made no logical sense as predictors, so we excluded them from the model. For example, regarding association with *Safe City*, conjoint attributes such as Hotel Type, Car Rental, and Vacation package Price are excluded from the model.

Exhibit 11: Logit Coefficients as Predictors of Perceptions
(3 of 12 Statements, for Illustration)
(First Level of Each Attribute Constrained to Zero via Dummy Coding)

	Romantic	**Pampered**	**Too Expensive**
Las Vegas, NV	**0.00**	**0.00**	0.00
Orlando, FL	**-0.71**	**-0.12**	0.09
Anaheim, CA	**-0.85**	**-0.54**	0.04
San Francisco, CA	**0.41**	**-0.14**	-0.14
Chicago, IL	**-0.81**	**-0.14**	0.22
New York, NY	**0.20**	**-0.23**	-0.15
Washington, DC	**-0.56**	**-0.22**	-0.33
3 nights	0.00	**0.00**	**0.00**
5 nights	-0.26	**0.46**	**0.98**
7 nights	-0.23	**0.41**	**1.68**
Moderate (2 star hotel)	0.00	**0.00**	**0.00**
Upscale (3 star hotel)	0.30	**0.55**	**-0.41**
Deluxe (4 star hotel)	0.21	**0.89**	**-0.48**
Luxury (5 star hotel)	0.32	**1.13**	**-0.62**
Business (with meeting/business	**0.00**	**0.00**	0.00
Resort (usually with spa, golf, etc.)	**0.53**	**1.03**	-0.17
Boutique (with distinct	**0.30**	**0.52**	-0.24
None included	0.00	0.00	**0.00**
Compact car rental	0.08	0.05	**-0.15**
Full-Size/SUV car rental	-0.16	0.14	**-0.32**
Low Price	0.00	0.00	**0.00**
Med Price	0.07	0.30	**0.73**
High Price	0.02	-0.22	**1.16**
Alternative Specific Constant:	-1.02	-2.29	-1.81

Given these utility scores, we can predict the percent of agreement that any vacation package (described using one attribute level from each attribute) would be associated with each of the perceptual statements. Consider the *Romantic* perceptual statement. Referring to the utilities in Exhibit 11, the percent of respondents that would agree that [Orlando, 3 nights, Luxury (5 star hotel), Resort (usually with spa, golf, etc.), Compact car rental, Med Price] was a romantic vacation package is:

Orlando	-0.71
3 nights	0.00
Luxury (5 star hotel)	0.32
Resort (usually with spa…)	0.53
Compact car rental	0.08
Med Price	0.07
Alternative Specific Constant	-1.02
Sum:	-0.73

[7]Probability agree this package is "Romantic" = Exp(-0.73) / [Exp(-0.73) + Exp(0)] = 32.5%

Though not shown in Exhibit 11, the standard errors associated with conjoint attribute levels predicting the perceptual statements range from about 0.14 (for 3-level attributes) to 0.28 (for the 7-level attribute). If we had used 800 respondents instead of 200, the standard errors would be halved (doubling the precision) with standard errors ranging from 0.07 to 0.14. For aggregate logit scores involving choice data, standard errors of this magnitude are higher than what we typically are accustomed to for CBC data (where we typically see standard errors of 0.05 or less). Considering the precision of the results from the analysis of determinance (Exhibit 10) and the analysis to compute weights that predict perceptual statements (Exhibit 11), we see that the latter analysis is much more demanding on sample size. For obtaining relatively precise results for this latter analysis, which is the core focus of perceptual choice experiments (Exhibit 11), perhaps n=800 to 1000 would be worth the cost and effort.

Earlier, we referred to Johnson's rule of thumb for suggested sample size. For this conjoint attribute list and particulars of our questionnaire design involving 12 perceptual items and a conjoint attribute involving at most 7 levels, that formula suggests sample size between n=583 (minimum) and n=1167 (preferred). Perceptual choice experiments can be quite demanding on sample size!

Choice Simulations with "Why" Insights

Next, we built a market simulator to predict choice likelihood for different vacation packages (the standard CBC simulator based on HB scores) as well as to use the conjoint attributes to predict the degree of agreement on the perceptual dimensions we included in the questionnaire. We decided to report perceptual results only for the most determinant dimensions via a heat map[8], where the width of the column is proportional to determinance (Exhibit 12).

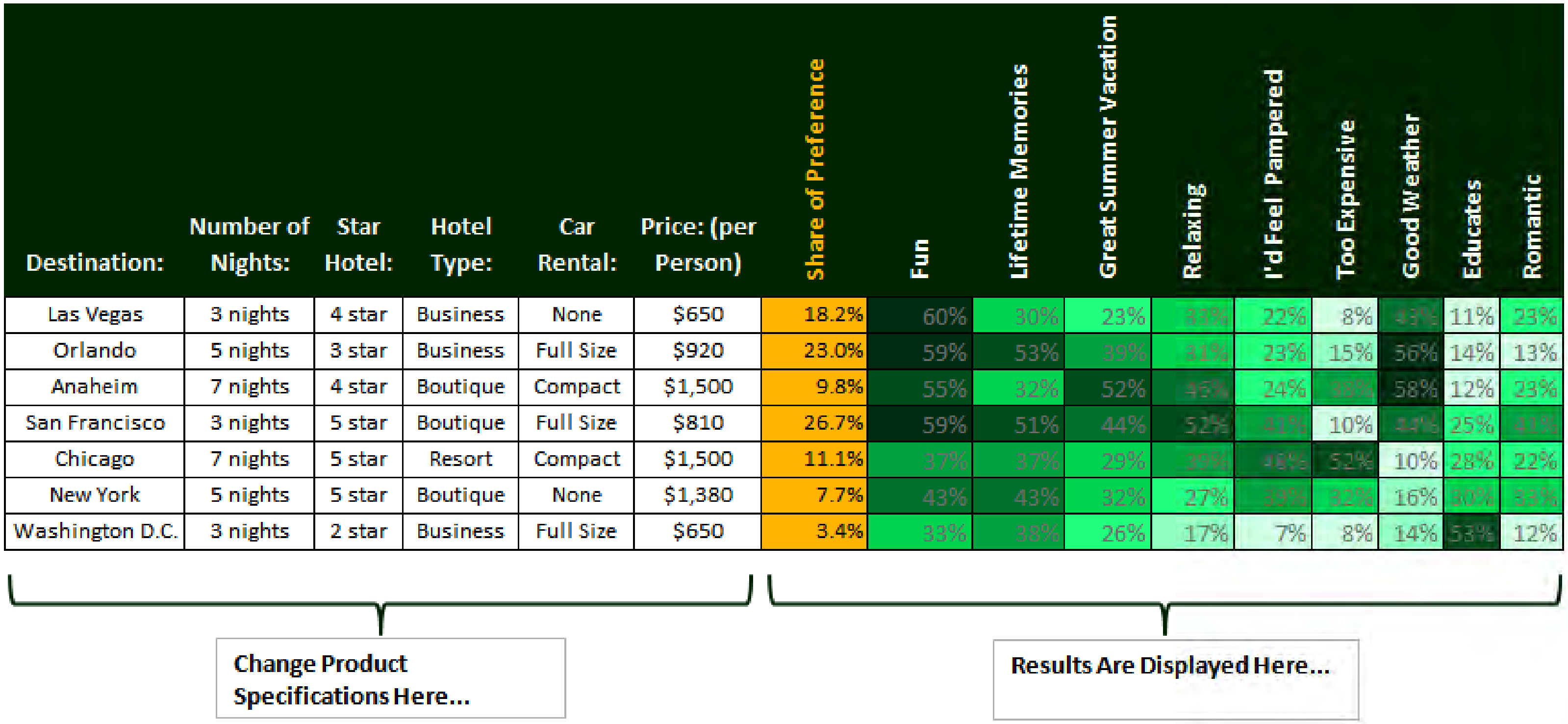
For our base case scenario, the 3-night San Francisco vacation package at a boutique-style 5-star hotel, with full-size car rental for $810 per person is the most preferred package. The perceptual choice experiment simulation gives some insight into why. This San Francisco travel package is perceived to be essentially the most *Fun* (59% agreement) and *Relaxing* (52%) vacation package of the seven. It scores high as well on other dimensions and relatively low on the *Too Expensive* dimension.

Washington D.C., on the other hand, only captures 3.4% share of choice (the lowest share). Although it is perceived by the sample of respondents to be the vacation package that is most

[7] In our binary logit formulation, the "not chosen" constant alternative was constrained as the zero-utility alternative (a vector of zeros in the design matrix), leading to the Exp(0) term in the denominator of this equation.

[8] Excel's ***Conditional Formatting + Highlight Cells Rules*** makes it relatively easy to create these heatmaps.

likely to *Educate and Expand Horizons*, that attribute has relatively lower determinance (as we computed earlier using aggregate logit). The Washington D.C. vacation packages scores lowest of the seven on the *Fun* dimension, which is the most determinant attribute.



Destination:	Number of Nights:	Star Hotel:	Hotel Type:	Car Rental:	Price: (per Person)	Share of Preference	Fun	Lifetime Memories	Great Summer Vacation	Relaxing	I'd Feel Pampered	Too Expensive	Good Weather	Educates	Romantic
Las Vegas	3 nights	4 star	Business	None	$650	18.2%	60%	30%	23%	33%	22%	8%	43%	11%	23%
Orlando	5 nights	3 star	Business	Full Size	$920	23.0%	59%	53%	39%	31%	23%	15%	56%	14%	13%
Anaheim	7 nights	4 star	Boutique	Compact	$1,500	9.8%	55%	32%	52%	46%	24%	38%	58%	12%	23%
San Francisco	3 nights	5 star	Boutique	Full Size	$810	26.7%	59%	51%	44%	52%	41%	10%	44%	25%	41%
Chicago	7 nights	5 star	Resort	Compact	$1,500	11.1%	37%	37%	29%	39%	48%	52%	10%	28%	22%
New York	5 nights	5 star	Boutique	None	$1,380	7.7%	43%	43%	32%	27%	39%	32%	16%	30%	33%
Washington D.C.	3 nights	2 star	Business	Full Size	$650	3.4%	33%	38%	26%	17%	7%	8%	14%	53%	12%

Change Product Specifications Here...
Results Are Displayed Here...

Exhibit 12: Choice Simulator Predicted Agreement on Perceptual Items

If we change the vacation package for Orlando, 5 nights from the business-style hotel to a resort/spa style hotel and rerun the market simulation, its share of preference increases from 23.0% to 32.7%. With that change in hotel type, perceptions (that are significantly influenced by hotel type) for the Orlando, 5 night package shift as follows:

Exhibit 13: Shifts in Perceptions Due to Changing Orlando, 5 Night Package from Business to Spa/Resort Hotel

Fun:	59%	⇨	66%
Lifetime Memories:	53%	⇨	57%
Great Summer Vacation:	39%	⇨	48%
Relaxing:	31%	⇨	43%
I'd Feel Pampered:	23%	⇨	40%
Romantic:	13%	⇨	19%

A market simulator built in Excel displays these changes to share of preference and perceptions instantaneously with updates to the heat map colors and values on the grid.

Summary and Conclusion

We have introduced an extension to CBC called *Perceptual Choice Experiments* that provides insights into *why* respondents make choices. The approach involves placing perceptual items and pick-any association tasks directly beneath standard CBC questions. We demonstrated how the insights could be visualized for managers via a heat map integrated within a what-if choice simulator. We also demonstrated how to estimate determinance weights for the perceptual items, allowing the researcher to prioritize the items and ignore any not related to choice.

Unfortunately, very few things in life come for free. Our Perceptual Choice Experiment slightly more than doubles the time for respondents to complete the eight-question CBC survey (see details in Appendix A). Rather than a median time of 20 seconds per CBC task, the CBC + perceptual choice experiment took respondents a total of 44 seconds per task (a total of nearly six minutes for an 8-question CBC). Also, perceptual choice experiments require large sample sizes (perhaps n=600 to 1200, for a typical experimental design) to obtain reasonably precise predictions of perceptual agreement for product concepts.

Bryan Orme

Appendix A: Comparison of Two Questionnaire Formats for Conducting Perceptual Choice Experiments

In September 2014, we conducted a split-sample experiment (with thanks to Survey Sampling International for providing the sample) to test which of two different questionnaire formats was better for conducting perceptual choice experiments. Respondents with HH income > $34,000 and who intended to travel out-of-state for a vacation in the next 12 months were invited to complete one of three different (randomly selected) questionnaires. The subject matter was vacation packages. The conjoint attribute list had six attributes: Destination (7 cities), #Nights (3 levels), #Stars for Hotel (4 levels), Type of Hotel (3 levels), Car Rental (3 levels), and Price (3 levels).

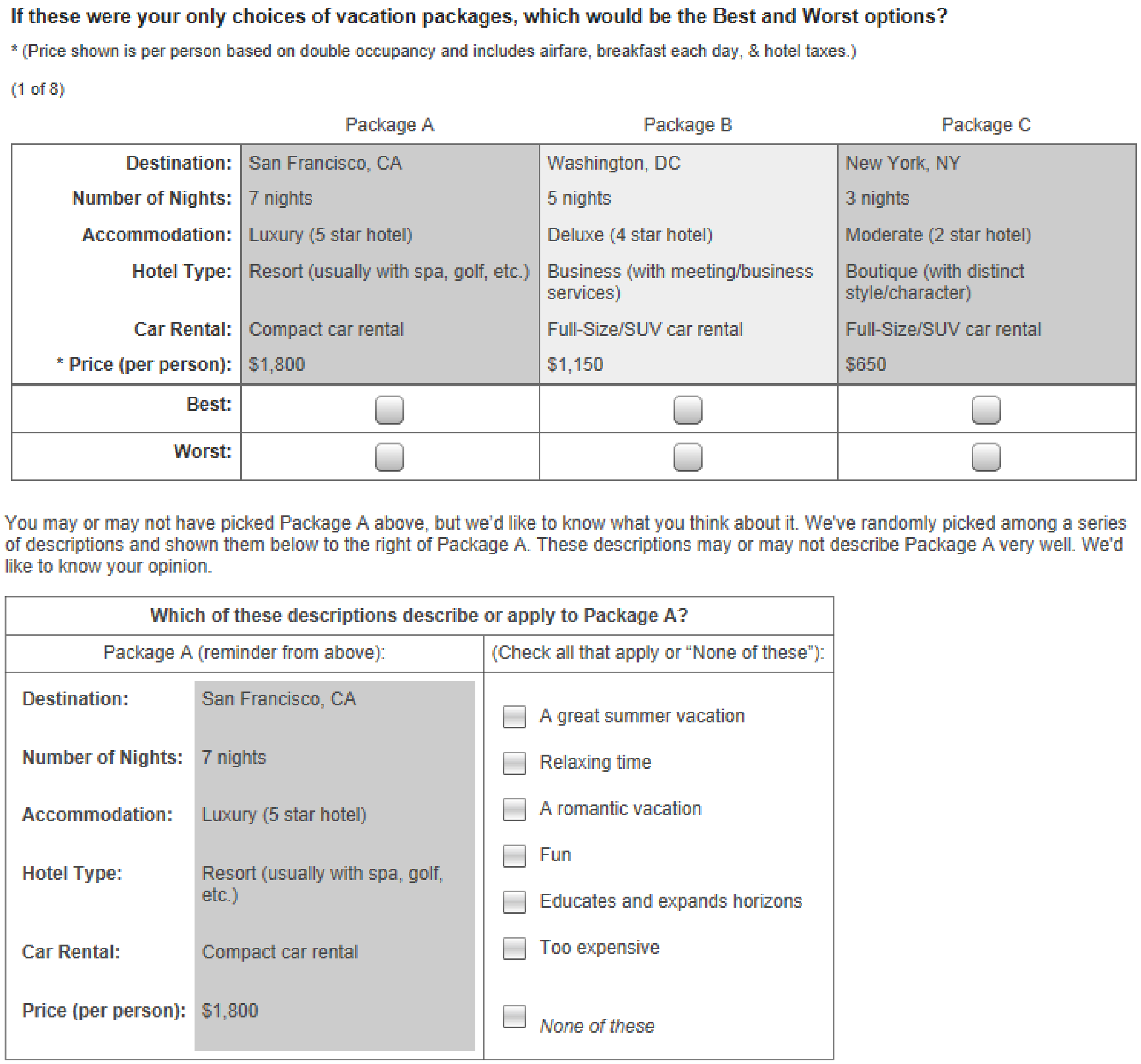
The perceptual choice experiment involved 12 statements, such as *A great summer vacation*, *Fun*, and *A romantic vacation*.

These three cells (different versions of the questionnaire) were as follows:

Cell 1: CBC + Single-Card Perceptual Choice Experiment (n=199)
Cell 2: CBC + Grid-Based Perceptual Choice Experiment (n=218)
Cell 3: CBC with no perceptual choice experiment (n=210)

The two perceptual choice formats were *Single-Concept Format* or *Grid Format* as shown below in Exhibits A1 and A2. Cell 3 was a control group that only completed a standard CBC exercise for comparison. The perceptual choice questions were asked beneath each of 8 choice tasks in the surveys for Cells 1 and 2.

Exhibit A1: Single-Concept Format[9]



If these were your only choices of vacation packages, which would be the Best and Worst options?

* (Price shown is per person based on double occupancy and includes airfare, breakfast each day, & hotel taxes.)

(1 of 8)

	Package A	Package B	Package C
Destination:	San Francisco, CA	Washington, DC	New York, NY
Number of Nights:	7 nights	5 nights	3 nights
Accommodation:	Luxury (5 star hotel)	Deluxe (4 star hotel)	Moderate (2 star hotel)
Hotel Type:	Resort (usually with spa, golf, etc.)	Business (with meeting/business services)	Boutique (with distinct style/character)
Car Rental:	Compact car rental	Full-Size/SUV car rental	Full-Size/SUV car rental
*** Price (per person):**	$1,800	$1,150	$650
Best:	□	□	□
Worst:	□	□	□

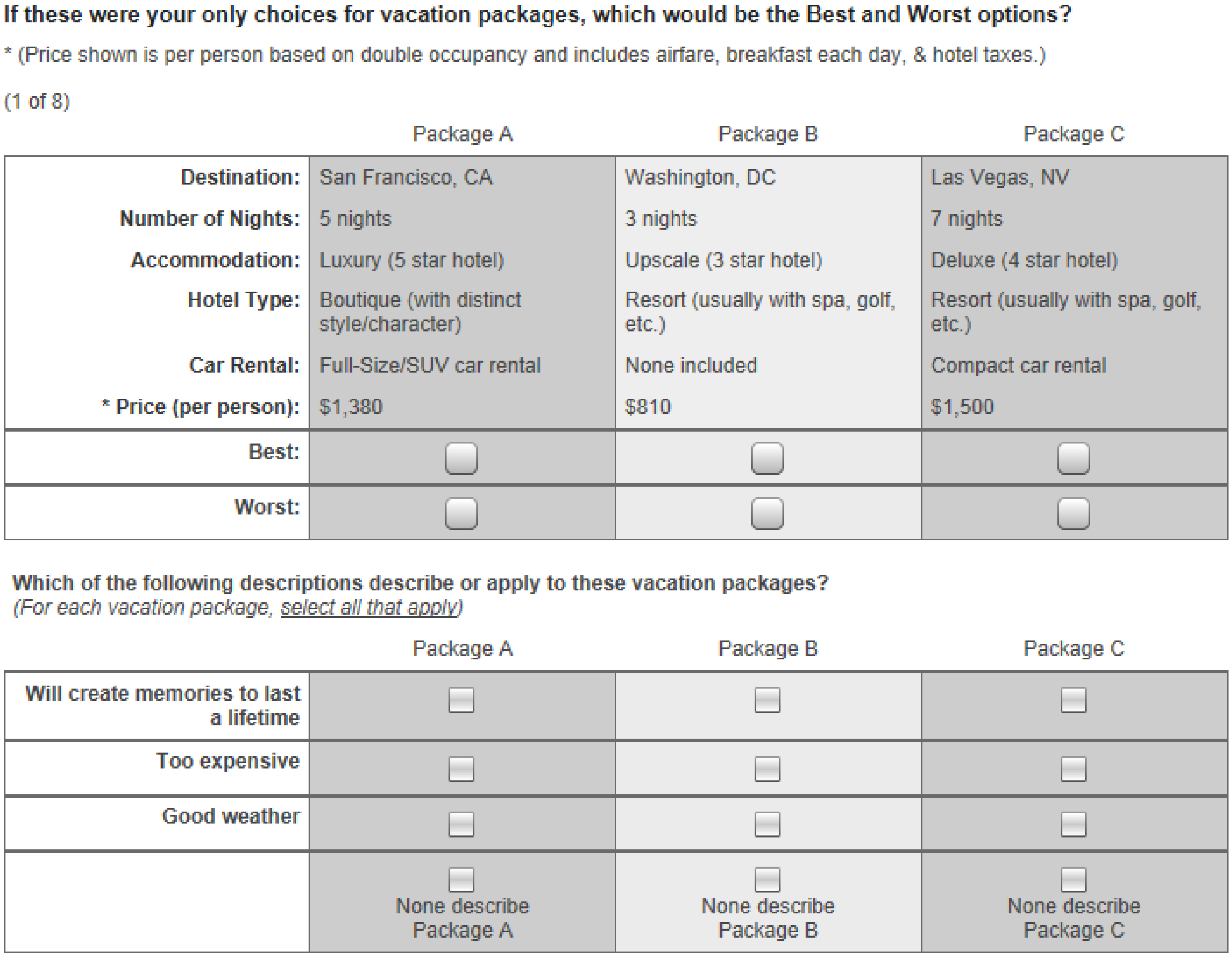
You may or may not have picked Package A above, but we'd like to know what you think about it. We've randomly picked among a series of descriptions and shown them below to the right of Package A. These descriptions may or may not describe Package A very well. We'd like to know your opinion.

Which of these descriptions describe or apply to Package A?

Package A (reminder from above):		(Check all that apply or "None of these"):
Destination:	San Francisco, CA	□ A great summer vacation
Number of Nights:	7 nights	□ Relaxing time
Accommodation:	Luxury (5 star hotel)	□ A romantic vacation
Hotel Type:	Resort (usually with spa, golf, etc.)	□ Fun
Car Rental:	Compact car rental	□ Educates and expands horizons
Price (per person):	$1,800	□ Too expensive
		□ *None of these*

[9] A random task is selected to ensure a level-balanced and orthogonal design for efficient binary logit modeling of perceptual choices. If only the respondent's selected concept from the CBC question was used in the perceptual follow-up, then the perceptual choice experimental design would be strongly biased in favor of preferred levels.

Exhibit A2: Grid Format



If these were your only choices for vacation packages, which would be the Best and Worst options?

* (Price shown is per person based on double occupancy and includes airfare, breakfast each day, & hotel taxes.)

(1 of 8)

	Package A	Package B	Package C
Destination:	San Francisco, CA	Washington, DC	Las Vegas, NV
Number of Nights:	5 nights	3 nights	7 nights
Accommodation:	Luxury (5 star hotel)	Upscale (3 star hotel)	Deluxe (4 star hotel)
Hotel Type:	Boutique (with distinct style/character)	Resort (usually with spa, golf, etc.)	Resort (usually with spa, golf, etc.)
Car Rental:	Full-Size/SUV car rental	None included	Compact car rental
*** Price (per person):**	$1,380	$810	$1,500
Best:	☐	☐	☐
Worst:	☐	☐	☐

Which of the following descriptions describe or apply to these vacation packages?
(For each vacation package, select all that apply)

	Package A	Package B	Package C
Will create memories to last a lifetime	☐	☐	☐
Too expensive	☐	☐	☐
Good weather	☐	☐	☐
	☐ None describe Package A	☐ None describe Package B	☐ None describe Package C

Anecdotal Pre-Test Evidence

Prior to fielding the split-sample study, we conducted an informal poll among employees at Sawtooth Software. About 2/3 preferred the grid-style (Cell 2) approach to the single-concept (Cell 1) approach. Those who preferred the grid-style approach commented that it seemed strange that the single-concept (Cell 1) approach randomly selected one of the concepts for evaluation on the perceptual items. The randomly selected single concept approach made them feel more at the mercy of an arbitrary process rather than empowered and in control of providing their opinions regarding concepts they both liked and didn't like from the CBC portion of the task. Although this is purely anecdotal evidence from a small and certainly biased sample of market researchers and software developers, it is interesting feedback.

Time to Complete Choice Screens

Median time per choice screen (task) for the three questionnaires is shown in Exhibit A3.

Exhibit A3: Median Seconds Per Task

	Task1	Task2	Task3	Task4	Task5	Task6	Task7	Task8	Average
Cell1	61.5	41.5	37	32	31	27.5	28.5	27	35.75
Cell2	71	50	44	41	37	36	36	34	43.63
Cell3	34	23	18	18	18	18	15	15	19.88

Adding the perceptual questions to the CBC experiment doubles the time to complete the choice screens. The grid-style approach (Cell 2) is a bit longer to complete than the single-concept approach, but with 50% more information collected (given our questionnaire design): with the single-concept layout, we showed 6 items per task x 8 tasks = 48 perceptual agreement check-boxes; the grid-style approach featured 3 items per task x 3 concepts per task x 8 tasks = 72 perceptual agreement check-boxes.

Qualitative Assessment of the Questionnaires

At the end of the survey, we asked respondents to evaluate their experience using a 5-point scale (1=Strongly Disagree, 2=Somewhat Disagree, 3=Neither Agree Nor Disagree, 4=Somewhat Agree, 5=Strongly Agree).

Exhibit A4: Qualitative Assessment of Questionnaire Experience

	Cell 1 CBC + Single-Concept Perceptions	Cell2 CBC + Grid Perceptions	Cell3 CBC Only (No Perceptual Questions)
This survey sometimes was confusing	2.13 (0.082) 18% agree	2.13 (0.084) 19% agree	1.90 (0.078) 14% agree
This survey was enjoyable	4.07 (0.062) 77% agree	3.99 (0.067) 76% agree	4.19 (0.059) 80% agree
This survey was too repetitive	2.64 (0.089) 32% agree	2.64 (0.084) 27% agree	2.29 (0.086) 23% agree
I found myself starting to lose concentration at least once	2.30 (0.088) 25% agree	2.43 (0.083) 22% agree	2.07 (0.079) 14% agree
This survey was too long	2.11 (0.079) 13% agree	2.27 (0.081) 16% agree	1.78 (0.073) 8% agree
The questions about which descriptions applied to different vacation packages were easy to answer	4.02 (0.067) 78% agree	3.91 (0.072) 75% agree	NA

(No statistically significant differences between first 2 columns.
Standard errors shown in parenthesis.)

The data suggest that respondents saw no difference between the single-concept and grid-style approaches on these qualitative dimensions.

Number of Perceptual Boxes Checked

If respondents clicked very few perceptual check-boxes (indicating agreement that the product concepts were described well by perceptual statements), we'd have little to go by to model how conjoint attribute levels led to agreement with the perceptual statements. We use binary logit to build the models, so maximal efficiency occurs if the item is selected 50% of the time. As the probability of agreeing with perceptual items tends toward either 0% or 100%, the binary logit models have very little information by which to estimate part-worth perceptual parameters (other than the constant).

The single-concept approach (Cell 1) led to 28% of the perceptual boxes checked. The grid-style approach (Cell 2) led to 35%.

Assembling the information reported to this point:

- Regarding time to complete, Cell 2 was 35.75 seconds / 43.63 seconds = 0.82 as efficient.
- Regarding amount of data collected, Cell 2 was 1.5 as efficient (9 checks vs. 6 checks per task).
- Regarding percent of agreement boxes checked, Cell 2 collected 35% / 28% = 1.25x more information.

Net, Cell 2 was: 0.82 x 1.5 x 1.25 = 1.54 as efficient as Cell 1 per time-equalized respondent effort.

Do Follow-Up Perceptual Questions Affect CBC Responses?

An important question is whether the presence of the perceptual follow-up questions leads to higher or lower quality responses to the standard CBC tasks at the top of the choice screen. Perhaps when respondents know that they will be asked to delve deeper by evaluating the perceptual aspects of each of the product concepts, they provide better CBC choices. To investigate this, prior to the CBC questions we asked respondents a series of self-explicated questions about four of the attributes (destination, hotel stars, hotel type, and car rental options). Respondents chose their preferred level for each of the attributes as well as whether each of the attributes mattered to them on a 3-point scale (Yes, it's very important; Yes, but not very important; No). For each respondent, we then compared the most preferred levels for the HB utilities estimated from the CBC tasks to the self-explicated preferred levels (but ignoring any attributes that were rated as not very or not at all important). The two groups of respondents who completed perceptual choice questions beneath each CBC task had hit rate matches between self-explicated and CBC/HB utilities 9% and 6% higher (for cells 1 and 2, respectively) than the control respondents who only completed the standard CBC tasks (Cell 3). Though these hit rates are directionally higher for perceptual choice experiment respondents, the differences were not statistically significant. The data suggest, but do not confirm, that respondents provide better quality answers to the CBC questions when they are asked follow-up perceptual diagnostic questions about the product concepts.

Summary

Our experiment suggests that the grid-based method (Cell 2) of data collection for perceptual choice experiments is better than the single-card approach (Cell 1):

- Factoring in the time to complete the questions and the amount of data collected, the grid-style approach is 1.54 more times efficient than the single-concept approach. In other words, for every second of respondent effort, it is 54% more efficient, while being no less tiring or confusing.

- The grid-based approach is more compact to present on the survey page.

- Individual-level analysis suggests that when respondents are asked to complete additional perceptual association questions, the quality of their answers to the standard CBC tasks on the same page may be slightly improved.

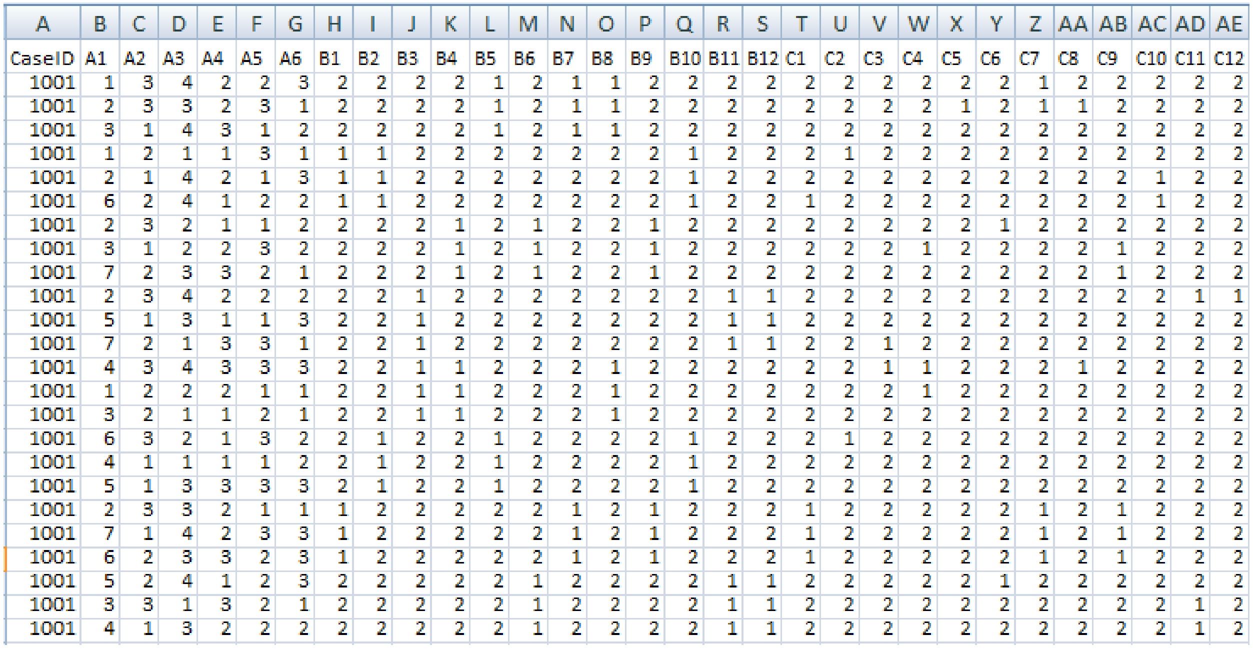
- Our qualitative assessment is that the grid-style approach seems more logical than to ask respondents perceptual questions about a randomly selected concept.

Appendix B: Data Preparation for Sawtooth Software's MBC (Menu-Based Choice) Software

Any software that can perform MNL or binary logit analysis may be used to estimate the perceptual choice models described in this paper. Among the Sawtooth Software tools, MBC (Menu-Based Choice) software is rather handy for performing the modeling.

The data should be prepared in a comma-separated values file (.csv file) as shown below:

Exhibit B1: Data Preparation for MBC Software



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
CaseID	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
1001	1	3	4	2	2	3	2	2	2	2	1	2	1	1	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2
1001	2	3	3	2	3	1	2	2	2	2	1	2	1	1	2	2	2	2	2	2	2	2	1	2	1	1	2	2	2	2
1001	3	1	4	3	1	2	2	2	2	2	1	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1001	1	2	1	1	3	1	1	1	2	2	2	2	2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	2	2
1001	2	1	4	2	1	3	1	1	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	1	2	2
1001	6	2	4	1	2	2	1	1	2	2	2	2	2	2	2	1	2	2	1	2	2	2	2	2	2	2	2	1	2	2
1001	2	3	2	1	1	2	2	2	2	1	2	1	2	2	1	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2
1001	3	1	2	2	3	2	2	2	2	1	2	1	2	2	1	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2
1001	7	2	3	3	2	1	2	2	2	1	2	1	2	2	1	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2
1001	2	3	4	2	2	2	2	2	1	2	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	1	1
1001	5	1	3	1	1	3	2	2	1	2	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2
1001	7	2	1	3	3	1	2	2	1	2	2	2	2	2	2	2	1	1	2	2	1	2	2	2	2	2	2	2	2	2
1001	4	3	4	3	3	3	2	2	1	1	2	2	2	1	2	2	2	2	2	2	1	1	2	2	2	1	2	2	2	2
1001	1	2	2	2	1	1	2	2	1	1	2	2	2	1	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2
1001	3	2	1	1	2	1	2	2	1	1	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1001	6	3	2	1	3	2	2	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	2	2
1001	4	1	1	1	1	2	2	1	2	2	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1001	5	1	3	3	3	3	2	1	2	2	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1001	2	3	3	2	1	1	1	2	2	2	2	2	1	2	1	2	2	2	1	2	2	2	2	2	1	2	1	2	2	2
1001	7	1	4	2	3	3	1	2	2	2	2	2	1	2	1	2	2	2	1	2	2	2	2	2	1	2	1	2	2	2
1001	6	2	3	3	2	3	1	2	2	2	2	2	1	2	1	2	2	2	1	2	2	2	2	2	1	2	1	2	2	2
1001	5	2	4	1	2	3	2	2	2	2	2	1	2	2	2	2	1	1	2	2	2	2	2	1	2	2	2	2	2	2
1001	3	3	1	3	2	1	2	2	2	2	2	1	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	1	2
1001	4	1	3	2	2	2	2	2	2	2	2	1	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	1	2

For our questionnaire layout as described in Appendix A, each respondent's data are coded in 24 rows (8 choice screens x 3 vacation concepts per screen).

The data layout is:

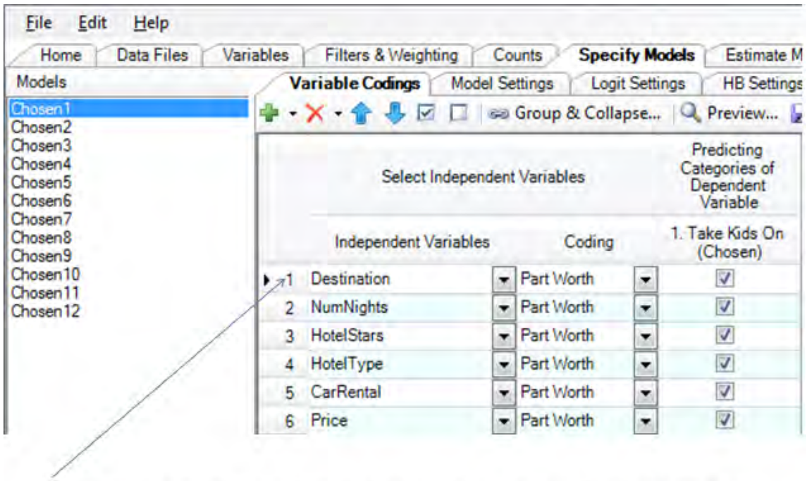
Fields	**Description**
CaseID	Respondent number
A1–A6:	Conjoint design, attribute level indices for attributes 1 through 6
B1–B12:	Availability flags for perceptual items 1–12, 1=available, 2=not available.
C1–C12:	Whether each of perceptual items 1–12 was selected, 1=Yes, 2=No.

For example, in choice task #1, respondent #1001 evaluated the conjoint concept: "1, 3, 4, 2, 2, 3" which means "Las Vegas, NV; 7 nights; Luxury (5 star hotel); Resort (usually with spa, golf, etc.); Compact car rental; $1,800" . . . with respect to perceptual statements 5, 7, and 8 (Good weather, Fun, and I'd feel pampered). The respondent clicked boxes indicating that only item 7 (Fun) described the conjoint concept.

To analyze the data using MBC, classify variables A1–A6 and B1–B12 as independent variables. Specify C1–C12 as dependent variables (where "2" is the off-state). Specify that Variable C1 is conditional upon B1 equal to "1" (is available); variable C2 is conditional upon B2 = 1, etc. for all twelve dependent variables.

The *Specify Models* dialog looks like the following, for each of 12 aggregate logit model specifications (modeling the dependent variable *Take Kids On* is shown below):

Exhibit B2: Variable Codings Dialog



Chosen1
Chosen2
Chosen3
Chosen4
Chosen5
Chosen6
Chosen7
Chosen8
Chosen9
Chosen10
Chosen11
Chosen12

	Select Independent Variables		Predicting Categories of Dependent Variable
	Independent Variables	Coding	1. Take Kids On (Chosen)
1	Destination	Part Worth	☑
2	NumNights	Part Worth	☑
3	HotelStars	Part Worth	☑
4	HotelType	Part Worth	☑
5	CarRental	Part Worth	☑
6	Price	Part Worth	☑

Note: only include independent variables that make logical sense as predictors of the perceptual evaluations! For example, Car Rental would not be a logical predictor of "Good Weather"

The MBC software automatically dummy-codes the independent variables, with the first level of each independent variable selected as reference (0-utility) levels. The aggregate logit output from MBC software is shown in Exhibit B3.

Exhibit B3: MBC Logit Output

```
Run includes 211 respondents (211.00 weighted).

1266 tasks are included in this model, for a weighted average of 6.0 tasks per respondent.

Total number of choices in each response category:
Category  Frequency  Percent
-----------------------------------------------------
       1      409.0   32.31%
       2      857.0   67.69%


Iteration    1  Log-likelihood = -744.16509  Chi Sq = 266.71847  RLH = 0.55554
Iteration    2  Log-likelihood = -741.40437  Chi Sq = 272.23993  RLH = 0.55676
Iteration    3  Log-likelihood = -741.37576  Chi Sq = 272.29715  RLH = 0.55677
Iteration    4  Log-likelihood = -741.37575  Chi Sq = 272.29717  RLH = 0.55677
*Converged after 0.16 seconds.

Log-likelihood for this model =       -741.37575
Log-likelihood for null model =       -877.52433
                                    ------------
                  Difference =        136.14858

Percent Certainty                =     15.51508
Consistent Akaike Info Criterion  =   1629.33661
Chi-Square                       =    272.29717
Relative Chi-Square              =     15.12762


         Effect        Std Err       t Ratio      Variable
  1      -2.24876       0.29872      -7.52793    ASC (1. Take Kids On (Chosen))
  2       2.12857       0.26948       7.89883    Destination_2 [Part Worth]
  3       1.53610       0.26802       5.73127    Destination_3 [Part Worth]
  4       1.09733       0.26986       4.06625    Destination_4 [Part Worth]
  5       0.42718       0.29629       1.44178    Destination_5 [Part Worth]
  6       0.84712       0.27678       3.06064    Destination_6 [Part Worth]
  7       1.29795       0.26998       4.80755    Destination_7 [Part Worth]
  8       0.15467       0.15400       1.00429    NumNights_2 [Part Worth]
  9       0.14704       0.15802       0.93053    NumNights_3 [Part Worth]
 10       0.08739       0.20502       0.42626    HotelStars_2 [Part Worth]
 11       0.10945       0.20079       0.54510    HotelStars_3 [Part Worth]
 12       0.00865       0.22360       0.03868    HotelStars_4 [Part Worth]
 13       0.08365       0.16983       0.49253    HotelType_2 [Part Worth]
 14       0.33321       0.15321       2.17488    HotelType_3 [Part Worth]
 15       0.04971       0.15586       0.31893    CarRental_2 [Part Worth]
 16       0.10914       0.15408       0.70833    CarRental_3 [Part Worth]
 17       0.03357       0.16401       0.20469    Price_2 [Part Worth]
 18       0.06858       0.17953       0.38197    Price_3 [Part Worth]
```

Since MBC software employs dummy-coding, the first levels of each categorical attribute are constrained to have utility = 0 (and are not shown in the report). For example, Destination #1 “Las Vegas” with a zero utility (the reference level) has a lower likelihood of predicting choice of being a good vacation package to *Take Kids On* than Destination # 2 (Orlando, FL) with a logit utility (Effect) of 2.12857.

REFERENCES

Alpert, Mark I (1971), “Identification of Determinant Attributes: A Comparison of Methods,” Journal of Marketing Research, Vol. 8.

Frazier, Curtis, Urszula Jones, and Katie Burdett (2006), “Brand Positioning Conjoint: A Revised Approach,” Sawtooth Software Conference Proceedings.

Gibson, Lawrence (2003), “Trade-Off vs. Self-Explication in Choice Modeling: The Current Controversy,” Sawtooth Software Conference Proceedings.

Glerum, Aurelie, Bilge Atasoy, and Michel Bierlaire (2014), “Using Semi-Open End Questions to Integrate Perceptions in Choice Models,” The Journal of Choice Modeling 10 (2014) 11–33.

Hutchinson, Harla (1989), “Gaining a Competitive Advantage by Combining Perceptual Mapping and Conjoint Analysis” Sawtooth Software Conference Proceedings.

Johnson, Richard (1999), “Product Mapping with Perceptions and Preferences,” Sawtooth Software Conference Proceedings.

Johnson, Richard and Bryan Orme (2003), “Getting the Most from CBC,” Technical paper available at http://www.sawtoothsoftware.com/education/techpap.shtml.

Orme, Bryan (2003), “Comment on Gibson,” Sawtooth Software Conference Proceedings.

Orme, Bryan (2010), “Menu-Based Choice Modeling Using Traditional Tools,” Sawtooth Software Conference Proceedings.

Pilon, Tom (1997), “Extensions to the Analysis of Choice Studies,” Sawtooth Software Conference Proceedings.

Poynter, Ray (1999), “But Why? Putting the Understanding into Conjoint,” Sawtooth Software Conference Proceedings.

Vriens, Marco and Curtis Frazier (2003), “Brand Positioning Conjoint: The Hard Impact of the Soft Touch,” Sawtooth Software Conference Proceedings.

PROFILE CBC: USING CONJOINT ANALYSIS FOR CONSUMER PROFILES

CHRIS CHAPMAN
KATE KRONTIRIS
JOHN S. WEBB
GOOGLE

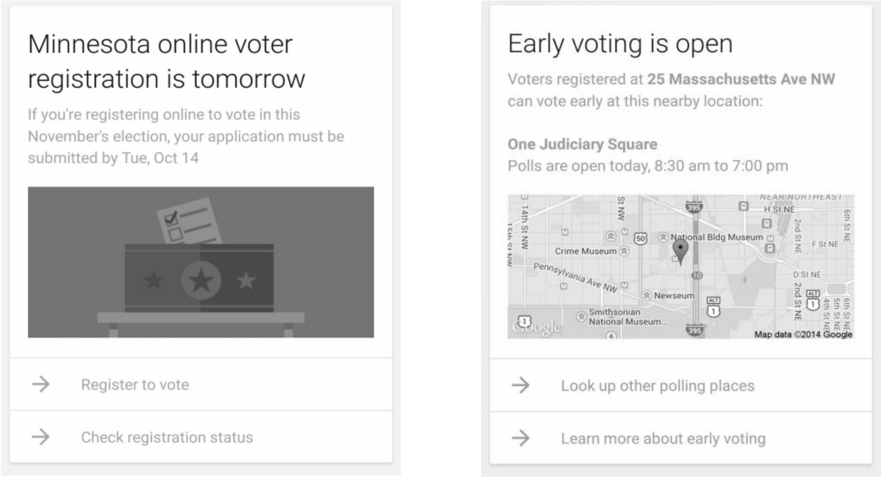
ABSTRACT

We investigate the usage of choice-based conjoint analysis (CBC) for sizing consumer profiles for a technology product area. Traditionally, technology research has often relied upon qualitative personas approaches that are difficult to assess quantitatively. We demonstrate that Profile CBC is able to find consumer profiles from tradeoffs of attributes derived from qualitative research, and yields replicable, specifically sized groups that are well-differentiated on both intra-method and extra-method variables. Thus, we conclude that Profile CBC is a potentially useful addition to analysts' tools for investigating consumer profiles.

INTRODUCTION: THE BUSINESS PROBLEM: SIZING CONSUMER PROFILES

The Google Social Impact team works on products and technical ecosystems for social good. This includes work on crisis response, civic innovation, and other social areas. For the project here, the team was interested to enhance civic engagement. As an example of the products this might inform, consider information served to users in advance of the November 2014 U.S. midterm election. The Civic Innovation team proactively served election information to Google Now users with four information designs; two such designs are shown in Figure 1.

Figure 1. Example Information Cards from Google Social Impact, October 2014



Minnesota online voter registration is tomorrow

If you're registering online to vote in this November's election, your application must be submitted by Tue, Oct 14

→ Register to vote

→ Check registration status

Early voting is open

Voters registered at **25 Massachusetts Ave NW** can vote early at this nearby location:

One Judiciary Square
Polls are open today, 8:30 am to 7:00 pm

→ Look up other polling places

→ Learn more about early voting

Serving these cards assumes that many users will find the information useful even though they might not have sought it. Previous qualitative research characterized such users as "interested bystanders," people interested in civic life yet who are not necessarily active participants or seekers of information (Krontiris et al., 2015).

Krontiris et al. (2015) documented civic *personas*, descriptions of prototypical (not actual) users. Personas are commonly used in technical product development to build product team awareness of users and to inspire design solutions. Personas may compile personal, behavioral, motivational, and product interest characteristics. An example excerpt is shown in Figure 2.

Figure 2. Excerpt from an Example Persona (Brechin, 2008)

> *Kathleen is 33 years old and lives in Seattle. She's a stay-at-home mom with two children: Katie, 7, and Andrew, 4. She drives the kids to school (usually carpooling with 2–3 other kids) in her Volvo wagon. Kathleen is thinking about buying the Sony rear-seat entertainment system she saw last weekend at Best Buy to keep the children occupied on the upcoming trip to see family in Canada.*

Before committing to projects that push civic information to users, the Google team wished to know how many people would benefit. Thus, the key business question was, "*How many interested bystanders are there [in the United States]*?" In other words, how many people might benefit from Google Now cards that proactively present information about civic events?

Difficulty with the Business Question

Unfortunately, as a qualitative description of a prototypical customer, a persona is not immediately sizable. In the present project, the qualitative research provided descriptions of purported representative interested bystanders but did not specify how many there were. This situation reflects two problems for personas: that, as pure descriptions, they are neither confirmable nor falsifiable (Chapman & Milham, 2006); and that, as composites of multiple dimensions, they fall prey to the curse of dimensionality. Once a persona comprises more than a few attributes, it is likely to match no one in an actual population (Chapman et al., 2008).

For these reasons, the first author had typically advised business stakeholders not to use qualitative personas in efforts to do market sizing. Instead, he has suggested that personas should be viewed as inspirational rather than descriptive. In this paper, however, we propose that choice-based conjoint analysis offers an appealing alternative that allows integration of multiple qualitative attributes while allowing quantitative sizing of groups.

Method: Choice-Based Conjoint Analysis Profiles, or Profile CBC

We addressed the problem of sizing the civic profiles using choice-based conjoint analysis (CBC), where the attributes were not product characteristics but were instead attitudinal and behavioral statements characteristic of persona attributes. The attitudes were derived from consumer characteristics that had been observed in the preceding qualitative research.

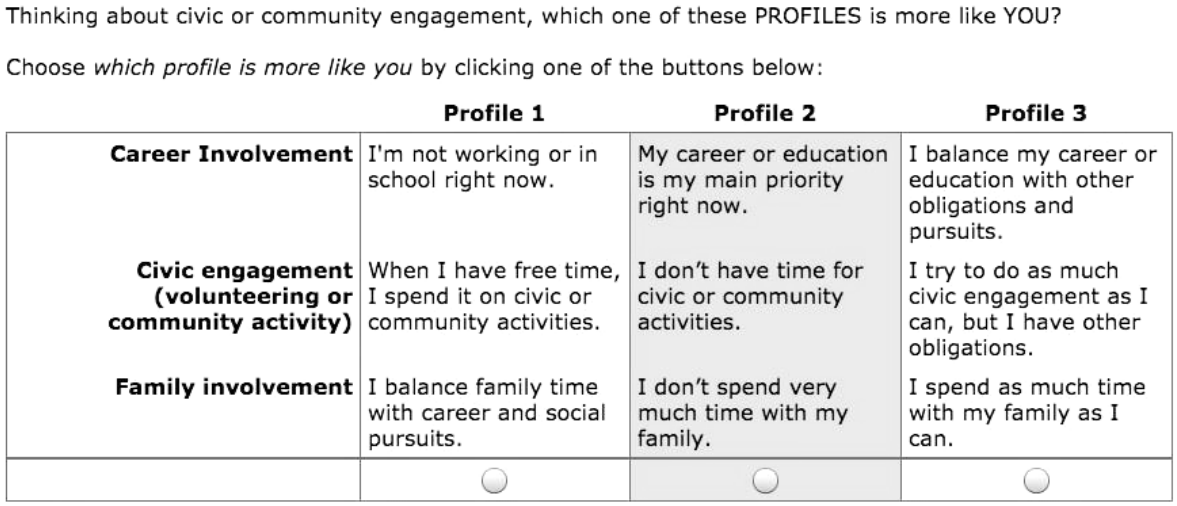
These characteristics were arranged into common areas (CBC attributes) comprising statements that could be considered to trade off against one another (hence, attribute levels). The list of characteristics included 8 areas (attributes) with 3–4 statements in each area (levels), for a total of 27 levels. Selected CBC attributes and example levels are shown in Figure 3.

Figure 3. CBC Attributes and example levels.
Attributes D–H are disguised in this paper.

Civic engagement	4 levels: I don't have time . . .; I try to do as much as I can . . .; etc.
Family engagement	3 levels: I don't spend very much time with my family; etc.
Career engagement	3 levels: My career or education is my main priority . . .; etc.
Attribute D	3 levels
Attribute E	3 levels
Attribute F	3 levels
Attribute G	4 levels
Attribute H	4 levels

This CBC design was fielded as a *partial profile CBC* (Chrzan and Elrod, 1995) such that each task presented *three concepts (profiles)*, where each profile comprised levels from *three* of the eight attributes. As we will describe below, we found this CBC format to be optimal for respondents' ability to perform the task. Also, as will be explained below, there was no "none" response option. An example task as fielded is shown in Figure 4.

Figure 4. An Example Partial Profile CBC Task, as Fielded.



Thinking about civic or community engagement, which one of these PROFILES is more like YOU?

Choose *which profile is more like you* by clicking one of the buttons below:

	Profile 1	Profile 2	Profile 3
Career Involvement	I'm not working or in school right now.	My career or education is my main priority right now.	I balance my career or education with other obligations and pursuits.
Civic engagement (volunteering or community activity)	When I have free time, I spend it on civic or community activities.	I don't have time for civic or community activities.	I try to do as much civic engagement as I can, but I have other obligations.
Family involvement	I balance family time with career and social pursuits.	I don't spend very much time with my family.	I spend as much time with my family as I can.
	○	○	○

Each respondent answered 12 tasks (with 3 profiles each). The attributes shown were randomly selected and ordered from all eight attributes and varied from task to task. The survey fielded a total of 500 variations of the 12-task questionnaire, created with the Sawtooth Software SSI/Web CBC module, and each respondent was randomly assigned to one of the 500 variants. Respondents were adults in the United States, obtained through an internet panel fielded by a third party market research supplier in October 2014. The data comprised N=2087 complete responses to the survey.

After data was collected, we identified profiles using aggregate latent class analysis of the conjoint utilities, conducted with Sawtooth Software CBC Latent Class (Sawtooth Software, 2004). As we discuss below, an alternative would have been to perform market simulation for specified profiles.

How Conjoint Analysis Solves the Sizing Problem

For experienced conjoint analysis analysts, the answer may be obvious: conjoint analysis provides utility estimates that allow determination of a probability estimate for each individual's match to a particular set of attributes (i.e., profile or persona), compared to other sets.

For analysts who are new to conjoint analysis, this works briefly as follows. For each respondent, a statistical model estimates a metric part-worth "utility" reflecting the preference for each of the attribute levels. The part-worths or utilities reflect the best estimate for likelihood to prefer one choice (in this case, one profile) in comparison to a specified set of alternative profiles, with likelihood proportional to the share of exponentiated summed utilities in the multinomial logit (MNL) model.

For illustration, consider utility values obtained for 2 levels (1 and 2) of two attributes (A and B), where each attribute level represents a statement as in Figures 3 and 4. Suppose for one respondent, utility(A1) = 0.5, utility(A2) = -0.4, utility(B1) = 1.1, and utility (B2) = 0.0. Suppose further, we are interested to compare two profiles for this respondent: profile 1 that comprises levels A1 & B2, and profile 2 that comprises A2 & B1.

Under MNL, the proportion of preference for each profile is expressed as share = *exp(*for one profile *x,* sum*(*utilities*(*profile *x))) /* sum*(*for all profiles *i, exp(*sum*(*utilities*(*profile *i)))*. In the present case, sum*(*utilities*(*profile *1)) = 0.5 + 0.0 = 0.5*, and sum*(*utilities*(*profile *2)) = -0.4 + 1.1 = 0.7*. This gives exponentiated values for profile 1 = *e^0.5 = 1.64* and profile 2 = *e^0.7 = 2.01*. Taking the share of preference ratios, the likelihood that this respondent matches profile 1 better than profile 2 is calculated as *1.64 / (1.64 + 2.01) = 45%*, and similarly the likelihood of better matching profile 2 is calculated as 55%.

Note that such allocation is only defined relatively within a specified set of profiles using various levels drawn from the same attributes; it does not answer the question whether some other, unknown profile might fit better. To identify the most likely profiles, rather than simulating them exhaustively we used latent class analysis to identify groups. In the discussion section below, we consider alternative methods to identify profiles.

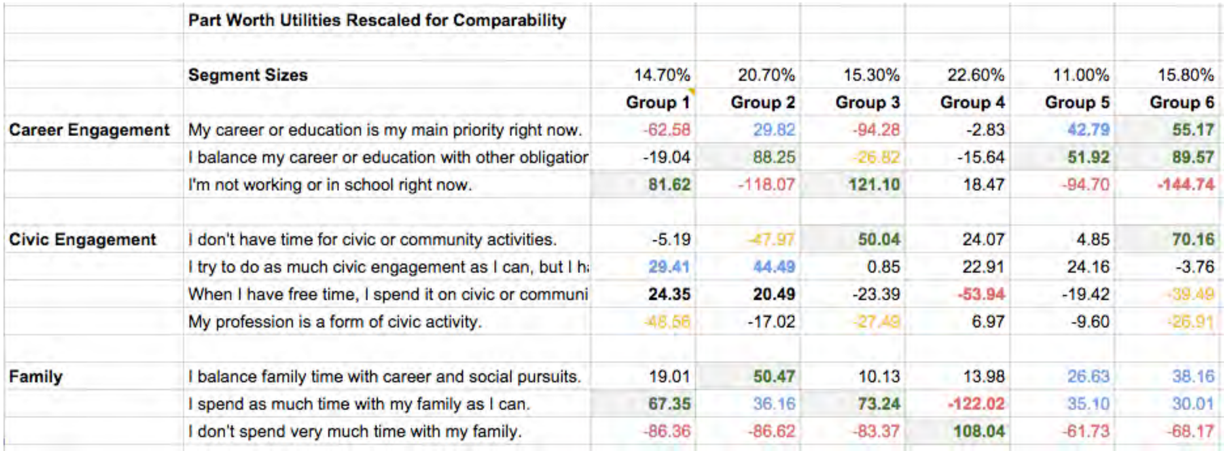
Such assessment is based on the respondent's own answers to the profile questions, and takes into account the contribution of each attribute for that respondent. In this way, it solves the difficulty of allocating respondents to profiles in the face of multidimensionality and imperfect matching. We leave aside details such as how part-worth estimates are calculated; for further discussion of MNL, we refer readers to Orme (2009).

Results: Answering the Business Question

Latent class solutions on the conjoint utilities were found for 2–10 classes, and a final solution of 6 classes was selected as preferable according to several criteria. In particular, the 6 class solution showed stronger fit indices (AIC and BIC) than solutions with fewer classes; the classes were qualitatively well differentiated and interpretable; the class sizes were relatively uniform, ranging 11%–23% of the sample; and solutions with more than 6 classes demonstrated weaker fit indices, less interpretable differentiation, or undesirably small groups (e.g., fewer than 5% of respondents in one of the classes). Among multiple versions of a 6 class solution proposed by CBC Latent Class, we retained the solution with the best fit index (AIC and BIC).

An overview of the 6 class solution results for the first three attributes is shown in Figure 5, which shows the mean part-worths (utility values) for each level of each attribute, by each group (note that the part-worths as shown are rescaled to be comparable and are not on the raw scale suitable for MLM calculation as shown above).

Figure 5. Excerpt of Part-Worth Means for the 6 Class Solution.
Part-worth values are shaded to indicate direction and magnitude, and are not raw utility values but have been rescaled to be comparable.



	Part Worth Utilities Rescaled for Comparability						
	Segment Sizes	14.70%	20.70%	15.30%	22.60%	11.00%	15.80%
		Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**
Career Engagement	My career or education is my main priority right now.	-62.58	29.82	-94.28	-2.83	42.79	55.17
	I balance my career or education with other obligatior	-19.04	88.25	-26.82	-15.64	51.92	89.57
	I'm not working or in school right now.	81.62	-118.07	121.10	18.47	-94.70	-144.74
Civic Engagement	I don't have time for civic or community activities.	-5.19	-47.97	50.04	24.07	4.85	70.16
	I try to do as much civic engagement as I can, but I h:	29.41	44.49	0.85	22.91	24.16	-3.76
	When I have free time, I spend it on civic or communi	24.35	20.49	-23.39	-53.94	-19.42	-39.49
	My profession is a form of civic activity.	-48.56	-17.02	-27.49	6.97	-9.60	-26.91
Family	I balance family time with career and social pursuits.	19.01	50.47	10.13	13.98	26.63	38.16
	I spend as much time with my family as I can.	67.35	36.16	73.24	-122.02	35.10	30.01
	I don't spend very much time with my family.	-86.36	-86.62	-83.37	108.04	-61.73	-68.17

In Figure 5, we see that the classes are well differentiated from one another across the rows. For example, in the "Career Engagement" attribute, Groups 1 and 3 often chose profiles without work or study, whereas Groups 2, 5, and 6 were likely to work. Additionally, each profile showed some attributes that were strongly loaded on it, within the columns. For instance, Group 4 is heavily identified as not spending time with family, and Groups 3 and 5 identified not having time for civic activities.

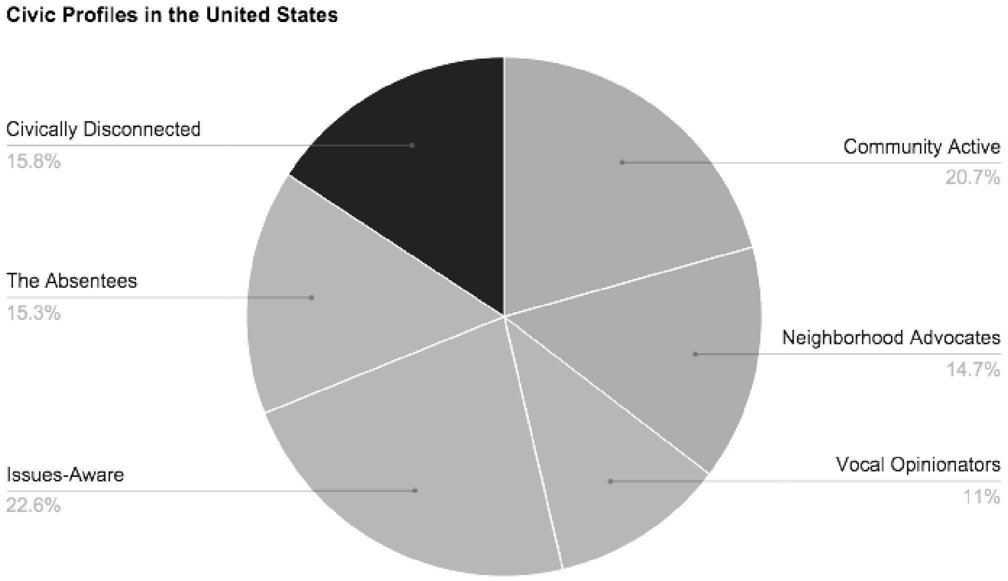
In short, the 6 class solution was interpretable, differentiated, and was free of the common but undesirable residual class (a class where no attribute is strongly associated, and the class is uninterpretable).

What about the business question? How many interested bystanders were there? Of the 6 classes, the utilities for 3 classes showed weak engagement in civic activities yet simultaneous high interest in civic happenings and information sources such as news. We identified these as matching the "interested bystander" profile; they correspond to groups 3, 4, and 5 in Figure 5.

Figure 6 presents the six groups with brief descriptive names and sizing. The interested bystander groups are the "Absentees," "Issues-Aware," and "Vocal Opinionator" groups, and comprise an estimated 48.9% of the respondents.

Given this breakdown of the groups' sizes, the business stakeholders concluded that there were enough interested bystanders to warrant further investigation of their needs and product design to meet those needs. Additional detail about civic engagement behaviors from the profiles (not shown here) provided more specific points to address with interested bystanders.

Figure 6. Sizing and Descriptive Titles for the Six Identified Profiles. Interested bystanders comprise the "Absentees," "Issues-Aware," and "Vocal Opinionator" groups.



External Correlates

A frequent outcome in segmentation projects is that class membership is strongly related to the basis variables used to classify people, in this case the conjoint utilities, yet the groups are weakly or not at all differentiated on other variables. In the present survey, we collected data separately from the CBC exercise on several other variables: household income, work status, gender, and self-reported frequency of voting. Figure 7 shows the group level means on those external covariates for each of the 6 civic profiles.

Figure 7. Mean by Civic Profile Class for Behavioral and Demographics Measures.

	Community Active	Neighborhood Advocates	Vocal Opinionators	Issues Aware	The Absentees	Civically Disconnected	*Range of group means*
Est'd mean income ($)	49,965	71,985	41,511	51,272	76,184	61,943	*34,673*
Employed full time	20%	66%	13%	41%	61%	59%	*53%*
Female proportion	58%	47%	66%	39%	47%	55%	*27%*
Report routine voting	55%	61%	37%	41%	48%	34%	*27%*

In Figure 7, we see that the six classes are once again well differentiated on the external variables. For example, full time employment ranges from 13% to 66% across the groups for a total 53-point spread from highest to lowest. There is a 27-point spread in gender and 27-point spread in reported voting frequency.

It is important to remember that these variables were not used in the profile determination and respondent assignment, and the clear differentiations here both confirm the importance of the profiles found and their external validity with regards to important civic behaviors. In other words, the Profile CBC method yielded profiles with important differences on other measures.

Discussion: Profile CBC Task Design

The present study reflects several rules of thumb for task design that the authors have formulated in the course of attempting Profile CBC in several product categories with several audiences. We offer these as best practice recommendations with the caveat that they are entirely based on our limited experience; we hope additional research will strengthen or modify them.

Our six suggested design principles are presented in Figure 8.

Figure 8. Suggested Design Principles for Profile CBC Questionnaires

1. Be careful to omit “must-have” attributes
2. Tasks should consist of 2 or 3 concepts
3. Concepts should present partial profile limited to 3 attributes
4. Tasks should not use a “none” option (especially single response “none”)
5. Tasks should not use allocation CBC
6. Careful investigation is needed before using ACBC

Principle 1—to omit “must have” attributes, means to ensure that all levels are actually suitable for trade-off by respondents. If a level is crucial to someone’s self-image or choice to the point that he or she would find it impossible to choose a conflicting profile, that level is better omitted (and perhaps could be used as an external validation measure instead). For instance, gender might fall into this category.

Principles 2 and 3—that tasks should present 2 to 3 concepts and no more than 3 attributes—reflect the cognitive difficulty of the task otherwise. With pre-testing, an analyst might cautiously relax these, but 3 profiles and 3 attributes is the maximum that we have found to be comfortable for respondents (cf. Patterson and Chrzan, 2003, for more on how respondents handle partial profile tasks). Principle 4—to avoid “none”—reflects the fact that respondents may use “none” with extreme frequency when selecting profiles if any level is an imperfect match. Because we are interested in their tradeoffs, it is difficult to interpret what “none” would mean in the context of Profile CBC.

Principle 5—to avoid allocation CBC—arises because of the cognitive complexity of “allocating” oneself across multiple profiles. Principle 6—to be cautious with ACBC—arises because of the difficulty in presenting screening tasks in ACBC. We have attempted but so far not succeeded in wording ACBC screening tasks in a way that works for respondents. If this problem were solved, we believe ACBC would have potential for Profile CBC.

Market Simulation: Alternative to Latent Class

In the present project, we began with qualitative personas, extracted their attributes, fielded a conjoint analysis study, and used latent class analysis (LCA) to determine profiles. This effectively discarded the original personas in favor of new ones (although largely similar to the previous personas) as found by LCA.

An alternative process would be not to use LCA, and instead to specify profiles after conjoint analysis whose attributes match those of the qualitative personas. We could then use the multinomial logit share formula or other market simulation techniques to determine the proportion of match for each of the specified profiles. This would be identical to the procedure outlined above in the section, "*How Conjoint Analysis Solves the Sizing Problem.*"

For the present study, we used the LCA profiles instead of market simulation for the qualitative personas for three reasons. First, the LCA results were similar enough overall to the personas that they were able to answer the business question and afforded the advantage of "letting the data speak." Second, LCA addresses gives quantitative guidance as to how many profiles there should be.

Finally, even when care is taken to select attributes, it is difficult to construct market simulations that precisely match a qualitative profile. For instance, suppose we are considering an attribute that is related to a persona but is of comparatively lesser importance than others. Should it be included in the market simulation or not? One could argue either way, and the choice will affect the share estimates. It cannot simply be determined by running both models because that would end up with cherry picking and steering the outcome. Because this situation may arise for many attributes across multiple profiles, it can lead to uncertainty in how to set up a market simulation.

With those caveats, we feel that market simulation is feasible and worthwhile when profiles are carefully constructed. Market simulation could also be used as a test of specific alternatives. For instance, if one asked, "Does this profile fit better than this other one?" it would be straightforward to put both into a market simulator and assess the relative shares of each.

Portfolio Modeling: Alternative to Latent Class Analysis

We used LCA to find the classes here, but there are several alternatives. As noted above, one simple alternative is to specify the classes directly and use a market simulator to size them. One possibility is to apply statistical procedures that have various assumptions other than the LCA methods used here. For instance, one might use Gaussian finite mixture models (cf. Benaglia et al., 2009).

Another alternative is a portfolio modeling approach that attempts to find the best set of profiles to match the respondents, as based on an overall fit criterion such as proportion of respondents matching or maximum likelihood. These may be performed through various iterative search techniques (cf. Chapman and Alford, 2010). For general optimization methods, a crucial question to consider is whether one needs answer the problem of a "none" parameter. If an algorithm simply maximizes total share of people allocated to groups, then it will always "succeed" by allocating 100% to a single group unless there is a criterion such as a none utility that prevents such allocation. However, as noted above, the "none" concept is difficult to express here; this problem—of how to express "none" and model it—is an area ripe for investigation.

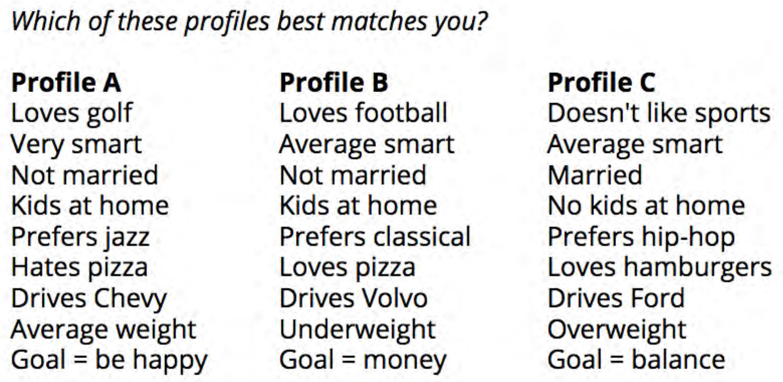
MaxDiff: An Alternative to CBC

MaxDiff is a potential alternative to choice-based conjoint analysis to field these tradeoffs. In particular, if the attributes on a profile are considered to be a list of characteristics that each might or might not apply and are not necessarily arranged into crisp groups (i.e., attributes), then a MaxDiff approach should work well. Additionally, MaxDiff might be conceptually simpler for respondents. Latent class analysis would work with MaxDiff responses nearly identically to the procedure we described here for Profile CBC.

Additional Notes on Partial Profile Designs

We have argued that partial profile tasks make Profile CBC possible. We believe that fielding a choice task with more than a few attributes that ask about self-identification all at once, such as the fictional one shown in Figure 9, is simply infeasible. In pre-tests, respondents balked at such a task.

Figure 9. An Example (Fictional) Profile CBC Task that Avoids Partial Profile Design but May Be Impossible for Respondents.



Which of these profiles best matches you?

Profile A	**Profile B**	**Profile C**
Loves golf	Loves football	Doesn't like sports
Very smart	Average smart	Average smart
Not married	Not married	Married
Kids at home	Kids at home	No kids at home
Prefers jazz	Prefers classical	Prefers hip-hop
Hates pizza	Loves pizza	Loves hamburgers
Drives Chevy	Drives Volvo	Drives Ford
Average weight	Underweight	Overweight
Goal = be happy	Goal = money	Goal = balance

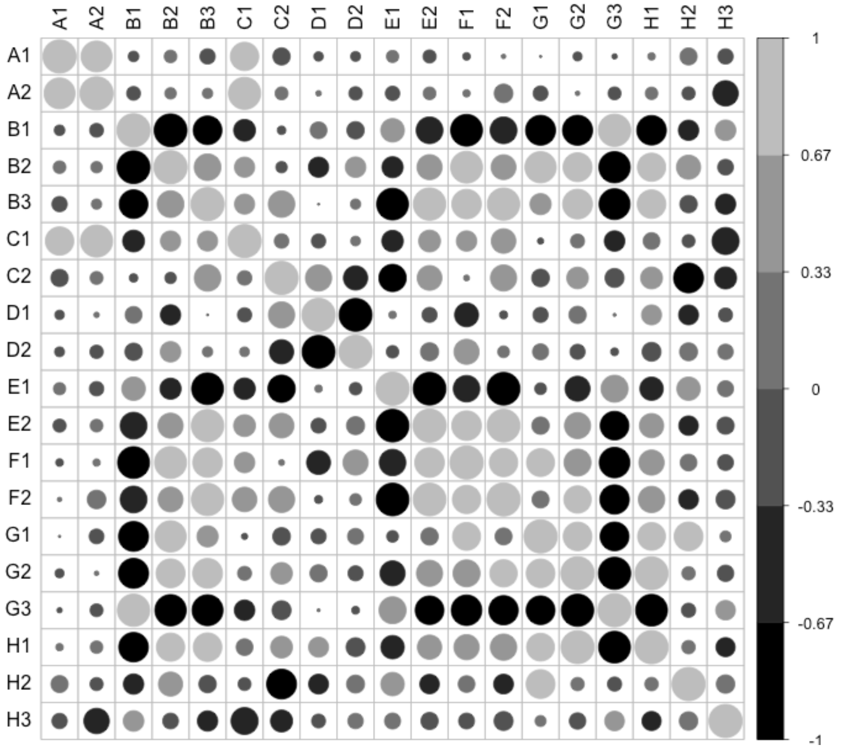
However, there are potential concerns with partial profile designs. For one, the required sample sizes will increase substantially; instead of a few hundred respondents, more might be required because each task provides less information. We suggest testing the design matrix to ensure there is adequate power for the intended sample size.

Because partial profile designs do not show all attributes in each task, they may underestimate the extent to which attributes are correlated. To see why this is, suppose that attributes A and B are very closely related. In a partial profile design, A and B often will not appear together. Thus, when each is observed without the other, it appears to contribute the full effect by itself in impact on the choice task, and this does not reflect its overlap with the other attribute. Additionally, because A and B often appear separately, there is corresponding less opportunity to assess tradeoffs between their various feature levels. This issue of potential attribute correlation should be examined at design time with respect to theory and previous findings, subjected to pre-testing before fielding a final survey, and examined post hoc for either excessive correlation or absence of expected correlation.

In the present study, we investigated attribute correlation qualitatively before fielding, and empirically after fielding the study. Figure 10 presents the Pearson's *r* correlation matrix for part-worth utilities found in this study, where the circle shading indicates direction (positive correlation is lighter and negative correlation is darker) and circle size indicates magnitude (plotting method from Wei, 2013). Overall, we see that several of the attributes are substantially correlated (e.g., attributes B, E, F, G, and H). These correlations were expected on theoretical bases for the attributes in question, and thus the correlations were confirmatory. Likewise, much of Figure 10 shows correlations of low magnitude (small circles) between levels; this was likewise confirmatory for attributes that were expected to have lesser levels of association.

Overall, we conclude that the partial profile method is likely required for Profile CBC, and that the problems of power and attribute correlation may be managed with attention and post hoc empirical inspection. To review more about partial profile concerns, see several papers in previous Sawtooth Software Conference proceedings (e.g., Huber, 2012; Yardley, 2013).

Figure 10. Correlation Matrix for the Attributes in the Present Study, N=2087. The final level in each attribute has been omitted. Circle size is proportional to absolute magnitude of correlation, and hue indicates direction.



Conclusion

The Profile CBC method outlined here demonstrates that choice-based conjoint analysis may be useful in situations where analysts seek to find and size clusters of respondents who identify with profile-like descriptions. Because Profile CBC allows incorporation of qualitative self descriptions as attributes, finds classes with a replicable procedure, and determines class size, it overcomes key limitations of purely qualitative personas. In the present study, we also observed that the classes showed substantial discrimination on external validation measures. Thus, when basic design cautions are observed, Profile CBC opens exciting new areas of exploration for conjoint analysts.

Chris Chapman

Kate Krontiris

John S. Webb

References

Benaglia, T., Chauveau, D., Hunter, D.R., Young, D. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 32(6), 1–29. http://www.jstatsoft.org/v32/i06/ (Last retrieved May 2, 2015.)

Brechin, E. (2008). Reconciling market segments and personas. Cooper Design. http://goo.gl/btJXsP (Last retrieved May 4, 2015.)

Chapman, C.N., and Alford, J.L. (2010). Product Portfolio Evaluation Using Choice Modeling and Genetic Algorithms. *Proceedings of the 15th Sawtooth Software Conference*, Newport Beach, CA, October 2010. https://goo.gl/fM86kp (Last retrieved May 4, 2015.)

Chapman, C.N., Milham, R.P. (2006) The personas' new clothes: methodological and practical arguments against a popular method. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50:5, 634–636. SAGE Publications. https://goo.gl/szQ54E (Last retrieved May 4, 2015.)

Chapman, C.N., Love, E., Milham, R.P., ElRif, P., and Alford, J.L. (2008). Quantitative evaluation of personas as information. Proceedings of the Human Factors and Ergonomics Society (HFES) 52nd Annual Conference, New York, NY, September 2008. http://goo.gl/4rLYEO (Last retrieved May 4, 2015.)

Chrzan, K., and Elrod, T. (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes." Presented at the 1995 Advanced Research Techniques Forum, Monterey, CA.

Huber, J. (2012) CBC Design for Practitioners: What Matters Most. *Proceedings of the 16th Sawtooth Software Conference*, Orlando, FL, March 2012. http://goo.gl/ieqgMK (Last retrieved May 4, 2015.)

Krontiris, K., Webb, J., Krontiris, C., Chapman, C. (2015). Understanding America's "Interested Bystander:" A Complicated Relationship with Civic Duty. Technical report, Google Civics Research Workshop, New York, NY, January 2015.

Orme, B.K. (2009). Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, 2nd edition. Madison, WI: Research Publishers.

Patterson, M., and Chrzan, K. (2003). Partial Profile Discrete Choice: What's the Optimal Number of Attributes. *Proceedings of the 10th Sawtooth Software Conference*, San Antonio, TX, April 2003. https://goo.gl/uNqjTt (Last retrieved May 4, 2015.)

Sawtooth Software (2004). The CBC Latent Class Technical Paper (Version 3). Sawtooth Software, Sequim, WA. http://goo.gl/n2ZmAu (Last retrieved May 4, 2015.)

Wei, T. (2013). corrplot: Visualization of a correlation matrix. R package version 0.73. http://CRAN.R-project.org/package=corrplot (Last retrieved May 4, 2015.)

Yardley, D. (2013). Attribute Non-Attendance in Discrete Choice Experiments. *Proceedings of the 17th Sawtooth Software Conference*, Dana Point, CA, October 2013. http://goo.gl/2Yg91Z (Last retrieved May 4, 2015.)

RUM & RRM: IMPROVING THE PREDICTIVE VALIDITY OF CONJOINT RESULTS?

JEROEN HARDON
KEES VAN DER WAGT
SKIM GROUP

INTRODUCTION

Different from the Random Utility Model (RUM), Random Regret Modeling (RRM) is based on the assumption that consumers do not choose the option that maximizes overall utility but by avoiding regret of not choosing another option. As RRM assumes the choice is based on which alternatives are offered, it takes the context of a choice into account, in contrast with RUM. RRM may be more representative for the way some consumers make their choices and thereby it may help to improve the predictive validity of a model. In this paper we show how we combine RUM and RRM to yield the best predictive validity, taking into account how some consumers choose to maximize utility while others choose to avoid regret.

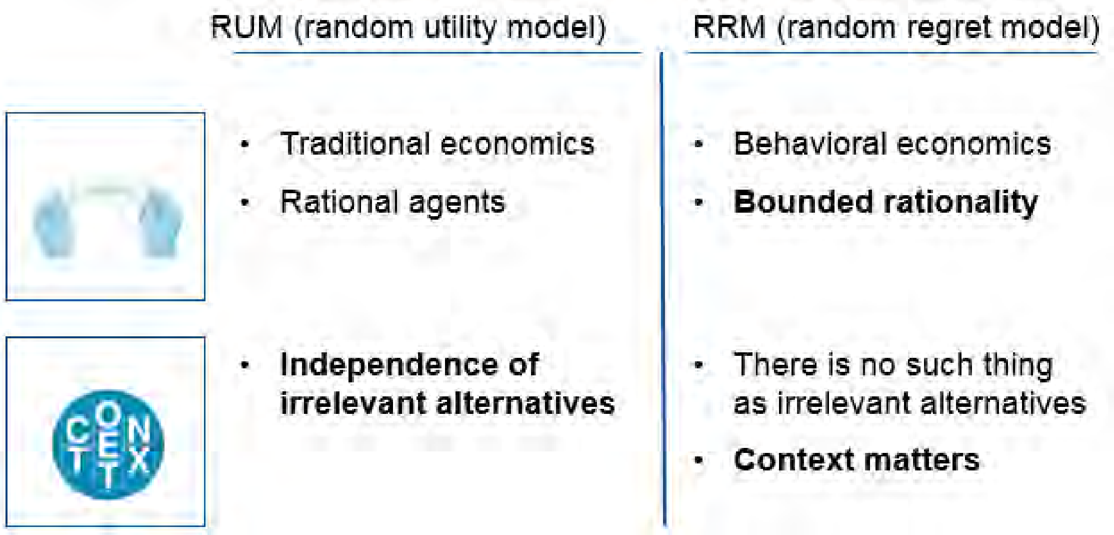
A SHORT INTRODUCTION TO RANDOM REGRET MODELING

Random Utility Modeling (RUM) is widely used to estimate the preferences or "utilities" of product characteristics. Those utilities help us do "what-if" analysis: What happens if we change the price of our products? What happens if competition does? How should we react on the competitor price changes? The model works well but it has its limitations. The Share of Preference model commonly used to play what-if games suffers from "Independence of Irrelevant Alternatives" (IIA). According to IIA, removing unchosen alternatives should not affect someone's choice.

However, if we improve a product by changing its characteristics in our "what-if" analyses (with RUM using Multinomial Logit for the likelihood function), it gains share from all other products proportionally to the other products' shares. Similarly, when a product loses share, it loses to other products proportionally with the other products' shares. Hence, IIA actually is an unrealistically simple assumption. In the real world, products compete with each other in a disproportionate way. If we improve an existing product, it usually gains most from a subset of products with which it competes most directly.

Random Regret Modeling (RRM) is one way of reducing this problem. It assumes that people compare a product with all alternatives on every characteristic, avoiding that they choose a product that is outperformed by an alternative on one or more characteristics. The RRM model assumes that as soon as people make tradeoffs, they run the risk of regret: usually there is at least one non-chosen alternative that out-performs a chosen product on one or more characteristics.

Figure 1. The Differences between RUM and RRM.



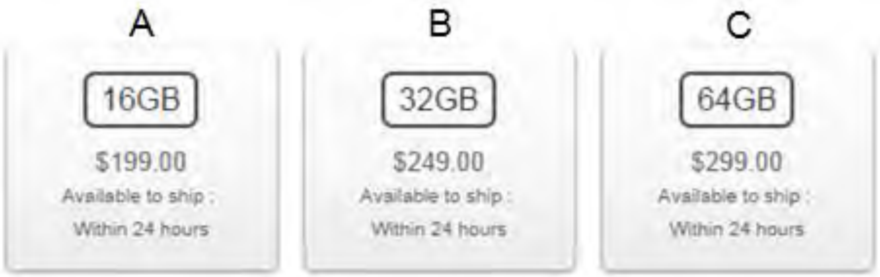
RUM (random utility model)	RRM (random regret model)
- Traditional economics - Rational agents	- Behavioral economics - **Bounded rationality**
- **Independence of irrelevant alternatives**	- There is no such thing as irrelevant alternatives - **Context matters**

The most important differences between the two models are shown in Figure 1.

Coding of RRM

RRM is coded differently compared to RUM, as the RRM coding is based on the alternatives and not at the product itself. Explaining how RRM is coded is easiest by means of an example. Let’s assume we are in the market to buy an iPod. There are 3 products, A, B and C. (Figure 2).

Figure 2. Three iPod Products.



A	B	C
16GB	32GB	64GB
$199.00	$249.00	$299.00
Available to ship :	Available to ship :	Available to ship :
Within 24 hours	Within 24 hours	Within 24 hours

When a product is superior on a certain characteristic, this does not lead to regret. When a product is inferior, this does lead to regret. In this example this would lead to: Product B has 16 GB more compared to product A, Product B has 32 GB less compared to product C. So product B has 0 + 32 = 32 regret on GB.

Product B is $50 more expensive compared to product A, Product B is $50 less expensive compared to product C. So product B has 50 + 0 = 0 regret on price. In order to overcome any potential scaling issues, the regret parameters are rescaled to be between 0 and 2.

RRM Limitations

Taking context into account sounds like a great idea, except it is not that easy. RRM has its limitations:

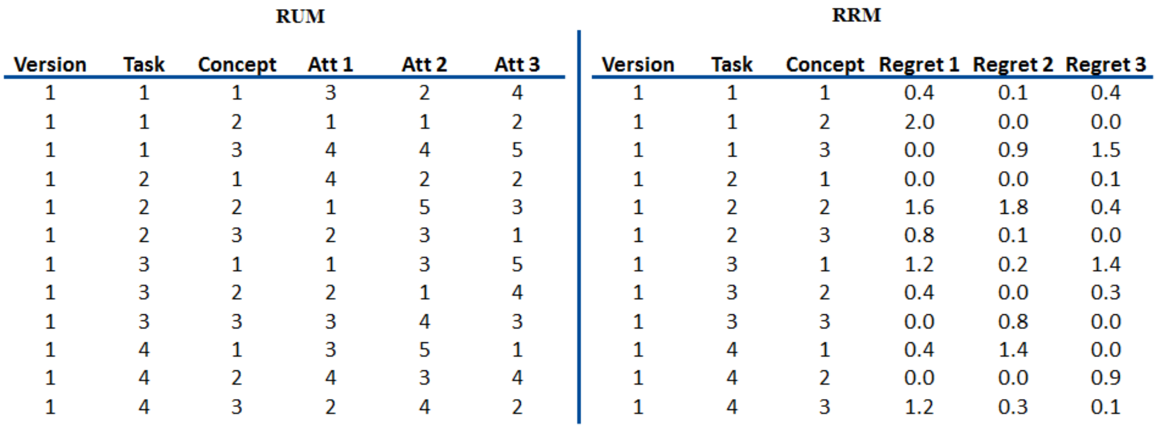
- You can only simulate the same amount of concepts or products you tested in the choice task. This is due to the way regret is coded.
- Using a regular none is not possible. This is due to the fact we cannot calculate the regret compared to the "none."
- Not that many studies/attributes are applicable.
    - Only ordinal attributes can be coded as regret.
    - Telecom studies seemed to be nice, but how to code regret of unlimited amount of minutes vs. 250min?
- For many studies we found that only price could be coded as regret.
- IIA is sometimes considered as an advantage by practitioners.
    - It allows simulation of a varying number of alternatives, when only a subset of alternatives is used in choice set with RUM.
    - However, for RRM, correct specification of a choice set is crucial.
- Conjoint designs in marketing research often are complicated.
    - For categorical variables the RRM and RUM predict equal market shares.
    - The concept of regret with respect to the none option is unclear, making it infeasible to estimate.
- Estimation with RRM restricts flexibility with forecasting.
    - The parameter size in RRM is correlated with the choice set size, since regret is composed of the summation of positive terms, see next paragraph.
    - Requires same size choice set for forecasting as for estimation.

THE "DUCT TAPE" SOLUTION (HYBRID SOLUTION)

We have many reasons to stay with the RUM coding, as it is the current workhorse model for conjoint analysis. It is well known, we know it works and it is well understood in the industry. Yet, RRM remains very interesting as it is a semi-compensatory model. Due to the compromise effect, RRM simulations can potentially outperform RUM simulations when there are attribute levels with intermediate utility. And last but not least, RRM does not impose IIA.

Because we see the advantage of both models, we figured why not use both at the same time. In Figure 3 you see the RRM way of coding and the RRM way of coding.

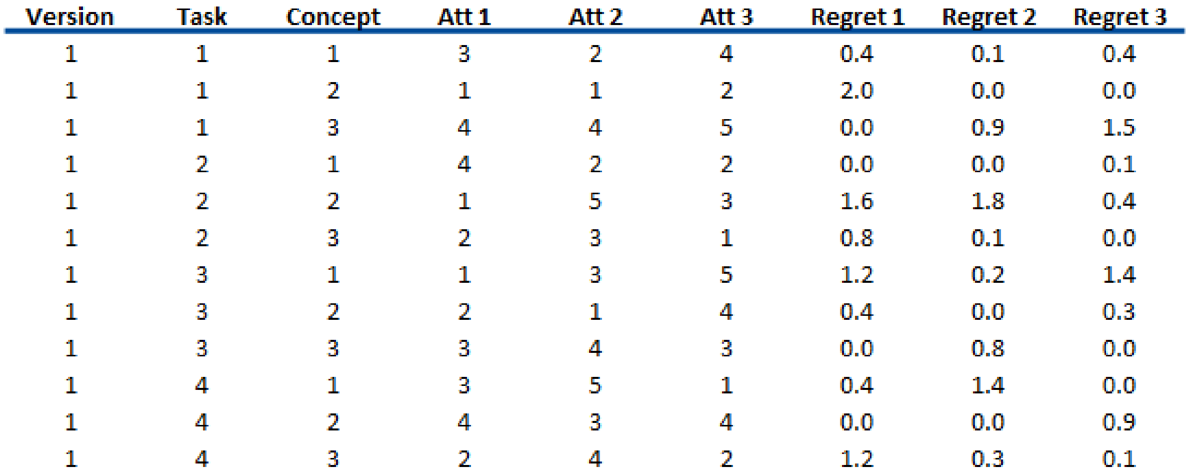
Figure 3. Coding of RUM and RRM



RUM						RRM					
Version	Task	Concept	Att 1	Att 2	Att 3	Version	Task	Concept	Regret 1	Regret 2	Regret 3
1	1	1	3	2	4	1	1	1	0.4	0.1	0.4
1	1	2	1	1	2	1	1	2	2.0	0.0	0.0
1	1	3	4	4	5	1	1	3	0.0	0.9	1.5
1	2	1	4	2	2	1	2	1	0.0	0.0	0.1
1	2	2	1	5	3	1	2	2	1.6	1.8	0.4
1	2	3	2	3	1	1	2	3	0.8	0.1	0.0
1	3	1	1	3	5	1	3	1	1.2	0.2	1.4
1	3	2	2	1	4	1	3	2	0.4	0.0	0.3
1	3	3	3	4	3	1	3	3	0.0	0.8	0.0
1	4	1	3	5	1	1	4	1	0.4	1.4	0.0
1	4	2	4	3	4	1	4	2	0.0	0.0	0.9
1	4	3	2	4	2	1	4	3	1.2	0.3	0.1

Our hybrid solution merges the two coding methods, which is represented in Figure 4.

Figure 4. Coding of the Hybrid Method



Version	Task	Concept	Att 1	Att 2	Att 3	Regret 1	Regret 2	Regret 3
1	1	1	3	2	4	0.4	0.1	0.4
1	1	2	1	1	2	2.0	0.0	0.0
1	1	3	4	4	5	0.0	0.9	1.5
1	2	1	4	2	2	0.0	0.0	0.1
1	2	2	1	5	3	1.6	1.8	0.4
1	2	3	2	3	1	0.8	0.1	0.0
1	3	1	1	3	5	1.2	0.2	1.4
1	3	2	2	1	4	0.4	0.0	0.3
1	3	3	3	4	3	0.0	0.8	0.0
1	4	1	3	5	1	0.4	1.4	0.0
1	4	2	4	3	4	0.0	0.0	0.9
1	4	3	2	4	2	1.2	0.3	0.1

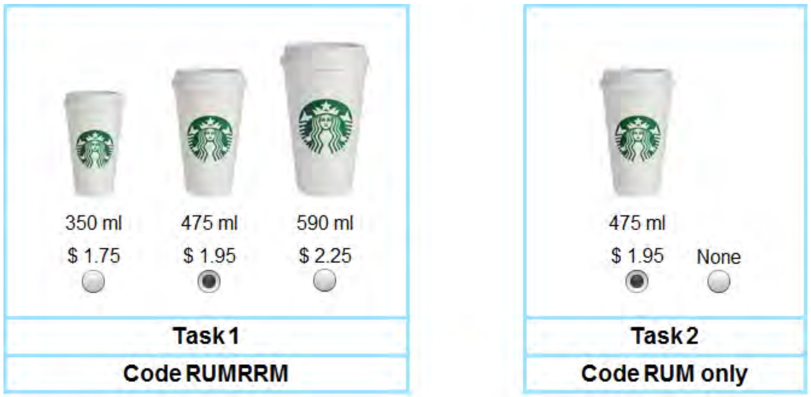
We had several reasons in favor of this hybrid method:

- It allows having the best of both worlds, while keeping the coding and the estimation process fairly easy.
- We see a lot of advantages:
    - We like context effects, and they are modelled
    - We like the dual response none, and this can be modelled (see next paragraph)
    - No need for new software as all can be modelled with standard software (i.e., CBC/HB)
    - The regret parameters can be estimated by means of user specified coding in CBC/HB

Dual Response None

Before we indicated RRM was not possible if a none was present. With the hybrid solution we found a way to include a so-called dual-response none. We will need to code the dual-response none slightly different, compared to what the standard Sawtooth Software provides. We code each task with the dual-response none as two tasks, where the first task is just the choice between the products. On this task we code both RUM and RRM. The second task always consists of the chosen concept and a none parameter. On this task we code RUM only. The assumption here is that the actual purchase decision is based on RUM, and RRM does not have an impact. The idea is illustrated in Figure 5.

Figure 5. Coding the Dual Response None Using the Hybrid Model



Two Test Studies

We used 2 datasets to check our hypothesis:

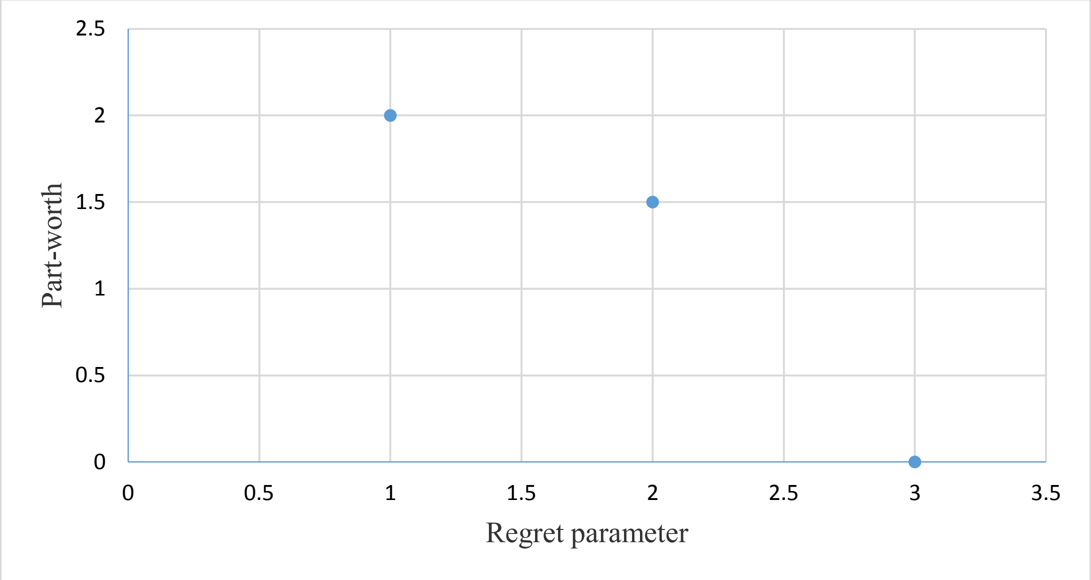
1. Health insurance study
2. Tablet study

We had data from 1 more study, but the data structure made us not able to use our hybrid model. The structure of the data was as follows:

- 8 attributes (7 with 3 levels, 1 with 2 levels)
- 8 tasks, all with 3 concepts
- Design strategy: Complete enumeration

While coding this study we found that when all levels of an attribute are on screen, the part-worth levels and the regret coding are 100% correlated. This means that the hybrid model cannot be estimated as the data is ill-conditioned. This is due to the 1-on-1 relationship this creates, see Figure 6 for an example.

Figure 6. The 1-on-1 Relationship between Part-Worth and Regret



Having all levels of an attribute always on screen without overlap will always result in the best level having no (zero) regret and the worst level at maximum regret. Idem for the other levels, the regret will always be the same, as the context will always be identical.

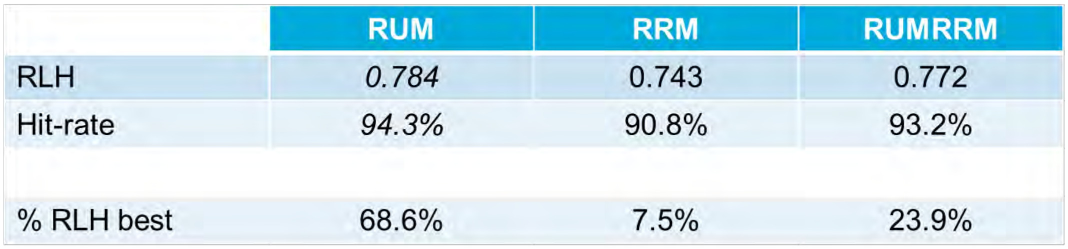
TEST STUDY 1—HEALTH INSURANCE

The structure of the data was as follows:

- 4 attributes of which 3 are nominal (with 4, 3, 4, 5 levels)
- 15 tasks, all with 3 concepts
- Design strategy: Complete enumeration
- 1245 respondents

We ran 3 separate models, RUM, RRM and the hybrid model. For the RUM coding we used part-worth and the RRM parameters were linear. In all runs we constrained the nominal attributes to follow logic order. In this study we compared RLH and hitrate as you can see in Figure 7.

Figure 7. The Results of the Health Insurance Study



	RUM	RRM	RUMRRM
RLH	*0.784*	0.743	0.772
Hit-rate	*94.3%*	90.8%	93.2%
% RLH best	68.6%	7.5%	23.9%

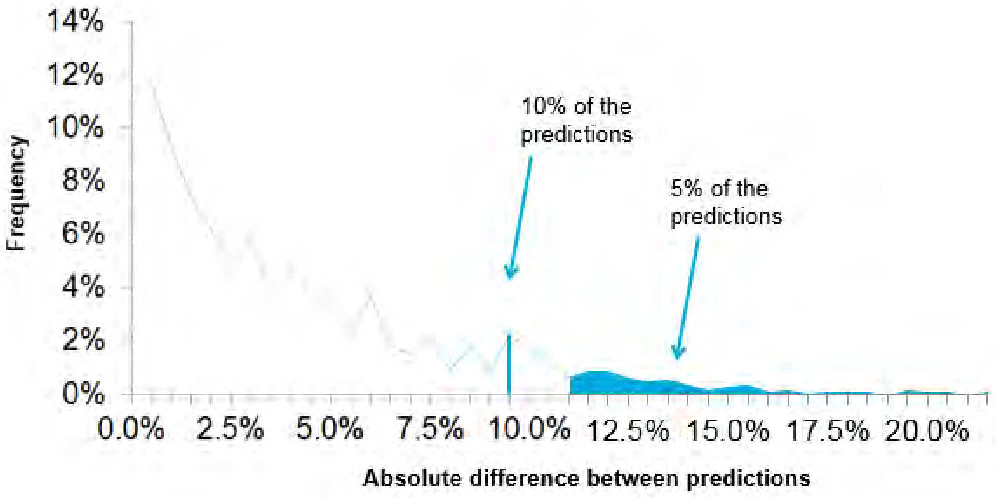
*Highest values in italic

The % RLH best is checking which RLH is best per respondent. RUM shows to have the most winning RLH scores, but looking at the other results we don't see a clear winner and the results are quite similar. We checked the correlation between the RLH scores of the different models, and the average correlation across the 3 methods is 0.97.

Having these similar results made us dig deeper into the data. We found that although the first choice hit rates we similar, the predicted preference share were not.

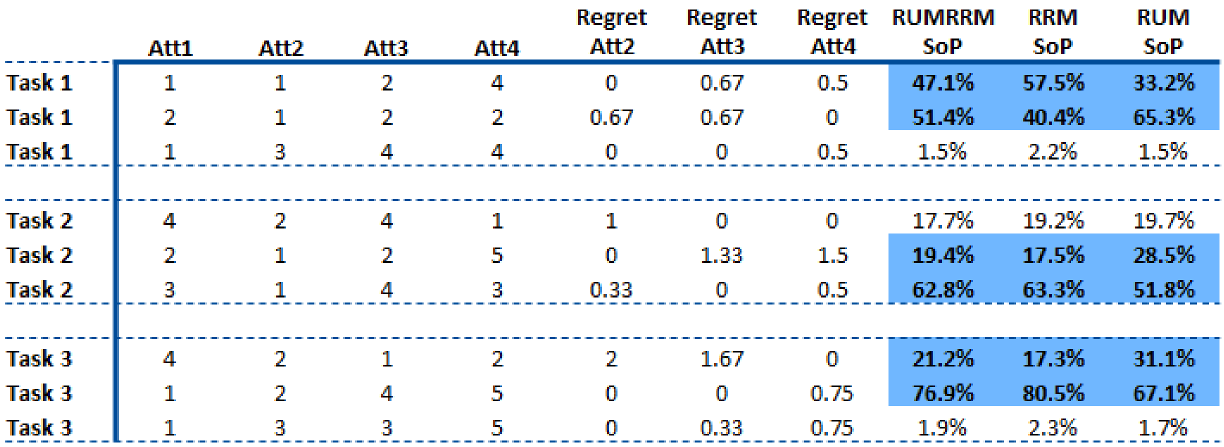
We compared the preference shares for each task of each respondent. We checked the absolute difference between the predictions, and found that in 5% of our predictions, we were 11.5% or more off. With 10% of our predictions this was 9.5%. This is illustrated in Figure 8.

Figure 8. Overview of Absolute Differences between Predictions



In Figure 9 we show 3 examples of predicted shares, based on the 3 models.

Figure 9. Example of 3 Task Predictions



	Att1	Att2	Att3	Att4	Regret Att2	Regret Att3	Regret Att4	RUMRRM SoP	RRM SoP	RUM SoP
Task 1	1	1	2	4	0	0.67	0.5	47.1%	57.5%	33.2%
Task 1	2	1	2	2	0.67	0.67	0	51.4%	40.4%	65.3%
Task 1	1	3	4	4	0	0	0.5	1.5%	2.2%	1.5%
Task 2	4	2	4	1	1	0	0	17.7%	19.2%	19.7%
Task 2	2	1	2	5	0	1.33	1.5	19.4%	17.5%	28.5%
Task 2	3	1	4	3	0.33	0	0.5	62.8%	63.3%	51.8%
Task 3	4	2	1	2	2	1.67	0	21.2%	17.3%	31.1%
Task 3	1	2	4	5	0	0	0.75	76.9%	80.5%	67.1%
Task 3	1	3	3	5	0	0.33	0.75	1.9%	2.3%	1.7%

As you can see on task 2, we predict the first choice the same for all 3 models, but looking closer we see that there are large differences in shares, and the rank-order is not the same. The same data could potentially lead to different recommendations.

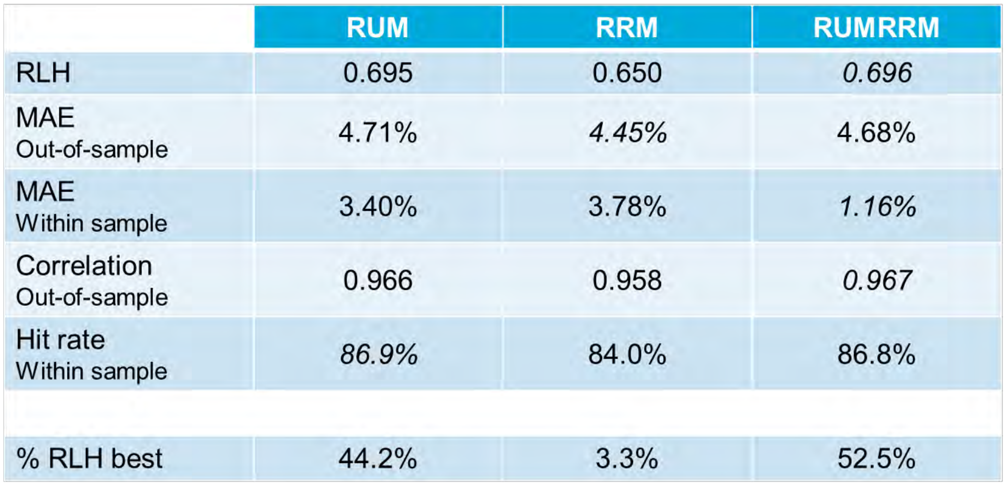
Test Study 2—Tablets

The structure of the data was as follows:

- 6 attributes, of which 5 are nominal (with 5,4,4,3,5,5 levels)
- 15 tasks, all with 3 concepts (3 holdout tasks)
- Design strategy: Overlap
- Total of 1247 respondents
    - 931 answered 12 CBC tasks, plus 3 holdout
        - half of respondents received choice sets constructed with a minimum overlap
        - the other half had a design that allowed level overlap
    - 316 respondents answered 15 tasks, constructed at random, hence a lot of overlap
        - This cell served as a holdout for out-of-sample validity checks

Here we also ran all three models, please find the results below in Figure 10.

Figure 10.The Results of the Tablet Study



	RUM	RRM	RUMRRM
RLH	0.695	0.650	*0.696*
MAE Out-of-sample	4.71%	*4.45%*	4.68%
MAE Within sample	3.40%	3.78%	*1.16%*
Correlation Out-of-sample	0.966	0.958	*0.967*
Hit rate Within sample	*86.9%*	84.0%	86.8%
% RLH best	44.2%	3.3%	52.5%

*Highest values in italic

Again the results are quite similar. We checked the correlation between the RLH scores of the different models, and the average correlation across the 3 methods is 0.95.

The same pattern occurs when we look at the absolute differences between predictions. In this case, in 5% of our predictions, we were 6.0% or more off. With 10% of our predictions this was 4.0%. This is illustrated in Figure 11.

Figure 11. Overview of Absolute Differences between Predictions

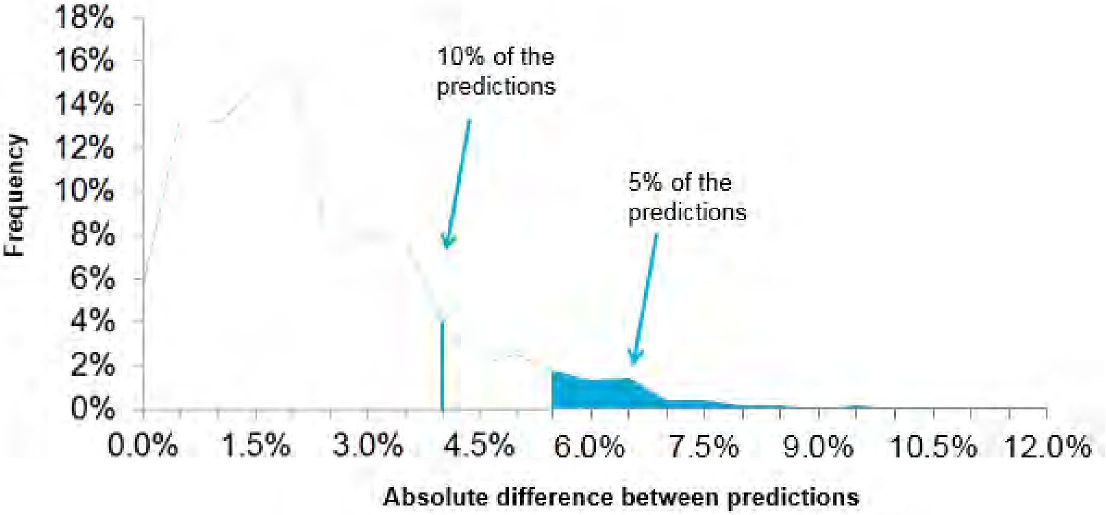
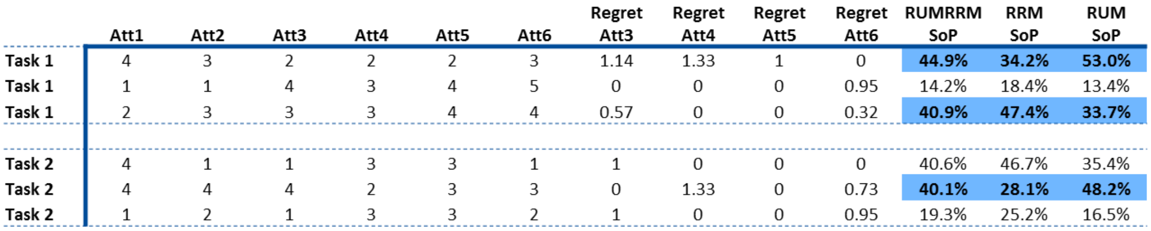


Figure 12. Example of 2 Task Predictions

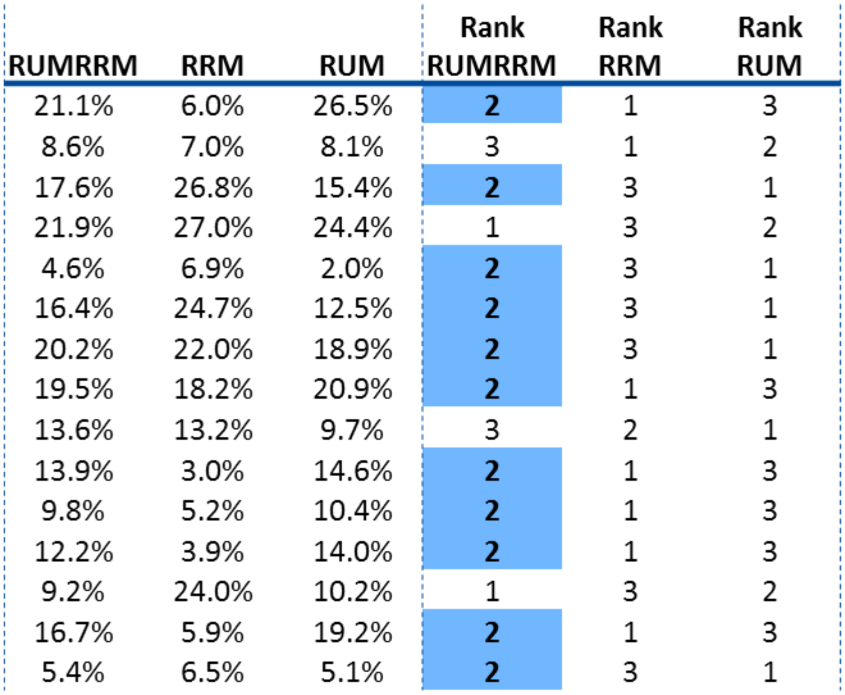


	Att1	Att2	Att3	Att4	Att5	Att6	Regret Att3	Regret Att4	Regret Att5	Regret Att6	RUMRRM SoP	RRM SoP	RUM SoP
Task 1	4	3	2	2	2	3	1.14	1.33	1	0	**44.9%**	**34.2%**	**53.0%**
Task 1	1	1	4	3	4	5	0	0	0	0.95	14.2%	18.4%	13.4%
Task 1	2	3	3	3	4	4	0.57	0	0	0.32	**40.9%**	**47.4%**	**33.7%**
Task 2	4	1	1	3	3	1	1	0	0	0	40.6%	46.7%	35.4%
Task 2	4	4	4	2	3	3	0	1.33	0	0.73	**40.1%**	**28.1%**	**48.2%**
Task 2	1	2	1	3	3	2	1	0	0	0.95	19.3%	25.2%	16.5%

Here as well, the same data leads to different "market leaders," and different recommendations (Figure 12).

Looking at out-of-sample predictions, we found the table in Figure 13.

Figure 13. Sum and Rank of Out-of-Sample Sum of Absolute Differences



RUMRRM	RRM	RUM	Rank RUMRRM	Rank RRM	Rank RUM
21.1%	6.0%	26.5%	**2**	1	3
8.6%	7.0%	8.1%	3	1	2
17.6%	26.8%	15.4%	**2**	3	1
21.9%	27.0%	24.4%	1	3	2
4.6%	6.9%	2.0%	**2**	3	1
16.4%	24.7%	12.5%	**2**	3	1
20.2%	22.0%	18.9%	**2**	3	1
19.5%	18.2%	20.9%	**2**	1	3
13.6%	13.2%	9.7%	3	2	1
13.9%	3.0%	14.6%	**2**	1	3
9.8%	5.2%	10.4%	**2**	1	3
12.2%	3.9%	14.0%	**2**	1	3
9.2%	24.0%	10.2%	1	3	2
16.7%	5.9%	19.2%	**2**	1	3
5.4%	6.5%	5.1%	**2**	3	1

You can see that in most cases (73%), the hybrid model is the safest bet. In Dutch we would say: The golden middle road.

Conclusions and Recommendations

To assume IIA or not, that seems to be the question.

- Traditionally with RUM, we assume Independence of Irrelevant Alternatives (IIA), while alternatives could be relevant.
- Now with RRM, we assume context matters, whereas the opposite might be the case.
- Taking a bit of both seems like a good compromise.

Summary of findings:

- RUMRRM, RUM and RRM do equally well (or badly) in terms of model fit and in- and out-of-sample prediction.
- Simulations can show quite different predictions, but we also do not know/cannot predict which one is best.
    - The exact same data can lead to different recommendations, which is scary.
- RUMRRM seems to provide a balance between RUM and RRM, so as long as we do not know/cannot predict which one is best, it might be your safest bet.
    - Although adding regret is not feasible for all studies.

Future Research

We will be looking into the effect of different design strategies, maximizing statistical robustness for regret parameters, while keeping the design as D-efficient as possible. The designs we used showed to be sub-optimal for estimating regret, as you can see in Figure 14.

Figure 14. Percentage of Regret Observations

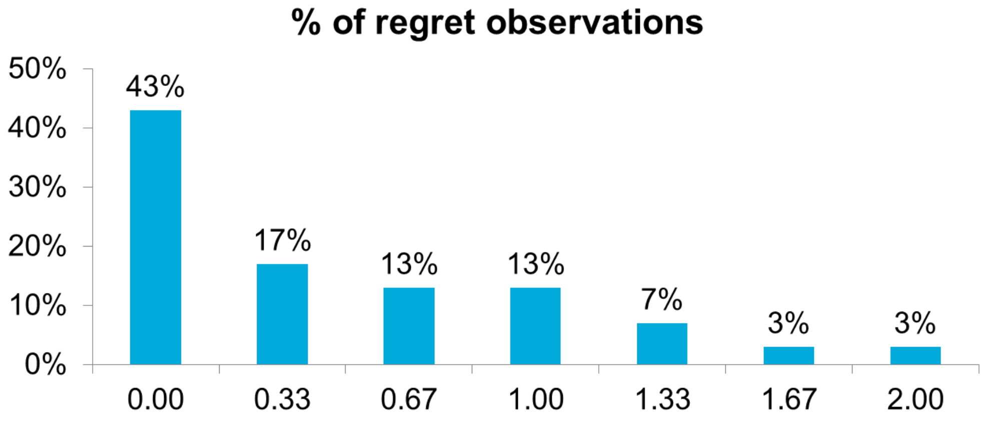


Figure 14 shows that the amount of observations for the higher portions of regret are much lower. This disbalance could lead to skewed estimations.

We will look into a way to apply RUMRRM on more concepts in the simulator than were shown on screen. We will do this by coding average regret, instead of the sum of the regret. This will make sure one is not extrapolating beyond the tested range.

Jeroen Hardon

Kees van der Wagt

References

The Random Regret Minimization Choice Modeling Paradigm: An Introduction with Empirical Tests (2014). Keith Chrzan and Jefferson Forkner; Sawtooth Software, Inc. (http://www.sawtoothsoftware.com/support/technical-papers/169-support/technical-papers/cbc-related-papers/1439-the-random-regret-minimization-choice-modeling-paradigm-an-introduction-with-empirical-tests-2014)

Sawtooth Software, Inc. (1999), CBC User Manual, Sequim: Sawtooth Software.

Sawtooth Software, Inc. (1999), The CBC/HB Module, Sequim: Sawtooth Software

Chorus, Caspar G. (2010) “A New Model of Random Regret Minimization.” European Journal of Transport and Infrastructure Research, 10:181–196.

Chorus, Caspar G. (2012a) Random Regret-based Discrete Choice Modeling: A Tutorial. Springer Briefs in Business (e-book).

Chorus, Caspar G. (2012b) “Random Regret Minimization: An Overview of Model Properties and Empirical Evidence,” Transport Reviews, 32, 75–92.

CAPTURING INDIVIDUAL LEVEL BEHAVIOR IN DCM

PETER KURZ[1]
TNS INFRATEST
STEFAN BINNER[2]
BMS MARKETING RESEARCH + STRATEGY

PROLOGUE: SOMEWHERE IN A CENTRAL TEST LOCATION . . .

At the beginning of this paper we want to take you to a real market research situation. Imagine you are with your client in a "central test location," a research studio where consumers are invited to participate in a survey. As such studios are usually set up for qualitative research purposes, there are observation rooms where clients or researchers can observe group discussions or in-depth interviews. When conjoint studies, a quantitative research method, are not conducted online, such test locations are often used in order to screen for the right target group, to use stimuli (such as dummies) and to conduct interviews in front of a computer. Of course such a set-up represents a great opportunity to observe interviews: either by the researcher, e.g., to conduct pre-tests, or by the client who wants to understand how consumers react to the choice tasks and to gain some insights as to their preferences.

We are now in such a central test location with our client for a large and important study. The respondents are instructed by the interviewers to always comment on what they select and to explain the motivation for their choice decisions. We are in the observation room listening to a respondent explaining her preferences while clicking through the choice tasks. As she goes through the choice tasks our client is impressed with how consistently she is making her choices. Our client is especially pleased that she has a high preference for a particular product feature as he is convinced that this feature is quite important for many of his customers.

A few weeks later we present the results of the study to the client and his team. To their general surprise (our client having already reported his experience during the interview to his colleagues), the product feature in question came out as being not desirable at all. Another feature was clearly the winner. Our client is irritated and tries to understand why the results do not fit the observations he made during the interviews. He asks about the interview with the respondent he found so interesting and wants to see her individual results. As we are prepared for all types of discussions, we have the individual results of all respondents on hand and we look for the individual results of this specific consumer. To our surprise, her individual part-worth utility for the feature she liked so much is negative, while the utility value of the alternative feature is positive. Our client raises his eyebrows and asks for an explanation. Didn't we promise him that we can derive individual utility values instead of an aggregated result? Didn't we tell him that by using HB we are able to deliver best results even if the model is quite complex and we cannot ask each respondent to complete very many choice tasks? Did something go wrong during the estimation process? How can we explain this result to our client without losing his trust in discrete choice modeling, and us?

[1] Head of Research & Development TNS Infratest (Peter.Kurz@tns-global.com)
[2] Managing Director, bms marketing research + strategy (s.binner@bms-net.de)

MOTIVATION FOR THIS PAPER

In the last few years many papers presented at market research conferences in regard to conjoint analysis or DCM focused on such topics as how (if at all) to apply covariates, how many choice tasks can be asked or how parameters (e.g., priors) should be set in the Hierarchical Bayes estimation and how relevant these parameters are for researchers.

Although these papers were sometimes controversial (especially since academics and practitioners often came up with different conclusions) some of the common conclusions from these discussions are:

1. In complex DCM designs one usually does not derive pure individual utilities, but kind of artificial or "pseudo individual" utilities which more-or-less represent the sample.
2. As long as one uses the resulting part-worth utility values for market simulations and not for segmentation or other additional analysis procedures it is believed that "enough" heterogeneity is captured and the simulation models work fine.
3. Even if there is not "enough" heterogeneity captured, it is certainly more than with aggregate logit models or other "practical" alternative approaches such as latent class analysis.

On the other hand when we are forced to use these "pseudo individual" utilities for segmentation or when we dig deeper into the data structure we sometimes find these individual estimates as being not really intuitive: We might not find the segments clients expect or we have observed while attending real interviews (as in our prologue). In the worst case, these "pseudo-individual" results, sold to our clients as real individual results, can lead to distrust in the simulation results and the value of the whole study.

Therefore we need a deeper understanding of how much heterogeneity we really capture with our DCMs using hierarchical Bayes techniques. This understanding will help us to further improve our research designs (e.g., sample structure) and estimation processes and guide us in how to interpret the resulting estimates and results.

QUESTION 1: CAN WE CAPTURE MORE HETEROGENEITY BY APPLYING COVARIATES?

Let us start with a short introduction to covariates and how they can be applied. If one uses "standard HB" as it is implemented in Sawtooth Software packages, default settings are applied and there are no covariates defined and one single multivariate normal prior is assumed. This leads to a "shrinkage" of respondents preferences towards the population mean and this effect is sometimes quite significant. This shrinkage tends to reduce differences between individual respondents and thus, between customer segments. The effect becomes more noticeable for small customer segments, unless they are boosted by using disproportionate sample structures. When such segment differences are reduced due to shrinkage, their chance of being identified and acted upon goes down as significance testing of segment level differences is based on the shrunken, less-different values.

The purpose of the application of covariates is to allow segments to keep their characteristics within a total sample estimation model. Instead of estimating disaggregate part-worth utilities with hierarchical Bayes based on one sample mean, which might reduce heterogeneity greatly compared to the underlying reality (Figure 1), the use of covariates in the upper level model aims to increase the heterogeneity of the utility distribution by shrinking respondents of different subgroups to their means based on their own subgroups rather than the total mean (Figure 2, for a single 2-level covariate).

Figure 1: Single Sample Mean

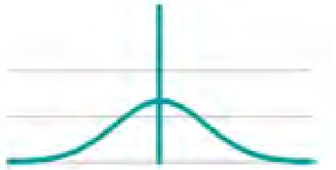
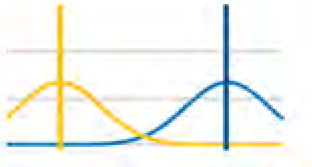


Figure 2: Multiple Sample Means



In theory this should do a lot to solve the potential problem of shrinkage of single respondents to the sample mean. However, in everyday work covariates have often not been found particularly effective in improving the overall model performance or in enhancing differentiation between subgroups in the simulations. If there is a sufficient number of choice tasks, covariates do not improve the model performance, because the lower level model dominates the solution. Even with a small number of choice tasks in most of our studies, covariates in general did not improve results; in some of our cases the covariates "washed out" and estimation converged to the same parameters as when no covariates were used.

We observed that in studies where each segment is represented by a sufficiently large sample size, HB without covariates converges towards same parameters as HB with a covariate model. But if the covariates are not really able to predict differences between sample segments, in the worst case they are just extra noise added to the model and can actually make it worse. Nevertheless the application of (the right) covariates in HB estimation will sometimes result in better distributed utilities and in an improvement of aggregate metrics, both at the total sample and subgroup level. The problem is just that we do not have or know the right covariates in all our projects.

Conclusion 1: Due to the lack of discriminating covariates, they are often not the solution to the issue of excessive shrinkage

QUESTION 2: WHY DON'T WE SIMPLY INCREASE THE NUMBER OF CHOICE TASKS IN ORDER TO COLLECT MORE HETEROGENEITY?

In our paper presented at the 2012 Sawtooth Software Conference we demonstrated there is a natural limit to how many choice tasks an individual respondent can answer. We called these limits Individual Choice Task Thresholds or simply ICTs. An ICT is the threshold past which an individual's further choices lead to poorer model fit rather than better, due to over-simplified responses or symptoms of exhaustion. In most of our studies we could see that in general, respondents had a diverse answering behavior and individually different choice task thresholds. For many respondents we got better or equally good hit rates and share predictions when we used only a smaller number of their choice tasks (the first ones, not the later ones) in order to avoid simplification.

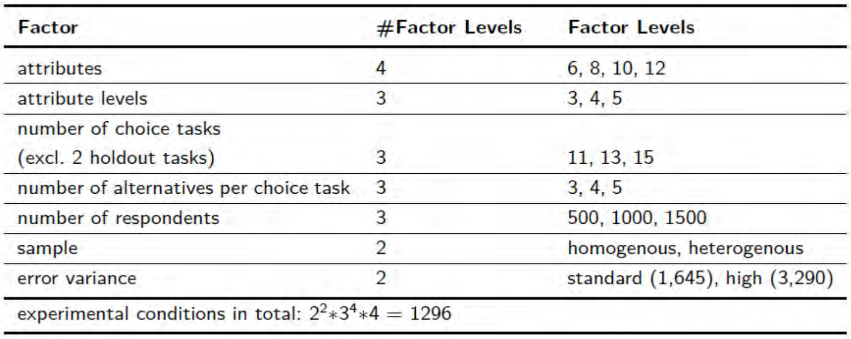
Therefore we concluded that "Less is more," meaning that we should ask fewer choice tasks in order to improve results. Furthermore we learned that more choice tasks could even be dangerous, resulting into misleading results and interpretation. The analysis of the individual posterior distributions showed that a large number of respondents tend to simplify the answering in later choice tasks. In the first half of the choice tasks we saw higher number of attributes with significantly non-zero utilities than in later ones.

Conclusion 2: due to the ICT, we cannot solve the shrinkage issue by simply increasing the number of choice tasks.

TOPIC 3: CAN SAMPLE SIZE COMPENSATE FOR THE LIMITED INDIVIDUAL INFORMATION WE COLLECT?

In order to answer this question Hein, Kurz & Steiner set up a research simulation experiment with 1,296 models:

Figure 3: Experimental Research with 1,296 Models (Hein et.al., 2013)



Factor	#Factor Levels	Factor Levels
attributes	4	6, 8, 10, 12
attribute levels	3	3, 4, 5
number of choice tasks (excl. 2 holdout tasks)	3	11, 13, 15
number of alternatives per choice task	3	3, 4, 5
number of respondents	3	500, 1000, 1500
sample	2	homogenous, heterogenous
error variance	2	standard (1,645), high (3,290)
experimental conditions in total: $2^2*3^4*4 = 1296$		

This simulation study clearly showed that for 6 and 8 attributes an increase of sample size could compensate for a decrease of the number of tasks (T) from 15 to 11 in terms of average RLHs. However, for a larger number of attributes (10 or 12, with 5 levels each), even a tripling of sample size from 500 to 1500 could not compensate for a relatively modest decrease of T from 15 to 13, in terms of average RLHs. We use T as the number of repeated measurements (number of tasks per respondent) in a choice model in our following explanations.

Furthermore the findings showed that a lack of individual information could not be compensated for by larger samples. HB does the best it can estimating part-worth utilities with a maximal amount of heterogeneity—but has no chance to provide good individual parameter accuracy if T is small. With small T one should consider using the upper level model for simulation.

Conclusion 3: Increased sample size is not a solution to cope with the limited number of choice tasks possible.

ISSUE 4: PARAMETER SETTINGS IN THE HB ESTIMATION

Each researcher has to define a hierarchical prior distribution of heterogeneity before estimating an HB model for discrete choice experiments. This distribution is usually intuitively chosen by the analyst. In practice most researchers currently use the multivariate normal distribution as the standard choice for their prior.

In a discrete choice data set, the observed choices y_{jt} are assumed to follow a multinomial logit distribution:

$$y_{jt} \sim \text{MNL}(X_{jt}, \beta_j)\ ;\ j = 1, \ldots, N;\ t = 1, \ldots, T;$$

where N is the number of respondents in the sample and T the number of choice sets each. The vector of part-worths β_j is different across respondents according to:

$$\beta_j = \Gamma z_j + u_j$$

where Γ is a matrix of coefficients relating the vector of part-worths β_j to a respondent's specific demographic variables z_j (i.e., the covariates). The product Γz_j accounts for the observed heterogeneity attributable to covariates, while u_j is a stochastic component representing the unobserved heterogeneity component. The distribution of u_j is of particular interest because it influences how β_j can vary across respondents independently of the covariates. In current practice, a standard multivariate normal distribution is almost always used as the default setup:

$$u_j \sim N(\mu, \sum)$$

is the "mid-level prior" governing how respondents differ, and its parameters come from the top-level prior that has "hyper-parameters" $\bar{\mu}$, υ and V set by the analyst (or by default, by the software):

$$\mu \sim N(\bar{\mu}, \sum \otimes a_\mu^{-1})$$
$$\sum \sim IW(\upsilon, V)$$

(The variance of the top-level prior is distributed as an Inverted Wishart variable [a multivariate generalization of the inverted chi-squared], scaled up by some identity-structured matrix. If the hyper-variance V is set very high, so the prior is very diffuse, the top-level model represented by the last two equations gives "free rein" to the data's determination of the parameters of the MVN for u_j in the middle-level model.)

If you believe in the above model of heterogeneity, the estimates result in consistent and efficient inferences about the unobserved population from the hierarchy, even for a small number of repeated measurements (T). But should you really believe in a MVN distribution of heterogeneity? For large T, the collection of in-sample posterior means are usually robust against misspecification of the upper level model. In other words, when T is large, lots of data overwhelms the priors, misspecified or not. However, in practical discrete choice experiments, T is almost always very small, so these assumptions matter more.

Consider a quick reminder of how hierarchical Bayesian methods work. Generally, three different levels are distinguished. The first level shows the hierarchical prior, with its parameters $\bar{\mu}$, υ and V, which are just chosen by the analyst (often just by accepting software defaults). The second level contains random effects or individual level coefficients β that represent consumer preferences for a given attribute. Finally, the third level is the data y. The index j represents the j^{th} respondent of a discrete choice data set and N denotes the total respondents participating in the survey. The index t stands for the t^{th} choice task and each respondent answers in total T choice tasks (Rossi, Allenby, McCulloch 2005).

According to the model, the β_j are determined by the hierarchical prior and then generate the data y (i.e., choices of the respondents). It implies that β_j for all unobserved consumers in the same market should be generated from the upper level model, assuming the upper level model is an acceptable representation of the population.

Using posterior means for this generalization can be risky from a theoretical perspective. There will be posterior uncertainty in the β_js for finite T that using posterior means completely ignores. On the other hand, the upper level model still provides the same insights about the population if N is large enough, even when T is small. This is a major theoretical reason to use a hierarchical model in the first place. A set of posterior means will only represent the population reasonably if both N and T become quite large. Large N overcomes any misspecification of the shape of the mid-level prior (i.e., MVN) and any problems in the hyper parameter values. Large T brings the posterior means into line with the data and reduces the influence of shrinkage. Large T also reduces the variance of respondents' posteriors, so that ignoring that uncertainty for each respondent is less problematic.

But in practice, most current choice simulators use posterior mean estimates of in-sample respondents to generalize beyond the sample of respondents interviewed and to predict preference shares for the population. This means that lower level model preferences are being used in the following form to generalize in simulations.

1. Use posterior draws $\{\beta^1{}_j, \ldots, \beta^R\}$ for j € $\{1, \ldots, N\}$ calibration individuals with R draws saved for each
2. Calculate posterior means $\hat{\beta}_j = 1/R \sum^R{}_{r=1} \beta^r{}_j$ for each respondent
3. Calculate preference shares for alternative i in a simulation scenario defined by attributes X_j:

$$S_{ij} = 1/N \sum^N{}_{k=1} p(y_i|X_j, \hat{\beta}_k)$$

In other words, posterior means of in-sample respondents are used to generalize beyond the sample and to predict choices of consumers in the hypothetical market.

So far we introduced in our theoretical discussion how to generalize from an HB model using the lower level model by creating a distribution of part-worths that is—perfectly—a 1:1 representation of the degree of heterogeneity in the population. In the following, we will use a simulation study to show what happens with the answers of our respondent from the introductory example during the estimation process. This explains how the lower level model inferences will be influenced in terms of replicating the true amount of heterogeneity in the population. Therefore, we compare a distribution of in-sample posterior means to the distribution obtained from the posterior of the hierarchical prior. This comparison will be simulated for different numbers of choice sets per respondent T and sample sizes N. As we will see, the number of choice tasks T in a discrete choice model data set has a strong impact on the results.

Our first simulation shows how inferences based on posterior means will be influenced when the number of repeated measurements T is small. For this simulation, we use a simple multinomial logit model setup (not hierarchical) that only includes one respondent (our woman with "rear wheel-drive" preference) without any heterogeneity.

Simulation Setup:

- MNL model (no hierarchy)
- Let β = (2,2) represent data generating preferences (the “true” utilities)
- Use β0 = (0,0) and Σ0 = $\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$ as informative priors
- p = 3 alternatives per choice task (as in our car example)
- T1 = 3 < T2 = 20 < T3 = 1000

Note that in these assumptions, the β_0 prior is not very close to the true β, but the prior variance Σ_0 is relatively tight. In other words, β is out in the tail of the MVN prior.

Figure 4

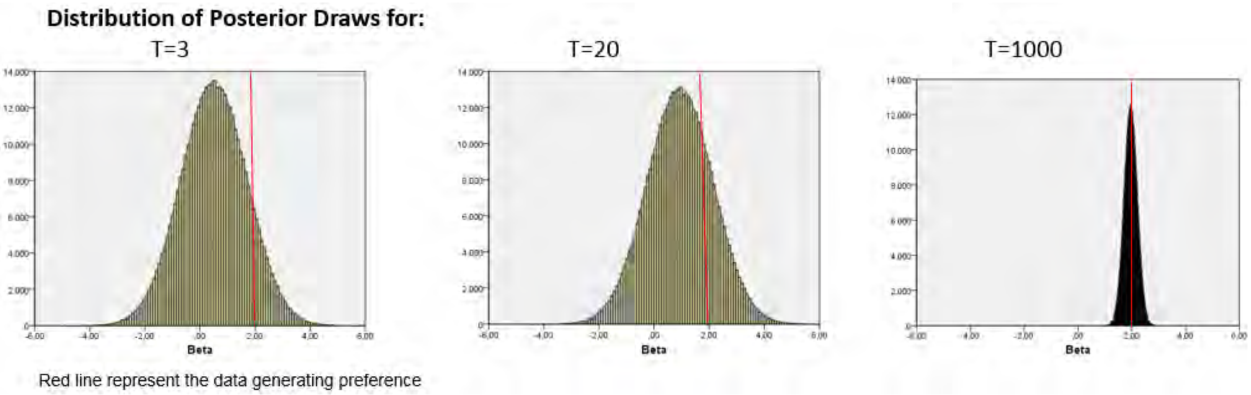


Figure 4 shows three graphs that correspond to the three different numbers of choice tasks T = 3, T = 20 and T = 1000 used in the simulation. The preferences, or betas, are plotted on the x-axis and their density on the y-axis (these graphs would be the same for either the first or second element of β). The vertical red line marks the true data generating preference (true β) that is equal to two. The solid black line traces the density function for the distribution of posterior draws. Figure 4 clearly shows that the posterior mean is little different from the prior when the number of repeated measurements T is small (T=3) and that, in this case, the posterior mean would not be very informative about the true location of the respondents preference. However, as the number of repeated measurements increases, the posterior mean becomes more and more accurate as to the true location of this respondent’s preference.

The informative prior we used here is meant to mimic what we get from a population of respondents with relatively little heterogeneity. In our real life example, the large number of front-wheel-drive likers means relatively little heterogeneity. In this situation, posterior means of individual level coefficients will be shrunk toward the prior (the overall distribution of respondents) unless T is really large. This is the reason for the shrinkage from rear-wheel-drive preference to front-wheel-drive preference for our respondent.

In the next simulation we extended the results from Figure 4 and applied them in the context of hierarchical models. The purpose of the following simulation is to confirm the previous finding that posterior mean inferences will be shrunk strongly towards the prior if T is small, meaning that lower level model estimates are unable to discover the true heterogeneity distribution in the population under such conditions. On the other hand, we will show that inferences on the basis of the posterior of the hierarchical prior will be unbiased regardless of the number of repeated measurements T (for a more comprehensive explanation of this topic, see: Pachali, Kurz, Otter 2014).

Let $\beta_i \sim$ MVN $(\bar{\beta}, V_\beta)$ with

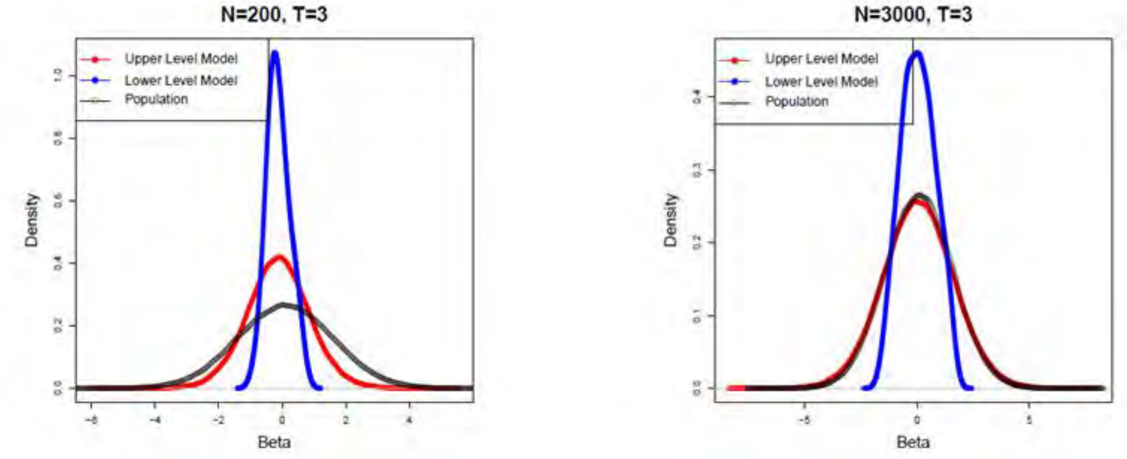
$$\bar{\beta} = (0,0.1,0.2,0.3,0.4) \text{ and } V_\beta = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2.5 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

We use this heterogeneity distribution to generate small (N = 200) and large (N = 3000) samples and then have each sample member participate in a short (T = 3) or a long (T = 60) discrete choice survey.

Consider combinations of small or large T and N:

N = 200, T = 3
N = 200, T = 60
N = 3000, T = 3
N = 3000, T = 60

Figure 5: Posterior Densities for Level "Rear-Wheel-Drive" with $\beta_2 \sim$ N (0.1, 1.5) and T = 3

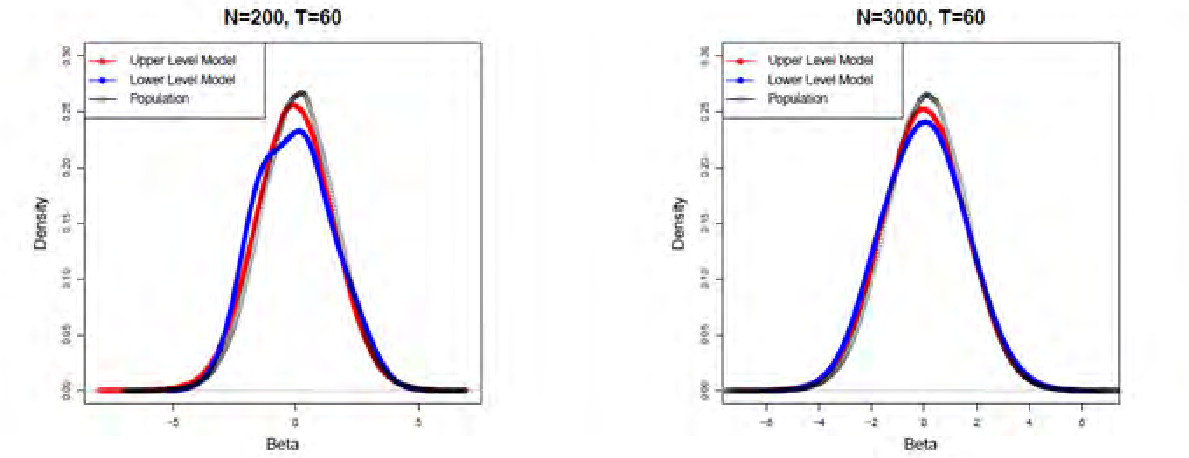


The posterior means severely underestimate the actual heterogeneity if T is small.

Figure 5 shows the distributions of posterior draws for the second part-worth. The figure on the left corresponds to the small sample N = 200 taking a short conjoint survey T = 3. The part-worth utility or beta is plotted on the x-axis while the y-axis shows the density. The black line (the lowest, flattest line) depicts the true distribution of heterogeneity for the part-worth "rear-wheel-drive." The blue line (the tallest, peakiest one) depicts the distribution of heterogeneity inferred from posterior means or the lower level model. Finally, the red line (the one in the middle) depicts the distribution of heterogeneity inferred from the posterior of the hierarchy, or the so-called upper level model. The figure on the right corresponds to the larger sample N = 3000 taking again a short conjoint survey T = 3. The color codes are the same (the black and red lines are nearly on top of each other).

Figure 5 shows how relying on posterior means of individual level coefficients severely underestimates the true amount of heterogeneity in both cases when the number of repeated measurements is small. This means even with a large sample size we are not able to capture enough heterogeneity that the "rear-drive" part-worth of our respondent isn't shrunk too much towards the mean.

Figure 6: Posterior Densities for "Rear-Wheel-Drive" with β_2 ~ N (0.1, 1.5) and T = 60



When the number of repeated measurements T is high, the information in the individual level posteriors is essentially independent from the sample size N.

However, the differences in Figure 5 vanish once long conjoint surveys are considered where each consumer provides a lot of information. So we see that if we obtain enough individual information—i.e., a long interview for our respondent—our model captures enough heterogeneity to represent the correct "rear-drive" part-worth for our respondent, even if she has a preference very different from the population. Figure 6 shows that the form of generalization from the hierarchical model becomes less important if the number of repeated measurements increases and the information in the data get rich (in each half of this figure, the blue lower-level model line is the least peaked, the black population line is most peaked, and the red upper-level model line most peaked).

So far, it has been shown that lower level model inferences about the unobserved population are biased when the number of repeated measurements T is small. This bias is caused by systematically underestimating the true amount of heterogeneity in the population. On the other hand, upper level generalizations consistently estimate the true amount of heterogeneity in the population even if the number of repeated measurements is really small. However, this only holds true under the premise that the hierarchical prior distribution of heterogeneity is correctly specified. This is an important finding since T will almost always be small in practice due to time constraints associated with the questionnaire and due to the response quality problems we encounter if we exceed the ICTs.

Conclusion 4: The current practice of using posterior means to generalize from the HB model biases inference and decision when T is small. The bias is against heterogeneity and differentiation. In practice, T will always be relatively small, because clients are more demanding, models become bigger and bigger, and respondent time and attention is limited

(ICT). Generalizations from the upper level model are consistent, i.e., without bias and efficient, even for small T, so long as the hierarchical prior distribution of heterogeneity is not misspecified. Therefore we should invest more time in the specification of our hierarchical prior in order to derive better market insights.

Can We Capture Individual Level Behavior?

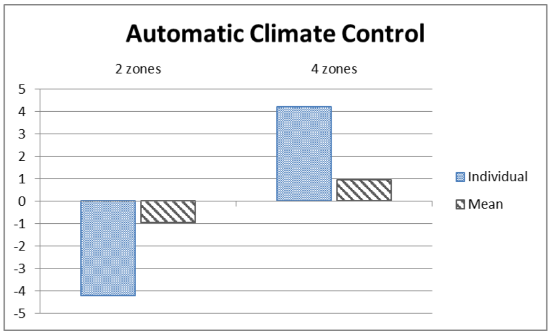
In day-to-day research practice researchers often talk about individual utility values estimated with hierarchical Bayes methods. However, pooling respondents together to get enough information and using multivariate normal distribution assumptions for estimating our models can result in much greater shrinkage than many practical users of HB are aware of.

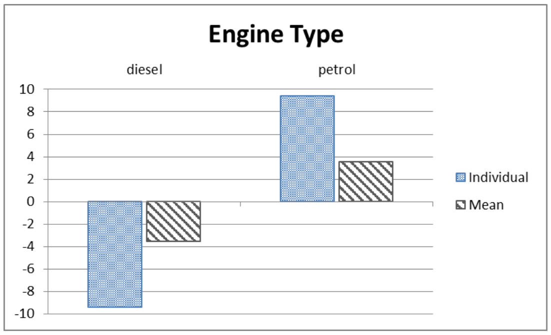
We analyzed a large-scale multinational study where we had a large database of recorded respondent observations in which they explained their preferences while taking the survey and compared these with their individual utilities. We aimed to get answers to the following questions:

Is it possible that individual preferences get washed out due to shrinkage?
Does it always happen?
Why does it happen?
What learnings can researchers take away?

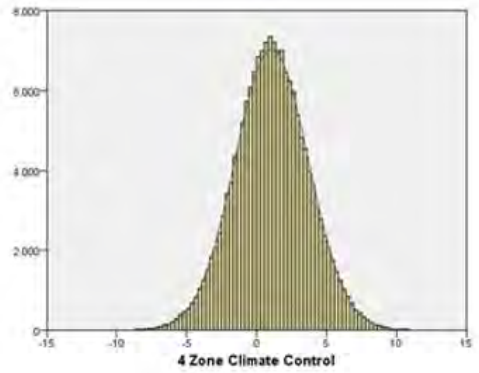
Let’s start with cases where we found individual behavior perfectly captured in the individual utilities.

The following two examples are ones where the individual opinion is consistent with population mean. In these two examples of two level attributes the individual preference shows in the same direction as the sample mean:

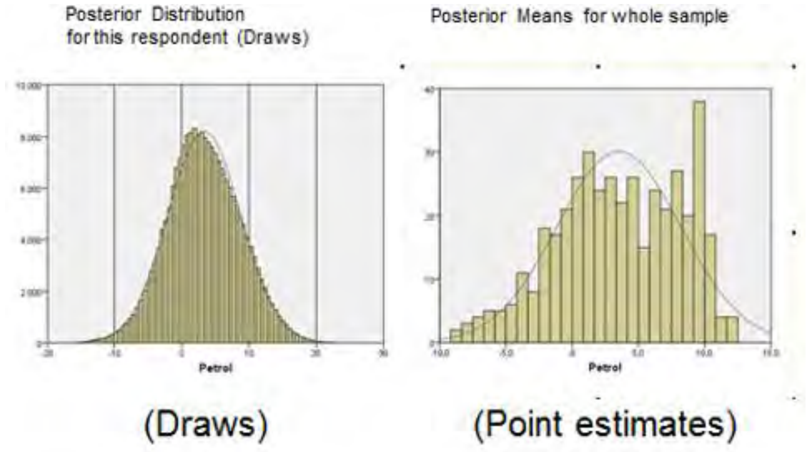




Furthermore the parameter captured a reasonable amount of variation as the following plot of estimation draws across all respondents shows:



Another interesting observation is that the mean over the last 1000 draws (posterior means) of the sample seems not be able to represent the true distribution (plotted line showing the nearly normal distributed real values):

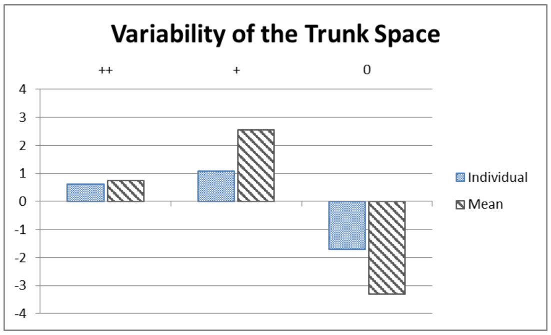


Let's now have a look at some cases where we did not capture individual level behavior:

In the first case the respondent's preference is not consistent with the population mean. This respondent gave a clear explanation of his preferences:

"I would say the first one because of variability of seats is good and more so the variability of the trunk space is better and the comfort of loading and unloading."

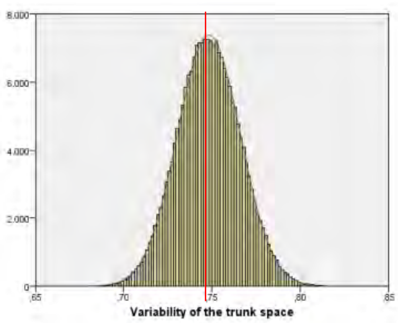
However, the individual utility values of this respondent look completely different:



There seems to be so much shrinkage that the order of preference between the first two levels of this respondent got reversed and there was not enough individual information to prevent that.

The posterior distribution for the sample of this case shows how narrowly distributed all draws are around the population mean:

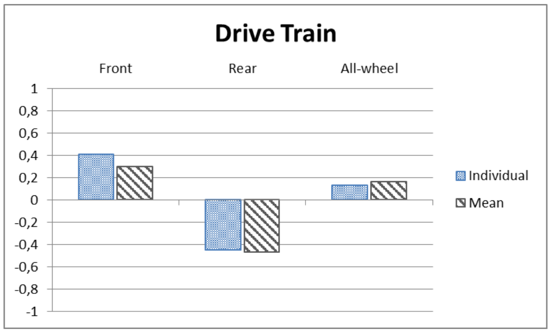
Posterior Distribution for ++ Level



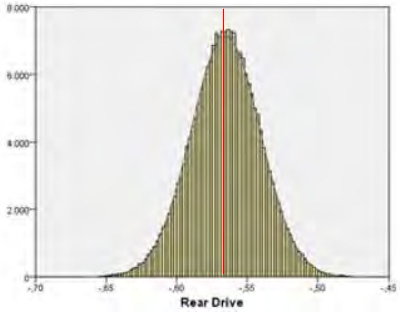
Another respondent among our examples has a clear preference:

"And plus it's rear wheel drive."—"And that's important to you?"—"Yes."

This respondent's preference is also not in line with the population mean. As there is again not enough individual information to capture the difference, a strong shrinkage towards the population mean can be observed:



And again the parameter estimates did not capture much variance:



What Did We Learn from this Analysis?

Point estimates (posterior means) don't always reflect the true variance (spread) in the population. However, we observed that posterior *draws* from the estimation process are usually doing better.

However, sometimes the draws too do not capture individual heterogeneity. This might be caused by several factors:

1. Too sparse individual information due to high model complexity (number of attributes and levels, alternative specific designs etc.) and limited number of choice tasks
2. Lack of covariates or (sometimes even worse) selection of the wrong covariates
3. Too small representation of a specific segment in the sample, caused by small share in the market (such as niche markets or exotic target groups)
4. Unusual individual answering behavior (i.e., outliers)

Bottom line: the failure to capture individuality is most severe for respondents who differ greatly from the population average. This is especially crucial when searching for segments or niches and when analyzing individual cases for any other purpose as well.

What If We Do Need to Search for Segments or Niches?

Before searching for segments or niches through individual results, researchers should apply more diagnostics to their studies and models. The following questions could provide some guidance:

- What does the tail-behavior of the distribution of the posterior draws look like? Are there any visual indications of shrinkage?
- Would we do better to apply mixtures of MVN's instead of using one single MVN?
- Are there effective covariates we could apply?
- Could we have collected more individual information, or reduced the number of parameters in the model?
- Can we simulate from draws or from the upper-level model, considering the distribution and structure of variance-covariance, instead of using point-estimates?

How to Avoid Mistakes in the Future?

From what we learned we advise practitioners to:

- Recognize that DCM—like almost every quantitative method—carries the danger of the mean fallacy error ("stuck in the middle").
- Always examine the individual distribution of your parameter estimates.
- Never forget that DCM models are generally excellent for the total market, but due to possible shrinkage effects they bear danger for small segments or niches
- Consider aggregated models such as Logit or distribution free LC, especially if there are attributes which are polarizing for a minority in the sample.
- Always try to understand the target group (consider pre-research in order to identify sub-segments or possible covariates).
- If one expects sparse data or specific sub segments the best practice depends on whether these segments are known up front:
    - If such segments are known in advance, apply covariates or sample-boost rare cells.
    - If such segments are unknown, apply diagnostics to the model (e.g., counts and parameter variations) and be aware of possible lack of individual level behavior in the model when running analysis and when drawing conclusions from the data.

Peter Kurz

Stefan Binner

REFERENCES

Allenby, G.M.; Rossi, P.E. (2006): Hierarchical Bayes Models, in: Grover, R.; Vriens, M. (Eds.): The Handbook of Marketing Research: Uses, Misuses, and Future Advances, S. 418–440, SAGE Publications Inc., Thousand Oaks.

Hein, M.; Kurz, P.; Steiner, W. (2013): Limits for Parameter Estimation in Choice-Based Conjoint Analysis: A Simulation Study, European Conference on Data Analysis 2013.

Johnson, R.M. (2000): Understanding HB: An Intuitive Approach, Sawtooth Software Research Paper Series.

Kurz, P; Binner, S. (2011): Added Value through Covariates in HB Modeling?, Proceedings of the 2011 Sawtooth Software Conference.

Kurz, P.; Binner, S. (2012): The Individual Choice Task Threshold: Need for Variable Number of Choice Tasks; Proceedings of the 2012 Sawtooth Software Conference.

Liakhovitski, D.; Shmulyian, F. (2011): "Covariates in Discrete Choice Models: Are They Worth the Trouble?" 2011 ART Forum Presentation.

Pachali, M.; Kurz, P.; Otter, T. (2014): How to Generalize from a Hierarchical Model, 2014 ART Forum Presentation

Rossi, P; Allenby, G.; McCulloch, R. (2005): Bayesian Statistics and Marketing, Wiley, Hoboken NJ.

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

Sentis, K. and Li, L. (2001): "One Size Fits All or Custom Tailored: Which HB Fits Better?" Proceedings of the 2001 Sawtooth Software Conference.

Sentis, K.; Geller, V. (2011): The Impact of Covariates on HB Estimates, Proceedings of the 2011 Sawtooth Software Conference.

OCCASION-BASED CONJOINT—AUGMENTING CBC DATA TO IMPROVE MODEL QUALITY

BJÖRN HÖFER
SUSANNE MÜLLER[1]
IPSOS

In many product categories, occasions play an important role in moderating consumer choice. While occasions have been integrated into other kinds of market research methods (e.g., segmentation) they are rarely used in combination with conjoint. Their integration can generate additional insights to support strategic marketing decisions. The challenge is to develop a methodical approach that keeps the costs and the burden for respondents acceptably low and is highly predictive of purchase behavior.

We have developed an Occasion-Based Conjoint (OBC) approach with an efficient way of data collection and augmentation. The data collection in the conjoint section differs from standard volumetric Choice Based Conjoint (CBC) only in the formulation of the choice task questions. The integration of occasions is shifted to the utility estimation and/or preference share calculation.

The comparison of different modeling alternatives regarding validity and practicability favors an approach with occasion-specific utility estimation. The comparison of this approach with standard volumetric CBC modeling, however, shows that—while the estimation of substitution effects can be improved—there is a small but relevant decline in internal validity (holdout prediction). Therefore further research is needed in order to find a likelihood formulation that can overcome this drawback.

MOTIVATION

Relevance of Occasions for Consumer Behavior

Consumer behavior is developing more and more in the way that consumers use multiple products from a product category. They look for products perfectly suited to specific occasions. This has resulted in one consumer using different products from the same category depending on the occasion. For example in the shampoo category, a consumer might use one product for every day, another after sports and a third one for special occasions. Similar examples can be found for many Fast Moving Consumer Goods (FMCG) as well as Over-the-Counter (OTC) drugs.

In these categories the multiplicity of usage occasions can—among other reasons—also be seen as one of the drivers of the purchase of multiple items in one shopping trip. Consumers buy having their usage occasions in mind. They know from experience their needs in specific usage occasions and choose products according to their anticipated usage occasions in the future. Their preferences in the choice situation are consequently not only based on factors present at the point of sale but also on their needs in the anticipated usage occasions. This has led to approaches that model the overall utility influencing choice as a function of partial occasion-specific utilities, e.g., a sum over usage occasions (Kim et al. 2002; Dube 2004).

[1] bjoern.hoefer@ipsos.com, susanne.mueller@ipsos.com

The high relevance of usage occasions in many product categories and their influence on choice behavior suggest that their integration has the potential to improve CBC studies in two major aspects:

1. **Model Quality:** Being one of the drivers of the purchase of multiple items, occasions are especially interesting for applications of volumetric conjoint or other choice experiments with multiple discreteness. As explained more detailed in the following paragraphs, we see the crucial improvement potential especially in the *estimation of the sales potential of new products* as well as in the *estimation of substitution effects*.
2. **Additional Insights:** For new product marketing, it is a great advantage to understand which usage occasions exist and at which of these potential new products would face little or no competition. Especially in markets with a high level of saturation, new not-yet-covered usage occasions can represent white spaces which, once they are addressed by new products, allow an intensification of current category users. The positioning and communication of existing and new products can be improved by focusing on the occasions that are especially relevant for the respective product.

More Realistic Estimation of Sales Potential of New Products

On which occasions products are consumed strongly influences the long term success of a product. The higher the number of occasions for which a product is relevant and the higher the frequency of these occasions the faster the product is consumed. This, in turn, will normally lead to a more frequent repurchase of the product.

In standard CBC, it is assumed that all products are equally relevant for all occasions. With a CBC approach that is modeled at the occasion level—this approach will in the following be called Occasion-Based Conjoint (OBC)—more differentiation regarding the occasions can be considered. The following table illustrates the differences of standard CBC and OBC. It shows the distributions of the purchase volume of one respondent for both approaches.

Table 1: The Effect of Occasion-Specific Preferences on Overall Purchase Volume

			Everyday	After Sports	Special Occasions	Total
Approach	**Total Volume (per year)**		**3200ml**	**600ml**	**200ml**	**4000ml**
CBC: without occasion differentiation	Share of Preference (in %)	Shampoo01	75	75	75	**75**
		Shampoo04	25	25	25	**25**
	Volume (per year)	Shampoo01	2400ml	450ml	150ml	**3000ml**
		Shampoo04	800ml	150ml	50ml	**1000ml**
OBC: with occasion differentiation	Share of Preference (in %)	Shampoo01	75	-	100	**65**
		Shampoo04	25	100	-	**35**
	Volume (per year)	Shampoo01	2400ml	-	200ml	**2600ml**
		Shampoo04	800ml	600ml	-	**1400ml**

CBC does not distinguish between occasions. Hence, the same preference share is assumed for each occasion. OBC takes the relevance for the different occasion into account. This results in differences in the preference shares for occasions for which not all products are relevant and consequently also in the total volumes per product (blue highlighted column). To simplify matters in this example, the same preference share distribution as in CBC is assumed for the occasion for which both products are relevant. In the occasion-based view Shampoo 01 has the highest share of preference for an unimportant occasion (i.e., “Special Occasions”). Unimportant occasions are those which have little influence on the overall purchase volume; this might be because they occur with low frequency or involve only little consumption per occurrence. As Shampoo 01 demonstrates, products which are mostly relevant for unimportant occasions tend to be overestimated with standard CBC modeling. This overestimation is especially risky for new products for which no market data exist. It can in extreme cases lead to wrong business decisions. Consequently, in categories in which preferences vary across occasions, market share forecasts can be done more accurately using OBC.

More Realistic Estimation of Substitution Effects

Methodically related to the estimation of sales potential is the quantification of substitution effects. Conjoint modeling can also benefit in this respect from the integration of occasions. It allows consumers only to switch between products that are relevant for the same occasion. Or in other words, products substitute for each other only in the degree to which they can serve the same occasions. This results in more accurate substitution effects which can be especially beneficial when it comes to assessing cannibalization effects of new product introductions. This is of course especially relevant for any kind of portfolio-optimization.

Additional Insights

OBC allows the occasion-specific simulation of preference shares. These preference shares can be indexed relative to the overall share of each product. The resulting index relevance values allow a quick assessment of the occasion-specific performance of the different products. The following table shows an example of index relevance values. The last line of the table contains the relative importance of the occasions. It sums up to 100% and tells us how the total volume is distributed over the occasions. This output can help to improve positioning and communication decisions. From this table you can, for example, conclude that shampoo 04 and 08 are highly relevant for special occasions. This insight can be used to emphasize this usage occasion in marketing campaigns to sell shampoo 04 and 08.

Table 2: Index Relevance Values demonstrate Occasion-Specific Product Performance

Product	Total Market Share Index	Everyday	After Sports	Special Occasions
Shampoo01	100	129	31	51
Shampoo02	100	139	25	4
Shampoo03	100	119	81	28
Shampoo04	100	38	105	443
Shampoo05	100	89	145	90
Shampoo06	100	132	48	9
Shampoo07	100	105	96	75
Shampoo08	100	44	48	504
Shampoo09	100	116	85	35
Shampoo10	100	130	9	81
Relative Importance		**68%**	**20%**	**12%**

Colors:	< 80	80–120	> 120

As demonstrated, OBC can improve model quality and allows the generation of additional insights. However, the integration of occasions into CBC studies without overburdening the respondents is a difficult task.

Method of Integrating Occasions into CBC Analysis

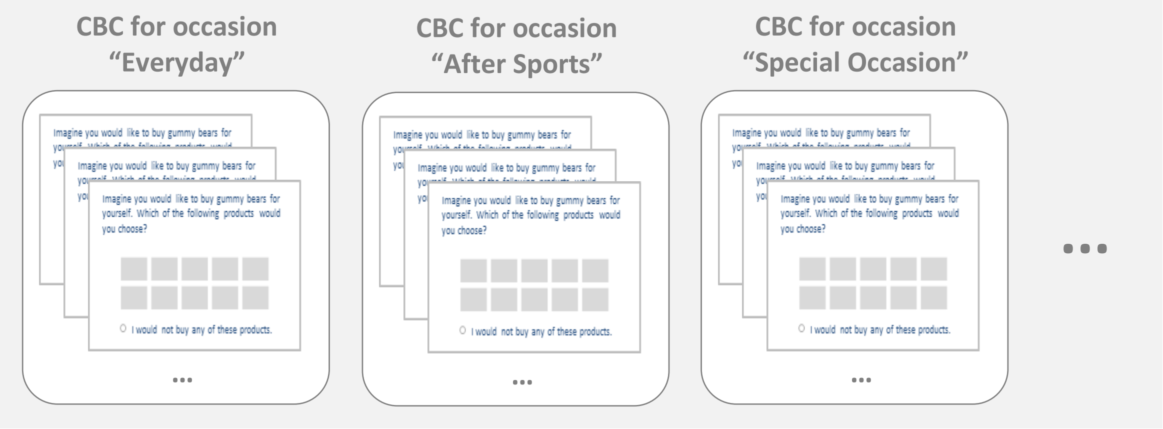
Data Collection—Measuring Occasion-Specific Preferences

The objective of our OBC approach is to measure the suitability of each product for the considered occasions. That means we want to learn how far each product satisfies the requirements of each occasion and quantify this by calculating occasion-specific preference shares.

The estimation of the preference shares is based on data from a volumetric CBC exercise. The necessary information can be collected in several ways. In the following, five possibilities are explained. They are listed in descending order according to the burden for the respondents (i.e., decreasing number of choice tasks).

1. **Occasion-specific choice tasks for all relevant occasions:** The most obvious way to receive the data base for occasion-based preference shares is to prepare one CBC exercise for each relevant occasion per respondent. The CBC exercises just differ in the question text which always refers to one specific occasion.

Figure 1: CBC Exercises for All Relevant Occasions

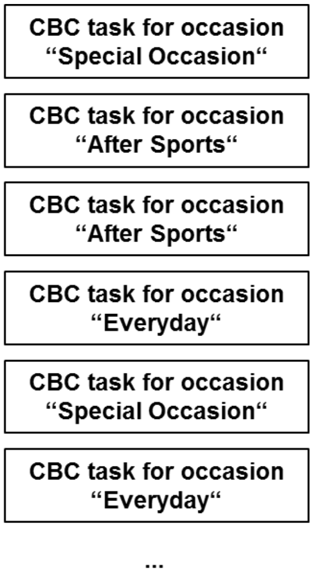

CBC for occasion "Everyday"
CBC for occasion "After Sports"
CBC for occasion "Special Occasion"
Imagine you would like to buy gummy bears for yourself. Which of the following products would you choose?
○ I would not buy any of these products.
...

As enough individual level data for each occasion exists, the occasion-specific preference shares would result from separate HB estimations for each occasion.

While this is methodically simple, it normally results in an unacceptably high number of choice tasks per respondent. For example 12 tasks per CBC exercise for 3 out of 6 occasions that are relevant for the respondent would already result in 36 tasks. This is unfeasible from a practical point of view regarding the questionnaire length and requirements for the respondents if they are finishing the tasks at all. That means there is sufficient data for generating individual level results on a reliable data base, but the collected information itself might be of low quality, because it is biased by habituation effects.

2. **Reduced number of occasion-specific choice tasks for all relevant occasions**: One way to reduce the number of screens per respondent compared to alternative (1) is to ask a reduced number of tasks for each relevant occasion. For example instead of asking 12 screens for each occasion just 6 tasks could be shown. With 3 relevant occasions, still 18 screens will remain.

Figure 2: Choice Task Flow in Random Order per Respondent

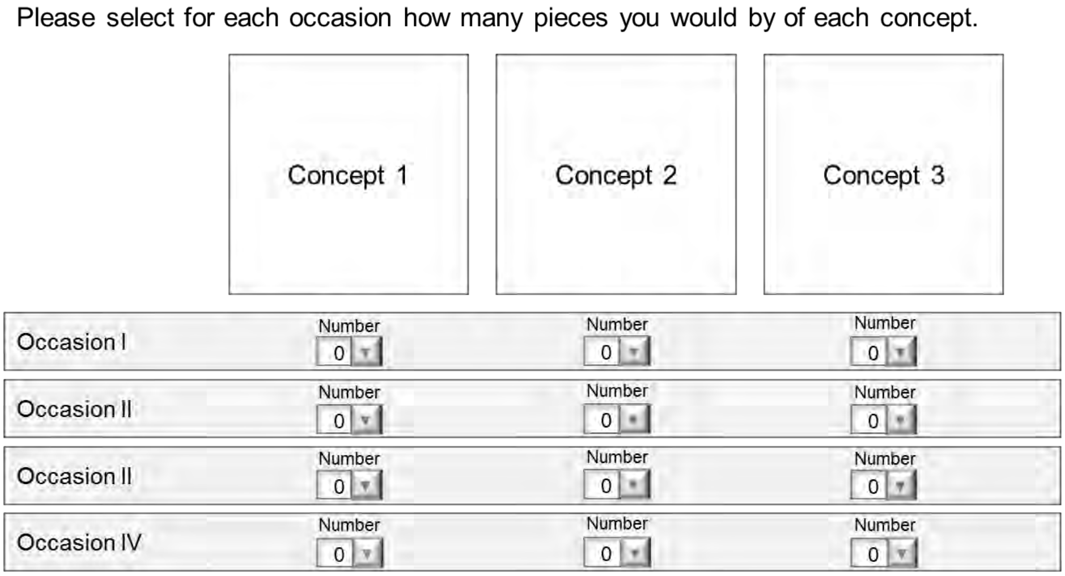

CBC task for occasion “Special Occasion“
CBC task for occasion “After Sports“
CBC task for occasion “After Sports“
CBC task for occasion “Everyday“
CBC task for occasion “Special Occasion“
CBC task for occasion “Everyday“
...

The CBC tasks for the relevant occasions might be asked in random order. Then you need to find a smart way of calling the respondent’s attention to the question text where the occasion of course needs to be changed and highlighted somehow. The question is how to make sure the respondent noticed the change in the considered occasion. The risk of biasing the collected data too much by confusing the respondents in terms of changing settings for the purchase act is not negligible.

When the number of screens per occasion is reduced and all other settings are kept the same, the data base for individual level estimation is automatically reduced. Of course the number of concepts per task might be increased. In our current example it would mean to double the number of concepts per screen to show the same number of concepts as in the not reduced scenario and to collect a similar amount of data. This might lead to an overfilled screen and overburdening the respondent with too complex choice tasks. In conclusion, an occasion-specific utility estimation at the individual level is hard to realize with a reduced number of screens.

3. **Occasion-specific choice for all occasions in each choice task:** An OBC approach showing the same number of screens as if just one CBC exercise without any occasion reference would be done (in our example 12 screens) is the most extreme one. To still integrate the occasion reference into the CBC exercise, all choices can be done on the same screen. This approach can be realized by arranging the screen in the following way:

Figure 3: Choices for all Occasions on the same Screen



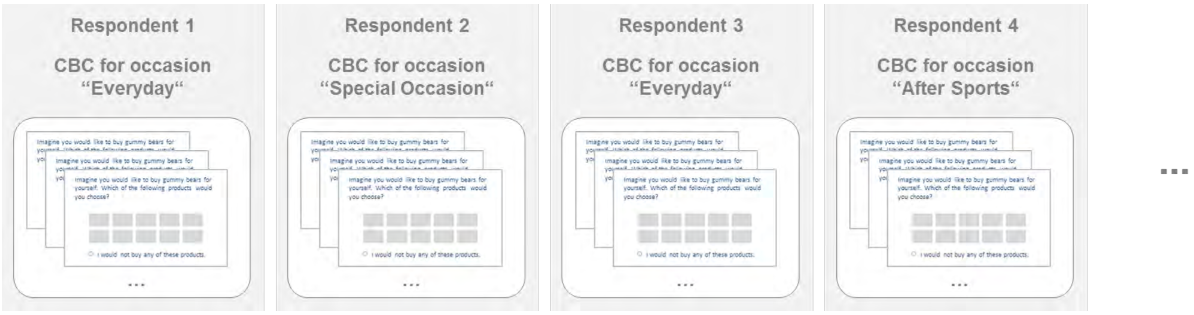
Please select for each occasion how many pieces you would by of each concept.

	Concept 1	Concept 2	Concept 3
Occasion I	Number 0	Number 0	Number 0
Occasion II	Number 0	Number 0	Number 0
Occasion II	Number 0	Number 0	Number 0
Occasion IV	Number 0	Number 0	Number 0

This is acceptable if you show just a small number of concepts per tasks (maximum of 4) and also a small number of relevant occasions for the category exist. Nevertheless you need to keep an eye on the complexity of the choices and assure that it is a realistically solvable task to distinguish between all of the occasions for each concept shown on the screen. In FMCG categories where usually big shelves are shown on a screen this approach would not be applicable.

4. **Occasion-specific choice for just one occasion per respondent:** Another alternative to keep the number of tasks as if an ordinary CBC exercise would be done is to refer in all choice tasks to just one specific occasion per respondent.

Figure 4: CBC Exercise Referring to One Occasion per Respondent


Respondent 1
CBC for occasion "Everyday"
Respondent 2
CBC for occasion "Special Occasion"
Respondent 3
CBC for occasion "Everyday"
Respondent 4
CBC for occasion "After Sports"
Imagine you would like to buy gummy bears for yourself. Which of the following products would you choose?
I would not buy any of these products.
...

The individual level information then of course is sufficient for reliable conjoint results, but the information per occasion needs to be enriched by increasing the sample size. With an adequate number of respondents per occasion, the HB estimation can be run separately per occasion as in alternative (1).

5. **Referring to all occasions in each choice task:** Another approach is to simplify alternative (3) by asking only one volumetric choice for all occasions together and not per occasion. This simplification of the tasks is especially beneficial in the case of having a higher number of concepts per screen for which it would be too tedious to

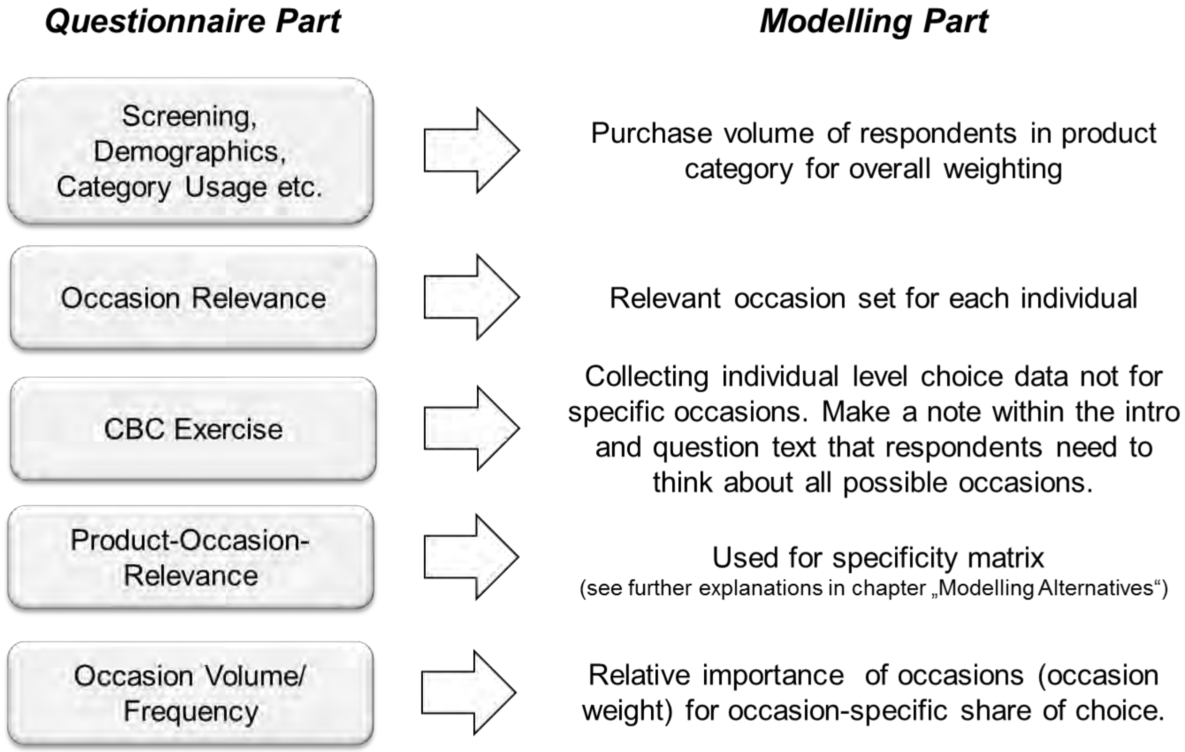
state choices for all occasions. The idea is to collect the general (not occasion-specific) preferences by referring to "all relevant occasions" within the question text and otherwise use normal volumetric CBC tasks. This data can then be augmented with occasion-related questions from the main questionnaire. The complementary occasion information helps to decompose the CBC data collected on a not–occasion-specific level into occasion-specific tasks per respondent.

For our methodological comparison we picked alternative (5), because it is the best approach to avoid overburdening or discouraging the respondents and that can be applied to FMCG categories with many SKUs/products as well. Furthermore, compared to alternative (4) it is more cost saving in terms of sample size for the client. As the data collection with this OBC approach does not differ from a standard volumetric CBC, let's have a deeper look in the collection of the complementary information that is necessary to receive occasion-specific preference shares:

1. **Occasion volume:** To weight each occasion in terms of how relevant it is per respondent, the volume consumed per occasion is needed. It can be determined by asking for the occasion frequency (number of consumption acts in a certain period of time, e.g., 1 year) and the amount consumed per occasion. By multiplying both components, the occasion volume results.
2. **Product-Occasion-Relevance**: To estimate the occasion-specific preference shares, we need to ask at which occasions the respondents use or could imagine using the different products from their relevant set. The most efficient way of doing this is a pick-any question.

This occasion-related information needs to be integrated into the questionnaire flow. After learning something about the category usage and the occasion relevance the CBC exercise follows.

Figure 5: Overview of Questionnaire Flow and Related Modeling Parts



Questionnaire Part

Modelling Part

Screening, Demographics, Category Usage etc.

Purchase volume of respondents in product category for overall weighting

Occasion Relevance

Relevant occasion set for each individual

CBC Exercise

Collecting individual level choice data not for specific occasions. Make a note within the intro and question text that respondents need to think about all possible occasions.

Product-Occasion-Relevance

Used for specificity matrix
(see further explanations in chapter „Modelling Alternatives")

Occasion Volume/ Frequency

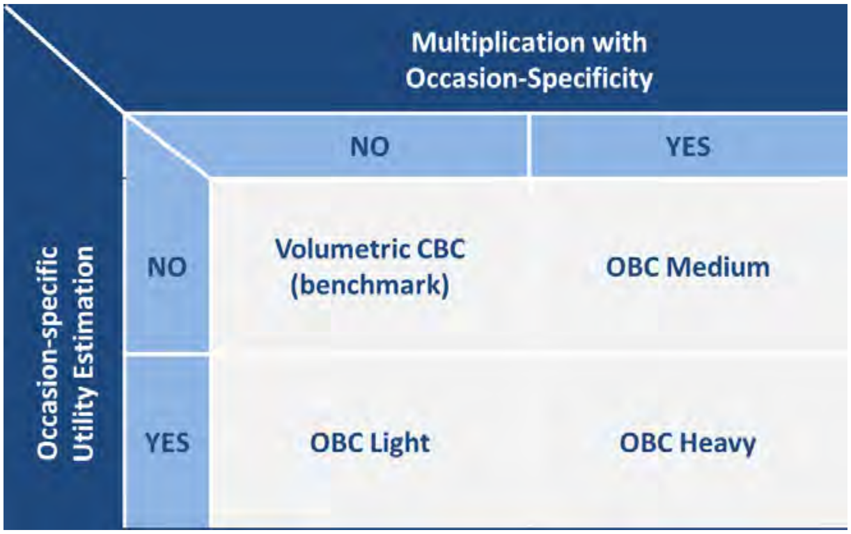
Relative importance of occasions (occasion weight) for occasion-specific share of choice.

The product-occasion-relevance might be challenging for the respondents in the way that you have a big number of products/SKUs that need to be allocated in terms of relevance for each occasion. A smart way to reduce this list of products is to extract those chosen at least once within the CBC and to ask the relevance per occasion only for these. So it makes sense to integrate the more specific occasion related questions after the CBC exercise.

Modeling Alternatives

We tested integrating the occasions at two different points of the modeling process: (1) by running an occasion-specific utility estimation and (2) multiplying with the occasion-specificity (product-occasion-relevance). As Figure 6 shows, this yields 4 modeling alternatives.

Figure 6: Overview of Modeling Alternatives



		Multiplication with Occasion-Specificity	
		NO	YES
Occasion-specific Utility Estimation	NO	Volumetric CBC (benchmark)	OBC Medium
	YES	OBC Light	OBC Heavy

If occasions are neither integrated by (1) nor (2) we have a standard (volumetric) CBC approach. If we use at least the occasion-specific utility estimation, we are talking about a "light" integration of occasions into the modeling of preference shares, hence we call the approach "OBC Light."

The second dimension to augment CBC data for receiving occasion-specific preference shares is the product-occasion-relevance. Asking it efficiently with a BOMPAT question it is defined the following way:

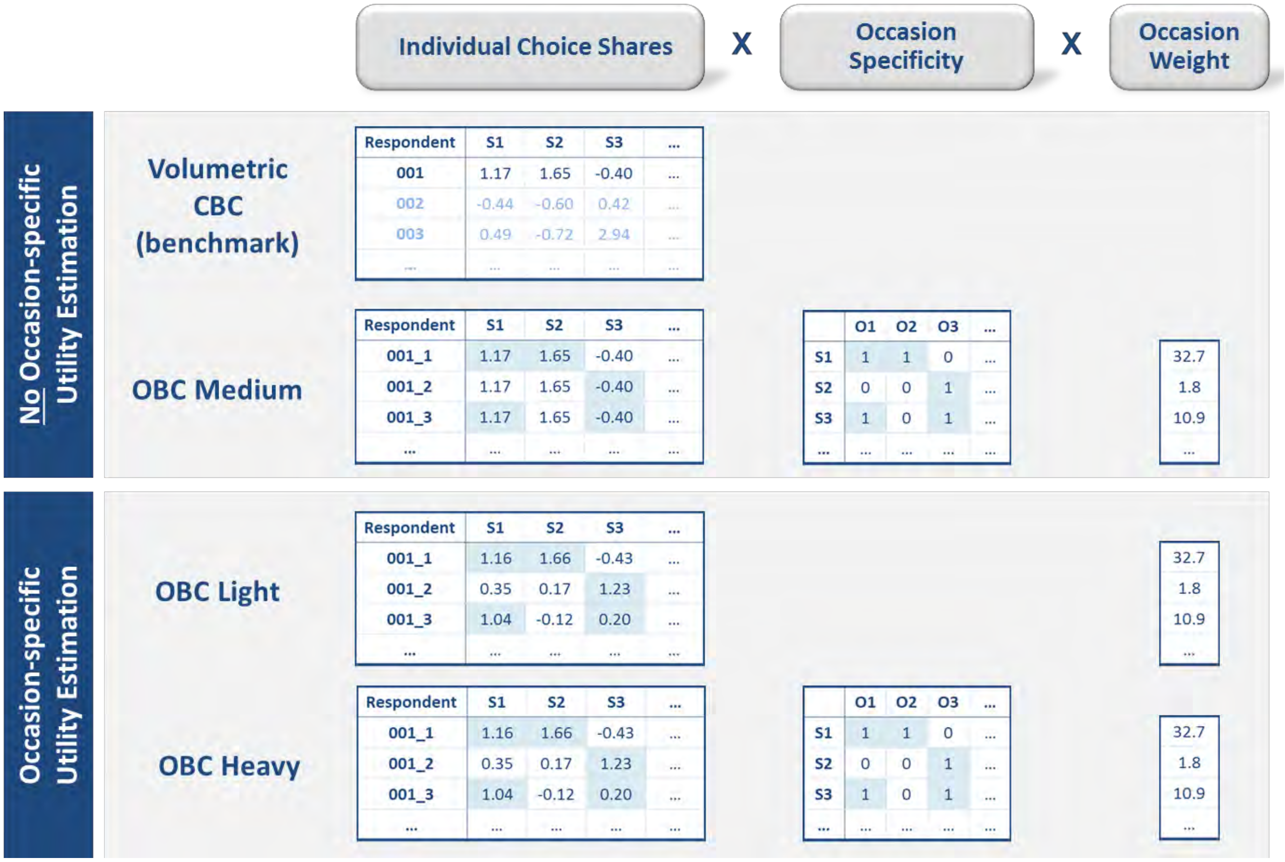
$$S = (s_{ij}) \qquad \text{with } s_{ij} = \begin{cases} 1 & \text{if SKU}_\text{j} \text{ fits to occasion}_\text{i} \\ 0 & \text{else} \end{cases}$$

We call this matrix a "specificity matrix" because it reflects at which specific occasions each product/SKU is used and can hence be assumed to satisfy the needs at that occasion. The simplest way to state this is to set it to 1 if it is relevant for the occasion and to 0 if it is not. Of course, also a continuous coding can be used if the necessary metric information for doing so exists. In the following, we assume a 0/1-matrix as defined above.

The OBC Medium approach includes just the integration of the specificity matrix into the model. Since no occasion-specific utility estimation exists the non-occasion-specific utility values will simply be replicated for receiving a data set on occasion level. Multiplying this data set with the specificity matrix is a relatively rough way of integrating the occasions into the model, so it is called OBC Medium. This approach has one drawback: As you are multiplying the utility values with zeros the model is less flexible in terms of calibration (e.g., according to market information).

Combining OBC Light and OBC Medium results in the OBC Heavy model. See the following Figure 7 to understand what the different levels of occasion integration mean for the main formula of OBC.

Figure 7: Composition of the 4 Modeling Alternatives



Individual Choice Shares
X
Occasion Specificity
X
Occasion Weight
No Occasion-specific Utility Estimation
Occasion-specific Utility Estimation
Volumetric CBC (benchmark)
OBC Medium
OBC Light
OBC Heavy

Volumetric CBC (benchmark)

Respondent	S1	S2	S3	...
001	1.17	1.65	-0.40	...
002	-0.44	-0.60	0.42	...
003	0.49	-0.72	2.94	...
...	...	...	...	...

OBC Medium

Respondent	S1	S2	S3	...
001_1	1.17	1.65	-0.40	...
001_2	1.17	1.65	-0.40	...
001_3	1.17	1.65	-0.40	...
...	...	...	...	...

	O1	O2	O3	...
S1	1	1	0	...
S2	0	0	1	...
S3	1	0	1	...
...	...	...	...	...

32.7
1.8
10.9
...

OBC Light

Respondent	S1	S2	S3	...
001_1	1.16	1.66	-0.43	...
001_2	0.35	0.17	1.23	...
001_3	1.04	-0.12	0.20	...
...	...	...	...	...

32.7
1.8
10.9
...

OBC Heavy

Respondent	S1	S2	S3	...
001_1	1.16	1.66	-0.43	...
001_2	0.35	0.17	1.23	...
001_3	1.04	-0.12	0.20	...
...	...	...	...	...

	O1	O2	O3	...
S1	1	1	0	...
S2	0	0	1	...
S3	1	0	1	...
...	...	...	...	...

32.7
1.8
10.9
...

To receive total preference shares for each product the occasion weights are used for building a weighted average over all occasions per respondent. The index-relevance-matrix from the section "Additional Insights" can be derived by comparing the total preference shares for each product with the occasion-specific preference shares.

Utility Estimation

As described in the previous section, we compare modeling alternatives with and without occasion-specific utility estimation. For the models without occasion-specific utility values everything works like in an ordinary volumetric CBC analysis for the HB estimation. For OBC Medium the standard (not occasion-specific) utility values would just be replicated and the resulting preferences are then adjusted by multiplying them with a product-occasion-relevance-matrix to get to occasion-specific preference shares. OBC Light and Heavy on the other hand are based on an occasion-specific utility estimation. Therefore a more extensive data preparation is necessary. The collected CBC data needs to be decomposed into separate choice task sets per occasion based on the BOMPAT questions for the specificity matrix and the occasion volume. After preparing the HB estimation input each task per respondent exists as often as occasions exist.

We tested two different options to do the occasion-specific utility estimation:

1. The first alternative is to include all choice task sets for each occasion in one run. Doing so without specifying any covariates, the HB estimation has no information on

which tasks belong either to the same respondent or to the same occasion. We tested different covariate settings to integrate the information, which choices belong to the same respondent or the same occasion. We integrated just the respondent covariate, just the occasion covariate and both at the same time as well. This leads to 4 different alternatives (including the run without covariate information) for each OBC model with occasion-specific utility estimation:

Table 3: Alternative Covariate Settings for Occasion-Specific Utility Estimation

Covariate	I.a	I.b	I.c	I.d
respondent		x		x
occasion			x	x

While trying to run the models I.b and I.d including the respondent covariate we were not able to successfully run the HB estimation with CBC/HB version 5.5.3. Therefore we ran OBC Light/Heavy I.b and I.d with the ChoiceModelR package in R.

2. The second alternative is to set up one run per occasion.

Within the following sections the results of all the modeling alternatives will be compared and summed up with a recommendation in which way occasions should be integrated into the OBC modeling.

Application and Results

Case Study

To test our approach to Occasion-Based Conjoint (OBC) and compare the alternative approaches to utility estimation and simulation we used a case study for a gummy bear manufacturer[2]. The manufacturer was planning to launch a new pack size and wanted to test several alternatives along with their pricing. The survey objective was to find out which pack size best completes the current product range (regarding market potential and cannibalization).

The sample consisted of 636 buyers of gummy bears within the past 3 months. There were quotas on age, gender and the number of children in the household. The data were collected in a 20 minutes online survey. The Conjoint used a typical FMCG set up consisting of only two attributes: (1) SKU (combining brand and pack size) and (2) price. Every respondent answered 12 random tasks and two holdout tasks. The following table summarizes the conjoint attributes, their coding and the number of levels (i.e., number of utilities estimated per respondent). It was the same for all modeling alternatives (Volumetric CBC, OBC Medium, OBC Light and Heavy).

[2] Category name is changed.

Table 4: Attributes and Number of Parameters in Conjoint Design

No	Attribute	Coding	# Level
1	SKU	Part-Worth	29
2	Price[3]	Log-linear	1
3	Price Thresholds[4]	User specified	1
4	None	Part-Worth	1
	Total[5]		**32 (31)**

In the following we will compare the modeling alternatives in three steps: (1) the comparison of the OBC Light alternatives (different approaches to occasion-specific utility estimation), (2) the comparison of the OBC Heavy alternatives (again different approaches to occasion-specific utility estimation), and finally (3) comparison of the winners for OBC Light and OBC Heavy with OBC Medium and Volumetric CBC. But before we go through these three steps of model comparison we give a short explanation of the validation criteria that we use throughout this process.

Validation Criteria

To compare the modeling alternatives and the alternative approaches to estimate the occasion-specific utilities, we looked at several criteria of internal and external validity as well as face validity. *Internal validity* was assessed by looking at aggregate as well as individual holdout prediction using aggregate MAE, aggregate R-Squared and the individual holdout hit rate of the models. To assess *external validity* we looked at the aggregate MAE and R-Squared of market share prediction. *Face validity* is generally described as the fact that a test measures what it is supposed to measure. It is often evaluated by checking whether the results meet certain expectations. We assessed face validity by looking at substitution effects; both holdout scenarios differ only in one SKU, one replaces the other. Both SKUs are from the same brand and differ only a little in packaging. Hence, a similar choice share should be expected for both. Comparing the choice shares of the two packaging alternatives between holdout tasks an index can be calculated with 100 indicating equal shares. In the two holdout tasks we observed an index of 96.9. In the following tables for each model the deviation of the simulated index from the observed index is given, the smaller the deviation the better. The simulated indices are also given in brackets.

Results for OBC Light

To see which occasion-specific utility estimation worked best for OBC Light the following table summarizes the results of the 5 alternatives. For alternatives I.a to I.d we estimated the utilities in one run with different covariates, for II (see section on utility estimation) we ran 6 separate utility estimations in CBC/HB.

[3] The Log linear price parameter is in the estimation constrained to be negative.

[4] Price Threshold parameters were constrained to be positive in the estimation.

[5] The number in brackets denotes the number of parameters to be estimated taking into account that one level can be deleted for attributes estimated based on part-worth coding.

Table 5: Results for OBC Light

		Occasion-Specific Utility Estimation				
		I.a	I.b	I.c	I.d	II
Number of Runs		**1**	1	1	1	6
Covariates	Respondent		x		x	
	Occasions			x	x	

Internal Validity: Holdout Prediction	aggregate MAE	**0.661**	0.641	0.648	0.627	0.636
	aggregate R^2	**0.923**	0.927	0.930	0.931	0.929
	individual hit rate	**47.7%**	47.7%	49.1%	47.8%	49.3%
External Validity: Market Share Prediction	aggregate MAE	**3.436**	3.434	3.413	3.421	3.441
	aggregate R^2	**0.013**	0.013	0.015	0.014	0.012
Face Validity: Substitution Effects	AE* (holdout: 96.9)	**0.050** **(92.0)**	0.104 (86.5)	0.117 (85.2)	0.114 (85.6)	0.099 (87.0)

We see little differences in internal and external validity. The small differences in external validity do occur because the market share prediction is generally very bad. This was most probably caused by the peculiarities of the category under study. The market definition used in the conjoint study was tailored specifically to the client brand. The market data that we used for comparison had therefore to be collected from different sources and they were hence not completely comparable to the preferences measured in the study. Therefore we based our choice of the winning approach mostly on the face validity criterion where we see stronger differentiation between the different procedures of occasion-specific utility estimation. Regarding the face validity, I.a, which is the pooled estimation without covariates, clearly outperforms the other alternatives.

Results for OBC Heavy

The same table as for OBC Light is given for OBC Heavy to see which occasion-specific utility estimation worked best for the OBC Heavy approach. This approach differs from OBC Light only in the fact that it uses the specificity matrix in the simulation model.

Table 6: Results for OBC Heavy

		Occasion-Specific Utility Estimation				
		I.a	I.b	I.c	I.d	II
Number of Runs		**1**	1	1	1	6
Covariates	Respondent		x		x	
	Occasions			x	x	
Internal Validity: Holdout Prediction	aggregate MAE	**0.751**	0.740	0.731	0.731	0.713
	aggregate R^2	**0.909**	0.912	0.915	0.914	0.917
	individual hit rate	**49.7 %**	49.7 %	50.4 %	49.7 %	50.6 %
External Validity: Market Share Prediction	aggregate MAE	**3.444**	3.439	3.418	3.425	3.456
	aggregate R^2	**0.013**	0.013	0.013	0.014	0.010
Face Validity: Substitution Effects	AE* (holdout: 96.9)	**0.049** **(92.0)**	0.106 (86.4)	0.103 (86.6)	0.099 (87.1)	0.100 (87.0)

For OBC Heavy we see almost the same picture as for OBC Light. There are little differences in internal and external validity. So we again decided based on the face validity to pick option I.a as the winner.

Validity Comparison of Alternative Models

Now in the last step we want to find out which method worked best overall. Standard volumetric CBC serves as a benchmark that is purely based on the results of the conjoint exercise and is not augmented with any occasion related information that was collected before and after the conjoint exercise.

Table 7: Results for Alternative Modeling Approaches

			Volumetric CBC	Occasion-Based Conjoint (OBC)		
				Medium	Light I.a	Heavy I.a
Utility Estimation	Occasion-Specific		no	no	yes	yes
	Covariate	Respondent	-	-	-	-
		Occasions	-	-	-	-
	Number of Runs		1	1	1	1
Multiplication with Occasion-Specificity			no	yes	no	yes
Internal Validity: Holdout Prediction	aggregate MAE		0.502	0.708	0.661	0.751
	aggregate R^2		0.958	0.911	0.923	0.909
	individual hit rate		50.8%	48.6%	47.7%	49.7%
External Validity: Market Share Prediction	aggregate MAE		3.427	3.453	3.436	3.444
	aggregate R^2		0.010	0.008	0.013	0.013
Face Validity: Substitution Effects	AE* (holdout: 96.9)		0.114 (85.5)	0.085 (88.5)	0.050 (92.0)	0.049 (92.0)

Regarding internal validity, the volumetric CBC clearly outperforms the OBC-alternatives. Also in term of external validity it appears to be a little stronger but again there are little differences to be seen on this criterion. Regarding face validity, however, volumetric CBC has the weakest performance of all approaches. So in terms of validity, it is only the face validity that provides justification to augment the CBC data with occasion-related information.

But as described in the motivation section, there are important other reasons for using an occasion-based approach, most notably it enables us to give additional insights that provide valuable support for strategic decisions. Occasions give important hints to the motivations behind the choice that cannot be provided by standard conjoint. So if we were to pick one of the versions of Occasion-Based Conjoint that we tested, we would pick OBC Light. All OBC models show a comparable performance in terms of validation criteria. OBC Medium, however, falls back in terms of face validity where OBC Light and OBC Heavy perform equally well. So it boils down to a choice between these two. And here practicability comes into play. OBC Light is simpler and more flexible compared to OBC Heavy because a specificity matrix is not necessary. The combination with the specificity in the simulation model can be problematic when it comes to calibration because effects of the utility calibration can be neutralized by the multiplication with the specificity. So it is advantageous to have all the information on occasion-specific preferences in the utilities and not in the utilities and the specificity matrix combined.

SUMMARY AND OUTLOOK

The integration of occasions into CBC has potential to improve the predictive power (e.g., sales potential and substitution effects) and to generate additional insight that can be used to improve marketing decisions for existing and new products.

The first obstacle to overcome to realize an Occasion-Based Conjoint (OBC) is the measurement of occasion-specific preferences without overburdening the respondents. We proposed and tested a particularly efficient way that collects overall preferences for all occasions in each choice task. These overall preferences have then later to be decomposed in the modeling process by using information about the relevant occasions and the relevance of products for these occasions ("Specificity").

The second obstacle is to develop a modeling, utility estimation and simulation procedure that yields stable and predictive preference shares. Here we compared three alternative approaches in order to yield a deeper understanding of how far and in which way occasions can and should be integrated into conjoint modeling. We found that what we call "OBC Light"—an approach that estimates occasion-specific utilities and uses these directly in simulations without multiplying them with the product-occasion-relevance—is the most promising approach. Yet further evidence, especially on external validity, is needed.

All our OBC approaches, however, were in terms of holdout prediction outperformed by standard Volumetric CBC. Although we have to take into account that we are using information from outside the conjoint exercise to predict a conjoint task, we of course hoped that OBC could improve the holdout prediction or be at least equally good. Ideally, if the occasions are a dimension of consumer behavior that is really shaping preferences it should be possible to not only improve face validity and external validity but also internal validity by augmenting conjoint data with occasion information. The fact that the holdout prediction of our OBC approach is not far behind Volumetric CBC indicates that this might be possible. One thing to work on—and here we have to thank Greg Allenby for his comments after our presentation—is the weighting scheme that we are using. The individual volume consumed per occasion does not necessarily equal the weight the occasion has in the purchase decision. How occasion importance can be estimated is accordingly the most important area for further research.

ACKNOWLEDGEMENTS

We would like to thank our discussant Greg Allenby from the Ohio State University and our reviewer Ken Deal from the DeGroote School of Business for their helpful comments on our presentation and article.

Björn Höfer

Susanne Müller

REFERENCES

Allenby, G. M. et al. (2002): Market Segmentation Research: Beyond Within and Across Group Differences. *Marketing Letters*, 13, 3, pp. 233–244.

Dubé, J.-P. (2004): Multiple Discreteness and Product Differentiation: Demand for Carbonated Soft Drinks. *Marketing Science*, 2004, Vol. 23, No. 1, Winter 2004, pp. 66–81.

Kim, J./Allenby, G.M./Rossi, P.E. (2002): Modeling Consumer Demand for Variety. *Marketing Science*, 21, 3, pp. 229–250.

Pinnell, J. (2005): Comment on Huber: Practical Suggestions for CBC Studies. *Sawtooth Software RESEARCH PAPER SERIES*.

R Development Core Team (2011): *R: A language and environment for statistical computing* [Computer software]. Version 2.13.0. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Sermas, R. (2012): ChoiceModelR: Choice Modeling in R.

Sha Yang/Vishal Narayan/Ravi Dhar (2007): An Integrative Model of Consumers' Product Consumption, Need States and Consumption Contexts. *18th Advanced Research Techniques Forum,* Santa Fe, 2007.

Yang, Sha/Greg M. Allenby/Geraldine Fennell (2002): Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions. *Marketing Science*, 21, 1, pp. 14–31.

PRECISE FMCG MARKET MODELING USING ADVANCED CBC

DMITRY BELYAKOV
SYNOVATE COMCON

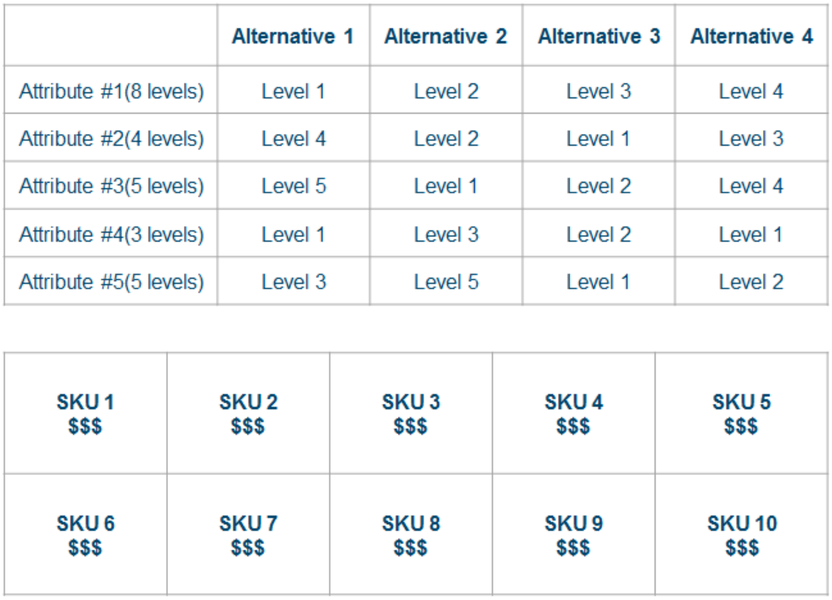
INTRODUCTION

Pricing studies in FMCG markets appear essential for clients. Corporate activities (promo, line optimization, etc.) and product prices depend on these studies. Thus, they have direct impact on sales revenue. Generally, the most common survey research methodology for pricing research is Choice Based Conjoint (CBC) (Orme 2003). However, these studies have several features to be considered for precise modeling. Unlike traditional CBC, we often can't describe a product by using a set of characteristics represented by the respective attributes as in FMCG markets; flavors, sizes and package types are specific to brands. Hence, on the one hand orthogonal design with all combinations available produces unrealistic products, which may confuse respondents. On the other hand, to represent real products too many prohibitions are to be specified and the design appears deficient. (Weiner & Sanches 2012). Therefore, SKU-price CBC suits these studies better. With this approach levels of one attribute correspond to SKUs, while another attribute corresponds to price points.

This type of CBC differs from the traditional one. First, we usually have a lot of SKUs available on the market (dozens or hundreds). While, in traditional CBC, a large number of attributes often poses a problem, in SKU-price CBC, however, the number of SKU-attribute levels becomes an issue.

As a result, respondents can see each SKU a few times only, unless each SKU is available per choice task. At the same time, their preference is very heterogeneous. Due to these reasons some approaches traditionally used in CBC become not suitable in such studies. Moreover, tasks for SKU-Price CBC differ from those for regular CBC (Figure 1). A lot of parameters tested are available in the traditional CBC task. SKU-Price CBC, however, lacks quite a lot of test options in the screen (only 10 SKUs from the list of 40 are present in the example). Thus, one choice at the traditional task gives us a lot more information. That makes SKU-Price CBC data really sparse.

Figure 1. Examples of Traditional CBC (Top) and SKU-Price CBC (Bottom) Screens



	Alternative 1	Alternative 2	Alternative 3	Alternative 4
Attribute #1(8 levels)	Level 1	Level 2	Level 3	Level 4
Attribute #2(4 levels)	Level 4	Level 2	Level 1	Level 3
Attribute #3(5 levels)	Level 5	Level 1	Level 2	Level 4
Attribute #4(3 levels)	Level 1	Level 3	Level 2	Level 1
Attribute #5(5 levels)	Level 3	Level 5	Level 1	Level 2

SKU 1 $$$	SKU 2 $$$	SKU 3 $$$	SKU 4 $$$	SKU 5 $$$
SKU 6 $$$	SKU 7 $$$	SKU 8 $$$	SKU 9 $$$	SKU 10 $$$

In addition, simulation differs a lot as well. First, we simulate the real market with continuous purchases. Second, we have to consider the fact that in most cases a respondent's several favorite SKUs cover his or her demand almost fully. So we have to recover the utilities on these SKUs for this respondent as precisely as possible. In addition, we need to recover their price functions accurately. Third, the analyst should pay a lot of attention to obtaining accurate price elasticity. In fact, plenty of secondary indices depend on the demand-price relationship. For customers, these figures link to their sales directly.

Finally, to compare the models obtained as well as estimation methods, we need a different approach for SKU-price CBC as it turns out that traditional measures of performance work poorly.

Research Approach and Measures of Performance

We can't select holdouts used traditionally. Recent Sawtooth Software papers showed that a couple of holdouts are not enough for a reliable model test (Orme 2015; Chrzan 2015). If we increase the number of holdouts, we'll have to reduce the number of tasks for estimation. This is not really good for sparse SKU-price CBC data. On the other hand, out-of-sample holdouts increase the sample size dramatically. Moreover, if we look at a respondent's answers (Figure 1) we can see that traditional CBC needs quite a lot of information to simulate the choice of the holdout, while with SKU-price CBC we only use utilities of the SKUs available in the task. At the same time, a respondent's favorite SKUs may be missing. That actually indicates low relevance of his choice with this holdout as in reality he would buy other products. However, when evaluating the model, his opinion is taken into account together with other respondents' choices. Therefore, using holdouts doesn't turn out to be the best measure of model or method quality with SKU-Price CBC. This makes it hardly possible to investigate this topic using empirical studies only. Moreover, with empirical studies conclusions are usually made based on just a few datasets, which could be not reliable enough. Hence, it is better to use systematic Monte-Carlo simulation study for comprehensive investigation.

However, even with MC-simulation study we cannot use the traditional measure of performance—the correlation between input and obtained (estimated) utilities. If we convert individual level utilities into preference shares, we'll see that only a few SKUs account for a major contribution, while the rest have near-zero shares. That is, from the modeling point of view utilities of -1 and -9 have the same "zero" impact. So, we have to recover utilities of a respondent's favorite SKUs accurately, while all others appear irrelevant. But if we consider correlations, they're important as well. We may have a situation when we fail to recover the best SKU's utilities, but manage to define precisely all the rest. In terms of modeling prospects, this is bad, although it results in a high correlation coefficient.

Therefore, in this paper I use the following measures of performance to support the modeling purposes:

- Mean Absolute Error (MAE) of Shares of Preference (SoP) between true and estimated current scenarios (all SKUs at the current prices)

- MAE of Price sensitivity calculated as follows:

$$\mathrm{M}= \frac{1}{N}\frac{1}{P-1}\sum_{i=1}^{N}\sum_{j=1}^{P-1} abs\,(\left[\frac{SoP_{ij}}{SoP_{ij+1}}\right] input - \left[\frac{SoP_{ij}}{SoP_{ij+1}}\right] obtained)$$

Where:

N—Number of SKUs,
P—Number of price-levels,
SoP$_{ij}$—Share of preference of SKU I at price level j (all other SKUs at current price),
Input—"true" simulated utilities,
Obtained—utilities after robotic respondents' answers estimation

This measure shows how, on average, we fail to predict price sensitivity (term very closely connected with price elasticity) in our study correctly.

All in all, we can conclude that SKU-Price CBC is a particular area of conjoint analysis with its own issues and solutions. In this paper I investigate two common challenges of SKU-price CBC—long lists of SKUs and accurate modeling of price elasticities. The paper is organized as follows. In the next section I examine possible ways to handle long SKU lists using an approach called "Consideration Sets." The later section deals with techniques to improve the price elasticity modeling accuracy.

Long SKU Lists

Recently clients asked us to include more and more SKUs for more realistic market simulation. If a study deals with 100 SKUs, even with design of 15 tasks per respondent and 15 SKUs per task, each respondent can see each SKU 2.25 times on average. This is absolutely not enough for accurate utility estimation at the individual level. Fortunately, this issue can be solved using an approach called "Consideration Sets." With the approach each respondent only evaluates the products relevant to him/her. Hence, a respondent does not waste time evaluating irrelevant products, the CBC exercise becomes more engaging and we get more helpful information when a respondent selects between his/her favorite SKUs. We can either ask a respondent to select manually which SKUs he/she would consider or the Set can be formed based on a hierarchy of the respondent's answers about category consumption (questions about brands, format, packages, flavors, etc. preference or usage). The second one is more useful in the case of long lists.

Using Consideration Sets we reduce data sparsity at the individual level by increasing the number of "useful" SKU appearances, at the same time assuming that SKUs outside this Set have "zero" attractiveness for this respondent. But the question is—how to code the SKUs not included in this Set? At the SKIM/Sawtooth Software European Conference, Pupke and Rausch (2013) presented the analysis of several ways to code the raw CBC choice-file. However, the conclusions for empirical and simulated studies differed and the authors admitted the necessity for further research. In my opinion, they used too homogeneous samples in simulations and only one holdout as a measure of performance. So I decided to continue this study.

Pupke and Rausch used the following methods of encoding (original description is kept):

1. Tasks as they were shown during the interview. Utilities of non-considered SKUs will be estimated based on the upper level model[1]
2. Each task will be enriched by additional concepts for all non-considered SKUs
3. like (1) with one additional task including all non-considered SKUs and NONE-option is chosen (see also York & Hall 2000)
4. like (1) with one binary choice task for each non-considered SKU (SKU vs. None-option)

4a. like 1) with one binary choice task for each SKU (accept or reject according to the consideration set)[2].

In my analysis, I have made some changes to these methods.

For Option 1 I suggest recoding utilities for out-of-set SKUs to -99 after estimation. Otherwise, utilities for out-of-set SKUs can be higher than those for in-set SKUs for a respondent. That leads to unrealistic switching between products in market simulations. For example in a beverage category small and big packages suit different purchase occasions. However, if a respondent's Consideration Set includes only small-size SKUs HB can draw high utility values for the most popular (at sample level) SKUs in big packs. As a result, when simulating the market, we increase the price for a can and see a growing share of big bottles. While in reality, a single small can is not an alternative to a big bottle. Recoding to -99 in this case resolves this problem. And in this way, more realistic switching between SKUs becomes an advantage of the Consideration Set approach.

Option 2 was used unchanged. I didn't use Option 3, because this method doesn't lower utilities for out-of-set SKUs sufficiently enough and Option 4 as it is a simplified version of Option 4a. For Option 4a, unlike the German authors, I used a user-specified "anchor" attribute level instead of "none" as it improves the results. Pupke and Rausch in their work weren't sure how well HB could estimate binary and conventional tasks together, so I added another option to the research:

5. like (1) with one task for each considered SKU. The in-set SKU is chosen over all out-of-set SKUs. Accordingly, the number of additional tasks equals the set size.

These coding methods were compared in MC simulation studies with 70 and 100 total SKUs (Consideration Set included 30 SKUs) using measures of performance described earlier. Table 1

[1] The same as estimation approach "dropped levels not available" within ACBC

[2] The same as estimation approach "dropped levels are Inferior" within ACBC

shows MAE of current scenario demand shares and Table 2 reports MAE of price sensitivities. In my opinion, a more accurate reproduction of price elasticities is more important. Firstly, the current shares can be calibrated with audit data, while calibrating elasticities is much harder. Secondly, as described earlier, a lot of secondary indices in price studies are based on price-demand relationships. Therefore, to reproduce it accurately is especially important.

Table 1. MAE of Demand Shares for Current Scenario

	1	**2**	**4a**	**5**
70 SKUs	0.29	0.227	**0.202**	**0.202**
100 SKUs	0.282	0.185	**0.147**	0.167

Table 2. MAE of Price Sensitivities

	1	**2**	**4a**	**5**
70 SKUs	0.129	0.128	**0.125**	0.138
100 SKUs	0.152	0.158	**0.149**	0.173

Anyway, Option 4a (ACBC “dropped levels are inferior”) outperforms all others on both performance measures. Option 1 (ACBC “dropped levels not available”) recovers elasticity quite well, while it come last in terms of the current demand shares prediction. Option 2 tends to overstate elasticities which was confirmed by commercial studies. Option 5 works well only for share prediction. Also, with this method HB estimation takes a very long time and the choice-files are huge! So, when using Consideration Sets I recommend coding choice-files adding a set of binary tasks comparing an SKU with an “anchor” attribute level.

ACCURATE ELASTICITY MODELING

As was described earlier plenty of important secondary indices in FMCG pricing studies are based on the demand-price relationship. Price elasticity is a good indicator of this relationship so in this section I investigate the accuracy of the elasticity modeling. The theoretical demand-price curve looks very simple (Figure 2a). In reality, though, things are more complicated. The first factor is psychological price thresholds (Figure 2b). Where we have round price values, demand falls more and violates smooth monotonic decrease[3]. The next problem is that different SKUs may react to changing prices differently (Figure 2c). Some show a demand drop faster than others. Another issue is called “asymmetrical elasticities”—it was shown by Arink et al. (2010) that price elasticity to lowered prices can differ from price elasticity to raised prices (Figure 2d). To generalize that, the curve can change the angle several times. For example, a flatter fall around the current price, and, on the contrary, a sharp change at the edges of the test price grid.

[3] Here and further on, prices are provided in Russian rubles to keep real thresholds. For reference, 1USD totals approx. 36 rubles.

Figure 2. Elasticity Modeling Issues
(a)

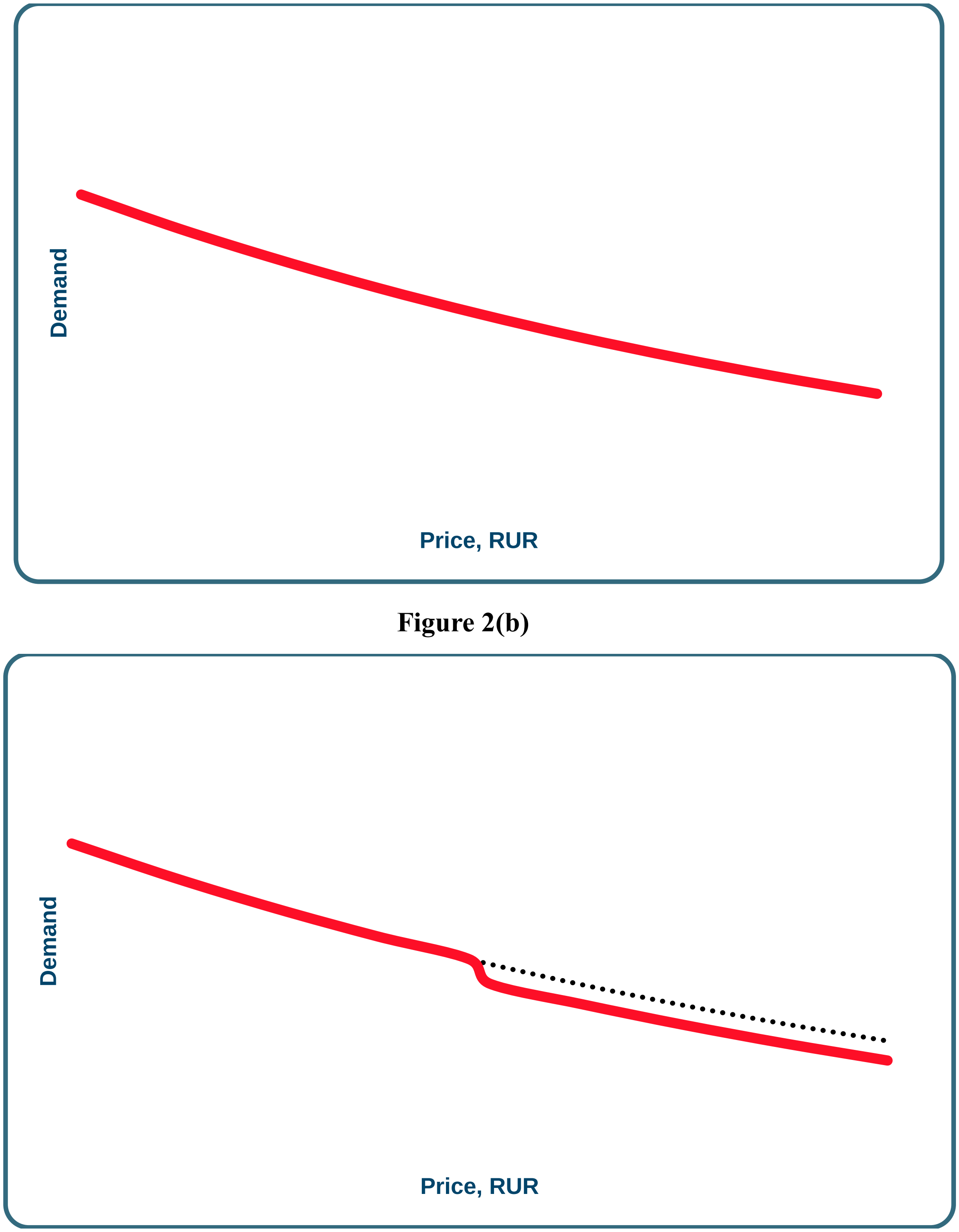


Figure 2(b)

Figure 2(c)

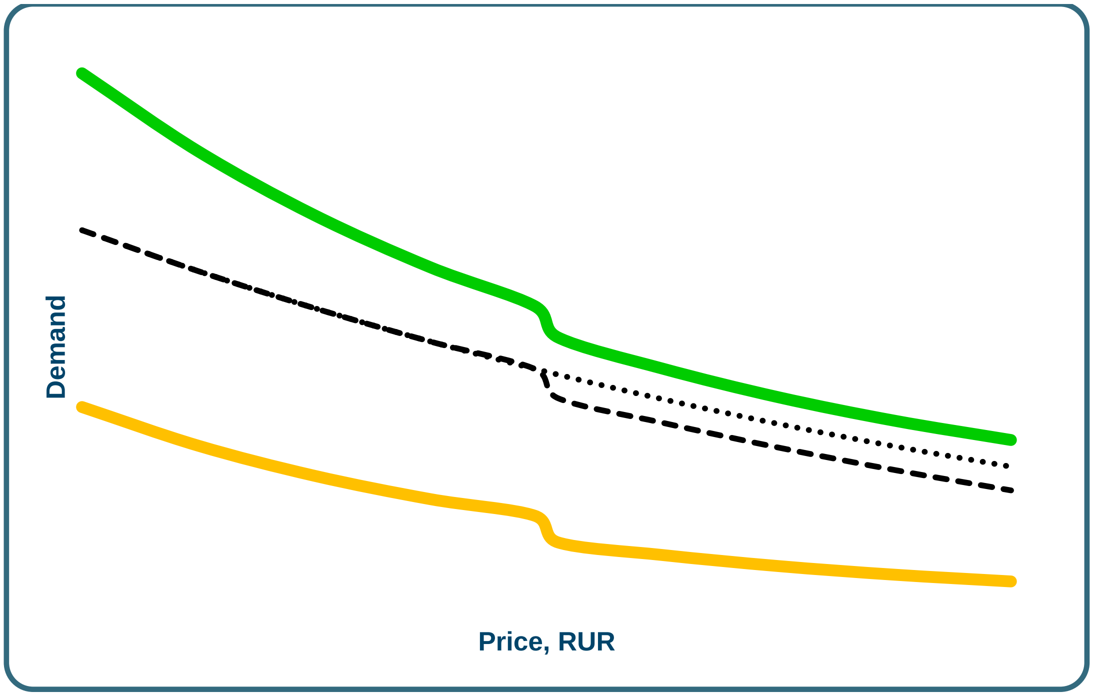
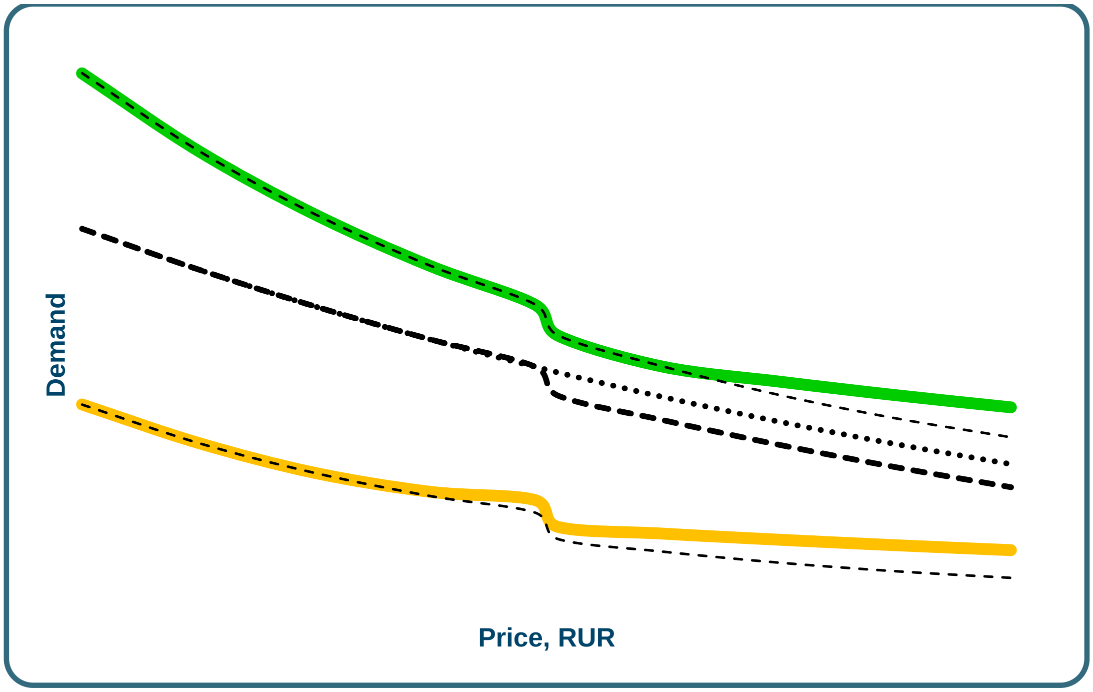


Figure 2(d)



We should consider these issues for accurate elasticity modeling in FMCG markets. In SKU-Price CBC the main feature responsible for the elasticity accuracy is a price-attribute. Therefore, this research focuses on techniques which allow us to consider the issues mentioned above in price-attribute estimation. Table 3 summarizes general ways to code the price-attribute as well as to set the pricing grid. It focuses on two main approaches to set the price grid for a study:

- ***Conditional Pricing.*** This method requires a researcher to specify the same number of price levels for each SKU—set as the same proportional price changes from the average (current) price for all SKUs.

- ***Alternative-Specific Pricing.*** This flexible approach allows a researcher to test a different number of price points for different SKUs and assign arbitrary test prices.

In estimations price attribute levels can be used as steps of ***conditional*** pricing, or as absolute values—***continuous*** pricing. We can use the slope or part-worth coding methods taking into account the main effects only or adding the SKU-price interaction.

Table 3. Summary of Price Attribute Encoding Options

	Slope (Linear, Log-linear, Piecewise)		**Part-worth**	
	Main effects	**SKU-Price Interactions**	**Main effects**	**SKU-Price Interactions**
Conditional	Assym. Elast.	Var. Elast. Assym. Elast.	Assym. Elast.	Thresholds Var. Elast. Assym. Elast.
Continuous	Thresholds (+-) Var. Elast. (+-) Assym. Elast.	⇧	Thresholds	*Leads to extremely sparse data*

The corresponding boxes show which of the issues described above a method can successfully deal with. All *conditional* methods can work with asymmetrical elasticity. Adding the *SKU-price interaction* does resolve the issue of varying elasticities. *Part-worth coding* with interactions is theoretically the most powerful method. In practice, however, we have too sparse data for it. In theory, *continuous with slope coding* is assumed to cope with all three issues, which, however, requires additional recoding efforts from a researcher and is not always possible in reality. The *interaction* model doesn't differ from those for *conditional* fundamentally. The main advantage of *continuous part-worth coding* is the ability to capture thresholds without any effort from researchers. As it calculates utilities for each tested price value, the model specifies demand drops at the thresholds automatically. Adding the *interactions* to the model doesn't seem reasonable. That will result in hundreds of additional parameters.

When choosing a method for dealing with the price attribute the main dilemma is to define the optimal level of the model complexity. A too simple model doesn't account for all effects required. If we make the model too complex, we increase the estimation error. This means that, again, we can't account for all the existing effects accurately as it performs like two combined equations with three unknown quantities. At the same time, the golden mean will be different for different parameters of the study. Since theoretically the most powerful coding options work poorly with sparse SKU-Price CBC data we should improve the simplest ones. The extended study is organized in the following way. Based on the Monte-Carlo simulation study, I describe improvements for thresholds incorporation and the capture of varying elasticities for part-worth coding. Implementing these techniques, I compare "part-worth coding" and then the "linear coding" methods.

For each of the conditions, after analyzing real projects, a population of 10,000 robotic respondents was generated to select the sample size required. Tested factors are reported in Table 4. I didn't use the scheme when all factor levels are tested with all. The fact is that some of these combinations are hardly possible in reality. In addition, in some radical circumstances, some methods may strongly outperform others. This can cause a bias when the final results are averaged. So I chose another strategy. Level 2 presented in Table 4 corresponds to the basic scenario—that is average real conditions. To test each factor level, we should substitute it in the place of the corresponding level in the basic design. The study used the same CBC design with 15 tasks and 12 alternatives per task and 5 equidistant price points across all tested factors combinations.

Table 4. Monte-Carlo Simulation Study Design Specification

Experimental Factors	**Level 1**	**Level 2**	**Level 3**
Sample size	500	1000	1500
SKU Number	20	40	60[4]
Varying elasticities	No	yes	
Thresholds	No	yes	high
Asymmetrical elasticity		no	yes

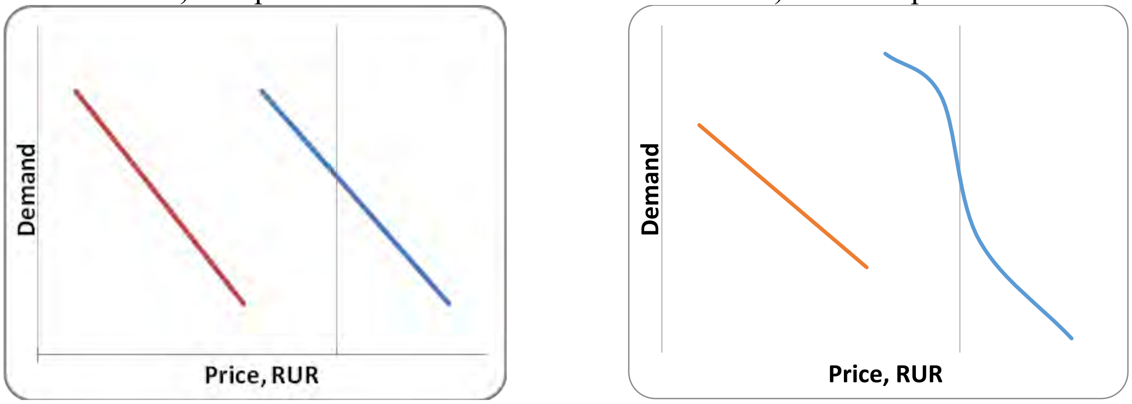
Thresholds Incorporation

The same level of the price attribute can be below the price threshold value for one SKU and above it for another (Figure 3a). If we calculate one price function, not including the SKU-Price interaction, we'll get the average scenario (Figure 3b). Moreover, for SKU 2, by interpolating the curve between two points, we get something that contrasts the reality a lot.

Figure 3. Thresholds Illustration

a) real picture

b) obtained picture

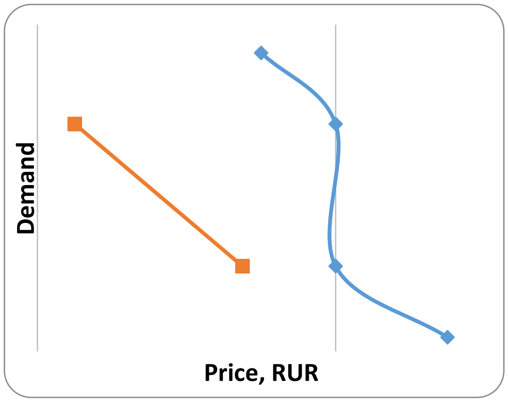


[4] Sample size of 1500 was used

We can embed the threshold by adding a user-specified attribute (so-called "threshold" attribute). If the SKU price is lower than the threshold value, it's coded as "0," otherwise as "1." Thus, the utility of this attribute provides additional demand drop when the price passes through the threshold value (Figure 4) and the total price utility is summed up from the utility of the corresponding price attribute level and the utility of the threshold attribute:

$$U_{price}=U_{pr.att.}+U_{thr.att}$$

Figure 4. Thresholds Incorporation Illustration



The problem is that we need to know the threshold price value. The simplest and most logical way is to assign threshold attributes to round price values. We can check the validity of this method. Table 5 provides the results for synthetic data. In one case, data generation didn't include the thresholds. In another case the threshold was embedded at the price of 50 rubles. Assigning the threshold attribute, in turn, to round values ranging from 46 to 55 rubles and calculating its utility with aggregate logit we can assess the demand drop obtained and find the price point most suitable for a threshold. As one can see, this approach allows us to recognize that the first case has no thresholds, while in another case, the price point of 50 rubles shows the most significant drop. That means this approach indicated true threshold.

Table 5. Threshold Search (Synthetic Data)

	Price, RUR	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**
Threshold *was not* implemented in true data	**Utility**	0.10	-0.04	0.06	0.03	-0.03	-0.01	-0.08	-0.06	0.03	-0.05
Threshold *was* implemented in true data		0.01	-0.07	-0.07	-0.07	**-0.21**	-0.17	-0.17	-0.08	-0.08	-0.06

Doing the same with real data (Table 6), we can see that the price around 50 rubles gives the largest drop too. Sometimes (like in Study #1) the next point after 50 rubles shows the most tangible drop. In this case, researchers need to define if they assign the threshold at 51 rubles, or leave it at 50 rubles or somewhere in-between. That refers to all assumed thresholds.

Table 6. Threshold Search (Empirical Data)

	Price, RUR	46	47	48	49	50	51	52	53	54	55
Study #1	Utility	-0.13	-0.14	-0.17	-0.19	-0.20	**-0.22**	-0.19	-0.16	-0.15	-0.15
Study #2		-0.05	-0.03	-0.11	-0.13	**-0.15**	-0.14	-0.11	-0.10	-0.11	-0.09

Continuing the investigation of the threshold incorporation approach, Table 7 compares part-worth main effects only method (without threshold incorporation) with five thresholds solution (to all round values from 30 to 70 RUR) and only three basic thresholds (30, 40, 50 RUR based on a test) using MAE of price sensitivities. It should be noted that the numbers can be compared only within a row (as different data was used for different MC design factors). For all but one case, adding "threshold" attributes improves the results. As a model with only one part-worth attribute is quite simple, the "5-threshold" option performs better.

Table 7. Threshold Attributes Effectiveness Investigation

	Part-worth, Main effects only	+5 thresholds	+3 thresholds
Base design	0.133	**0.125**	0.131
Sample size			
500	0.143	**0.126**	0.135
1500	0.123	**0.12**	**0.12**
Thresholds			
No	**0.108**	0.117	0.113
High	0.153	**0.127**	0.143
Varying elasticities			
No	0.103	**0.102**	0.105
Asymmetrical elasticity			
Yes	0.108	**0.106**	0.111
SKU Number			
20	0.127	0.122	**0.119**
60	0.115	0.113	**0.103**

Further on, based on my experience with empirical and simulated data I can suggest a few recommendations about using this approach:

- Remember to use constraints in HB.
- Avoid using a lot of “threshold” attributes. Only assign them to necessary thresholds.
- In order to verify the need of adding a threshold attribute for a certain price, do the following: calculate utilities for the threshold attribute in HB without adding constraints. Then, look at the sample average. If it is near zero, do not add the threshold.
- In addition, you should monitor data adequacy. If you have just a small amount of observations above or below threshold price it will be impossible for HB estimate the threshold attribute accurately.

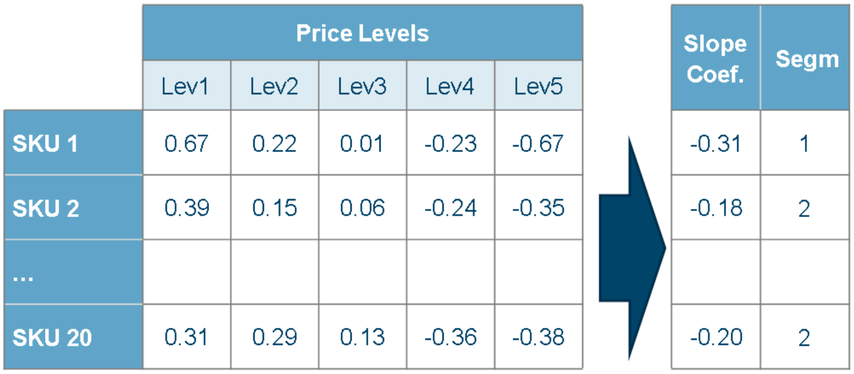
Capturing Varying Elasticities

In most FMCG studies the SKU list is quite long (more than 20 SKUs) so we cannot estimate SKU-price interactions precisely. To capture different elasticities for different SKUs without adding the “interaction” we can assign a price attribute for the whole group of SKUs with a similar elasticity (rather than for each SKU). Thus, we don’t have to calculate a lot of price attributes. At the same time, we allow SKUs from different segments to show different elasticities. In fact, people don’t have a built-in computer for shopping. Subsequently, they hardly have a unique price function for each SKU in their mind. They seem more likely to realize that they can pay more for one SKU, while they’ll only buy another SKU at a discounted price. That’s why adding several segment attributes seems enough.

To define the segment membership we can use our expert knowledge of the market (premium brands are less price sensitive than mass segment, unique propositions are less sensitive as there are no alternatives), or the actual values of product volumes and prices.

We can also use Raw CBC data. Figure 5 shows a simple algorithm to allocate SKUs to sensitivity segments. We calculate Aggregate Logit by adding the SKU-price interaction and get a set of utilities for each SKU-price combination. Then we calculate the slope coefficient and segment it. This simple approach (more complicated and precise approach based on HB estimation will be described later) doesn’t take much time and was used for segmentation in this study where hundreds of simulations were to be conducted.

Figure 5. Sensitivity Segmentation Algorithm



	Price Levels				
	Lev1	Lev2	Lev3	Lev4	Lev5
SKU 1	0.67	0.22	0.01	-0.23	-0.67
SKU 2	0.39	0.15	0.06	-0.24	-0.35
…					
SKU 20	0.31	0.29	0.13	-0.36	-0.38

Slope Coef.	Segm
-0.31	1
-0.18	2
-0.20	2

Table 8 examines how segment numbers affect the results using this segmentation algorithm. “Main effects + 5 thresholds” solution from the previous step is compared with different segments numbers (from 2 to 5). When using segments, only 3 thresholds are added, as with this number of calculation parameters, 5 thresholds turned out to be “too many.”

Table 8. Segments Number Investigation

	Main effects, +5 thresholds	**Alternative-specific, sensitivity segments, +3 thresholds**			
		2 seg	**3 seg**	**4 seg**	**5 seg**
Base design	0.125	0.116	0.113	**0.11**	0.134
Sample size					
500	**0.126**	0.169	0.177	0.179	0.162
1500	0.12	0.113	**0.111**	**0.11**	0.129
Thresholds					
No	0.117	**0.115**	**0.115**	0.123	0.133
High	**0.127**	0.136	**0.127**	0.154	0.163
Varying elasticities					
No	**0.102**	0.11	0.111	0.112	0.115
Asymmetrical elasticity					
Yes	0.106	0.115	**0.102**	0.11	0.121
SKU Number					
20	0.122	0.138	0.141	**0.15**	0.158
60	**0.113**	0.126	0.133	0.138	0.145

On average, the 3 segment solution performs better. With the growth of data sparsity, *main effects* (one segment) works better. Based on the experience of using segments I can provide some tips:

- In my opinion, using three segments is the golden mean. You have to be careful, though. Even if all SKUs have the same price elasticity using sensitivity segments approach you can get slightly distinguishable utilities for the resulting segments, although not as much as in the case with real presence of different elasticities.

- In real projects, you can somewhat adjust an SKU segment membership based on expert knowledge. If we look at the resulting segments, we can understand what they mean and what SKUs get to "wrong" segments. That increases the accuracy of estimation. I didn't do that with simulated data, so in reality the method of segments will prove even more superior to main effects.
- As we complicate the model, we have to sacrifice something. In this case, five thresholds didn't work as efficiently as three did. This shows that you have to compare what is more important to you: to capture lots of thresholds (if any) or different elasticities.

Part-Worth Coding Methods Comparison

The final comparison of part-worth methods is summarized in Table 9. To those discussed earlier—"Main effects + 5 thresholds" and "Sensitivity segments 3 segments"—I added "SKU-Price Interaction" for conditional pricing, and "alternative-specific continuous pricing" (where we obtain part-worth utility for each of the tested absolute price values) with 25 price points.

Table 9. Part-Worth Coding Methods Comparison

	Main effects +5 thresholds	**Sensitivity segments, 3 segments**	**SKU-Price Interaction**	**Alt. specific continuous**
Base design	0.125	**0.113**	0.18	0.114
Sample size				
500	**0.126**	0.177	0.219	0.135
1500	0.12	0.111	0.139	**0.108**
Thresholds				
No	**0.108***	0.115*	0.143	0.143
High	0.127	0.127	0.19	**0.095**
Varying elasticities				
No	0.102	0.111	0.144	**0.095**
Asymmetrical elasticity				
Yes	0.106	**0.102**	0.123	0.122
SKU Number				
20	0.122	0.141	0.155	**0.109**
60	**0.113**	0.133	0.205	0.127

The main finding of the comparison is that different methods suit different conditions. The general rule—the more sparse the data, the more simple model should be used. The *SKU-Price interactions* method provides the worst results (too complex model for really sparse SKU-price CBC data). *Main effects* as the simplest method is better in the case of larger numbers of SKUs or small sample size. The *Alternative-specific continuous* method captures thresholds naturally. If we look at the threshold factor, we'll see that it's far superior to other methods in the case of strong thresholds. At the same time, it loses to them without any thresholds. Based on my experience, for inexpensive categories such as drinks and chocolate (i.e., products that cost a few dollars) thresholds show low impact. With regard to cheese or elite alcohol, though, thresholds play an important role. For these categories, I recommend using alt. specific continuous. However, you should remember that the number of price points is important as well. Here I used 25, because, for example, 50 price points lead to lower accuracy. Researchers should evaluate whether this number of points is enough to cover the test price range normally. For example, when testing the whisky category, we encountered a great variety in the current prices for different SKUs. Using 30 price points was critically not enough. The price grid looked strange not meeting all research related needs. At the same time, working with elasticities, spread and asymmetry appears the weaknesses of this method. The *Sensitivity segments* approach (alternative-specific attributes for groups of SKUs based on conditional pricing) allows us to deal with varying elasticities successfully.

Linear Coding Methods Comparison

We can use this method for conditional pricing to code percentage of the current price, as well as for continuous pricing to use absolute values. The advantage allows us to use a flexible pricing grid. With both methods of coding, we can use the approaches to capture price thresholds and varying elasticities. Unfortunately, due to time limits, I couldn't test both encoding methods completely. So after some fast checking, I chose *conditional* coding as it works better in the presence of varying elasticities.

In addition, using this approach you can get more precise sensitivity segments definition (an alternative to the algorithm described earlier). If you add SKU-price interaction to HB estimation, the resulting sample means for SKU price slopes can be used to define sensitivity segments. It is a little more accurate comparing with aggregate logit but takes much more time, so I couldn't use it this study.

The results for the final linear coding methods comparison are reported in Table 10.

Table 10. Linear Coding Methods Comparison

	Main effects +5 thresholds	**Sensitivity segments, 3 segments**	**SKU-Price Interaction**	**Best from Part-worth coding**
Base design	**0.1**	**0.1**	0.115	0.113
Sample size				
500	0.115	**0.111**	0.151	0.126
1500	0.104	**0.098**	0.114	0.108
Thresholds				
No	**0.094**	**0.094**	0.111	0.108
High	0.117	0.118	0.161	**0.095**
Varying elasticities				
No	**0.08**	0.1	0.128	0.095
Asymmetrical elasticity				
Yes	**0.083**	0.084	0.131	0.102
SKU Number				
20	0.095	**0.073**	0.15	0.109
60	**0.099**	**0.099**	0.169	0.113

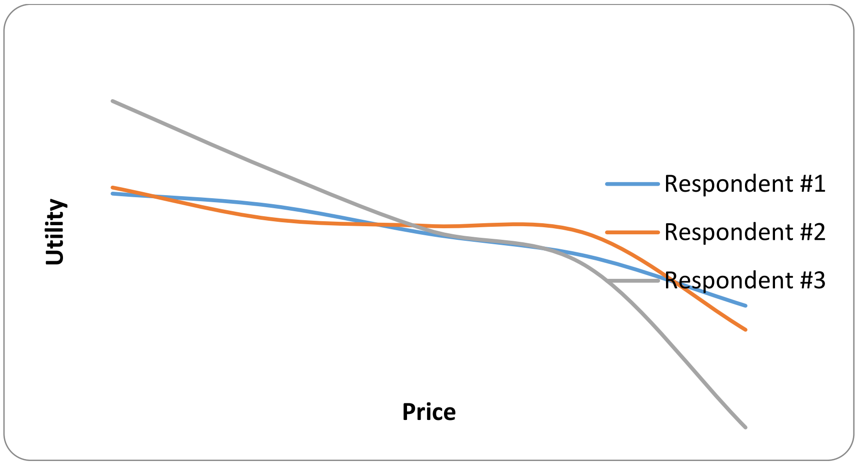
Even with linear coding, *interactions* method does not work well enough. So we need alternative-specific price attributes for sensitivity segments here as well. As we can see in this case, *Sensitivity segments* prevails over all other methods. *Main effects* are better only if thresholds influence a lot and, quite logically, without varying elasticities. At the same time, linear coding strongly outperforms part-worth coding. It's only in the case of high thresholds that alternative-specific continuous part-worth coding shows the best results.

Individual Conditional Price Threshold Investigation

The findings discussed above could be final and in which case I'd recommend using linear coding. But one issue was discovered during this research. For simulations, I used monotone linear price function with absolute value thresholds incorporated. After we take into account price thresholds by using a user-specified "threshold" attribute, the remaining price function becomes linear. Given that, linear coding is a priori in privileged position. It is quite uncommon in realistic situations when a simpler linear model is enough for correct modeling, while a more complex part-worth can only add an error but not improvements. But what if the conditional price function actually has a polyline form at an individual level? In other words, what if there's an individual conditional threshold for a person?

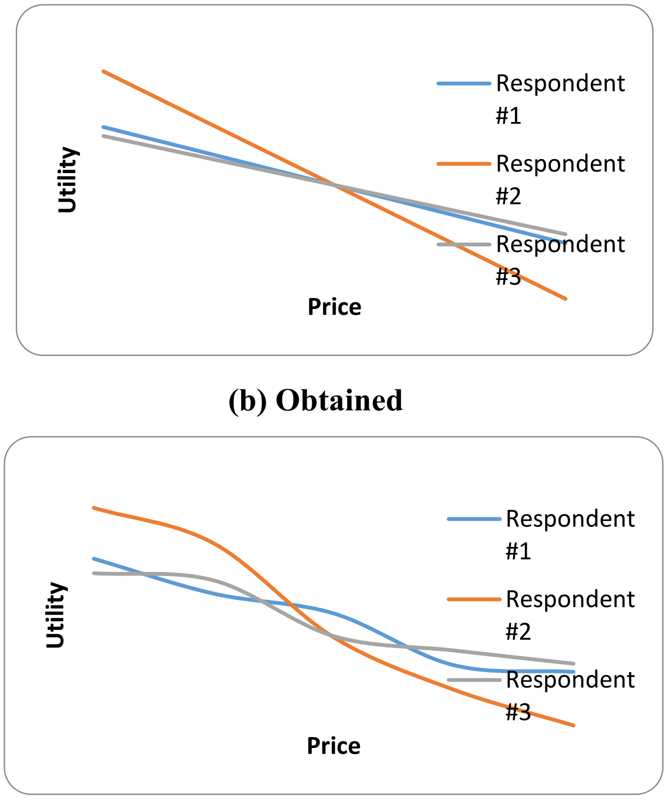
I can explain this effect in the following way. When dealing with the FMCG category, people know the approximate level of their favorite products' current prices. Thus, price absolute value isn't always as crucial. If one's favorite chocolate costs $1.96 on average, he would hardly reject to buy it when it costs $2.01. This price is only a few cents higher than he expected. At the same time, if he sees the product at the price which is strongly higher than the current level, he wouldn't buy it regardless of the absolute price. That means there's an individual level conditional threshold. That threshold could be at different levels for different people (for example + 20% for respondent #1 and + 30% for respondent #2). The same could be for lowering prices. In this case, part-worth coding could fit better.

Figure 6. Example of Part-Worth Utilities from an Empirical Study



If we look at the individual part-worth utilities from the empirical study (Figure 6) we can see that respondents' price functions are not linear. They have a different shape. But at the sample level, price function has a linear shape. Is it an individual conditional threshold or simply the result of a calculation error? I tried to check this with simulated data. I generated data for linear price functions (Figure 7a), but within the output I got nonlinear ones by using part-worth estimation (Figure 7b). I used various simulation parameters. The resulting price functions still had a nonlinear shape. We could think that the broken line shape is the result of a calculation error, but empirical data deviated from the linear shape even more. Another interesting point showed that the larger the sample size, the less deviation from the linear course is seen at the individual level. But even with big samples, we still had it.

Figure 7. Example of Part-Worth Utilities from a Simulated Study



(a) Input

(b) Obtained

For more detailed investigation of this effect I used the measure of nonlinearity obtained as follows (Table 11):

- Calculate slopes between adjacent price levels taking the utility difference (since conditional pricing grid price points are equidistant)
- Convert them to percentage for the normalization
- Take respondent's standard deviation as a measure of his/her nonlinearity
- Average them across the sample as well as the maximal slopes

Table 11. Measure of Nonlinearity Calculation

Respondent #i	**Level1**	**Level2**	**Level3**	**Level4**	**Level5**
Utilities	1.26	1.18	0.58	-1.22	-1.80
Difference	0.08	0.60	1.81	0.58	
Percentage	2.5	19.5	59.0	19.0	
Maximal slope	59	**St.Dev.**			20.8

The standard deviation is more correct statistically, while the maximum slope is more evident.

To evaluate the effect I took a real empirical study (1800 respondents; 16 tasks and 12 alternatives per task; 30 SKUs and 5 price points) and calculated price attribute utilities using two approaches: part-worth and linear. Both of the resulting data sets were used as input data for Monte-Carlo simulation study. Using these datasets I got two sets of synthetic CBC answers—based on linear and nonlinear input price functions. Both raw CBC choice-files utilities were calculated using part-worth coding. Table 12 compares the nonlinearity of the resulting price functions with the input one.

Table 12. Nonlinearity Investigation

	Maximal slope	**St.Dev.**
Initial data	51.1	17.2
By part-worth coding (nonlinear input)	51.8	17.5
By slope coding (linear input)	47.6	15.0

It seems that the empirical data was obtained based on non-linear individual price functions. I checked this effect with another project, and got the same findings. But it turned out that even if we have individual thresholds, we can only capture them with big samples. Thus, with a sample of 1000 respondents or under, we'd rather use linear coding.

CONCLUSIONS

When dealing with price studies in FMCG Market we should remember that they have their own features and require particular techniques. Some solutions traditionally used with CBC studies do not work well in the case of SKU-Price CBC with long SKU lists. Selecting a model we should keep in mind that data sparsity often doesn't allow us to specify all the effects assumed. Hence, sometimes less complex but stable solutions should be preferred.

Dmitry Belyakov

REFERENCES

Arink, M., Nef, V. & Favrelle, A. (2010) Janus and the changing face of pricing research. *Presented at the ESOMAR Congress 2010.*

Chrzan, K. (2015) How Many Holdout Tasks for Model Validation? *Sawtooth Software Research Paper Series.*

Orme, B. (2003) Special features of CBC software for packaged goods and beverage research. *Sawtooth Software Research Paper Series.*

Orme, B. (2003) Including Holdout Choice Tasks in Conjoint Studies. *Sawtooth Software Research Paper Series.*

Pupke, K. & Rausch, M. (2013) Coding of choice files when using a consideration set. *Presented at SKIM/Sawtooth Software European Conference.*

Weiner, J. & Sanches, M. (2012) Being creative in the design: performance of Hierarchical Bayes with sparse information matrix. *Proceedings of the 2012 Sawtooth Software Conference.*

SELECTION BIAS IN CHOICE MODELING USING ADAPTIVE METHODS: A COMMENT ON "PRECISE FMCG MARKET MODELING USING ADVANCED CBC"

THOMAS C. EAGLE
EAGLE ANALYTICS OF CALIFORNIA, INC.

This short paper is a result of a discussion of the paper entitled, "Precise FMCG Market Modeling Using Advanced CBC," I gave at the 2015 Sawtooth Software conference. The paper presented was very good. It covered a lot of issues in modeling fast moving consumer goods (FMCG) quite well. There were two issues that were of particular importance to me: 1) the handling of large numbers of SKUs, and 2) the modeling of the elasticity effects in such models. I divide this discussion into these two components.

SELECTION BIAS AND THE HANDLING OF LARGER NUMBERS OF SKUS

The author considered the issue of modeling large numbers of choice alternatives (100+) in the area of FMCGs. I, too, have spent quite of bit of effort modeling such large numbers of SKUs. The paper suggested one approach by which the SKUs are decomposed into a more manageable number of attributes and levels. This approach is similar to that originally proposed by the Fader and Hardie (1996) JMR paper, "Modeling consumer choice among SKUs." Rightfully, the author rejected this approach because such decomposition of the SKUs into independent attributes and levels ignores the distinct possibility that the entire uniqueness of the SKU is not captured by these attributes. In other words, the modeling of the unique SKU is capturing all the interaction effects that may exist among the attributes decomposing the SKUs.

Instead of the above approach, the author suggests performing an adaptive two stage task with respondents. Stage 1 asks the respondents to self-select a subset of SKUs from the entire set of SKUs for use in a Stage 2 choice model derived from a collection of customized choice tasks. It is adaptive in the sense the respondent controls the selection of SKUs as opposed to the researcher controlling this selection. This is a process whereby a respondent self-selects an evoked consideration set, which is then used to create the choice tasks the respondent would see in Stage 2.

My primary concern with such self-selection of SKUs is that the sample used in the modeling of the Stage 2 choice model is no longer random or under the control of the researcher. That is, observations are no longer random representations of the population of interest. They become determined by the very outcomes we wish to model! This is a form of selection bias written about by the Nobel Prize winning economist James Heckman. Heckman's article published in 1979, "Sample selection bias as a specification error," clearly points out the bias that can result in models where the sample is derived from the dependent variable. Boehmke (2004) states in his entry in the *Sage Encyclopedia of Social Science Research Methods*:

> *Selection bias is an important concern in any social science research design because its presence generally leads to inaccurate estimates. Selection bias occurs when the presence of observations in the sample depends on the value of the variable of interest. When this happens, the sample is no longer randomly drawn from the population being studied, and any inferences about that population that are based on the selected sample will be biased. Although researchers should take care to design studies in ways that mitigate nonrandom selection, in many cases, the problem is unavoidable, particularly if the data-generating process is out of their control. For this reason, then, many methods have been developed that attempt to correct for the problems associated with selection bias. In general, these approaches involve modeling the selection process and then controlling for it when evaluating the outcome variable.*

Any adaptive process we employ in our research based upon the outcome variable is susceptible to selection bias. In more technical terms, the error term in the Stage 2 choice model is no longer independent of the Stage 1 selection process. While the parameters may accurately predict results using internal validation, the literature clearly shows the parameters derived with selection bias in them, and inferences made from those parameters, will not predict as accurately to the overall population as a model accounting for the selection bias. Feinberg et al. (forthcoming) clearly demonstrate the parameters of choice models are affected by the limiting of the choice alternatives to those the respondents selected themselves. Both the classical and Bayesian choice models reflected these affected parameters.

We can ameliorate this bias. Heckman and others proposed a two stage modeling approach. Stage 1 models the probability the observation will be part of the sample. Stage 2 uses the predictions from stage 1 as additional variables in the stage 2 model to separate out the correlated errors. This is usually in the form of an instrumental variable. Including such a term makes the stage 2 model conditional upon the selection made in stage 1.

Another approach is to simultaneously model the consideration set process (i.e., the self-selection of SKUs by respondents) and the choice model conditional on the consideration set process. Feinberg et al. (forthcoming) discuss one simultaneous approach (i.e., capturing the correlation among the two components of the model). An earlier example of such simultaneous modeling is the paper by Terui, Ban, and Allenby (2011). Lastly, another paper that considers this in more traditional "frequentist" estimated choice models is that by Carson and Louviere (2014), though I believe the concepts espoused in this paper extend even to Bayesian estimation. While the Terui et al. paper does not address self-selection per se, it does demonstrate the simultaneous modeling of consideration sets and the conditional choice model. The Feinberg et al. (forthcoming) and Carson and Louviere (2014) explicitly focus on alternative selection and the bias it introduces.

This topic is "old hat" in aggregate choice modeling literature in Econometrics (e.g., Dubin and Rivers, 1989; Greene, 2014) and Transportation (e.g., Zhou & Lyles, 2014) but largely ignored in marketing and especially ignored by marketing practitioners. This seems especially true when I brought up this issue at the conference and most practitioners there had not heard of selection bias or had thought about how many of the adaptive practices that have been proposed were subject to selection bias. Of course, this sample of practitioners is subject to selection bias itself!

An underlying question is how large is the bias in the self-selection of choice outcomes in a choice model? To be even more specific, how large a problem is it in the hierarchical Bayesian models we practitioners fit? I do not know, nor do many others, I expect. It is an area for future research. Simple simulations relying on measures of internal validity (such as using holdout tasks and in-sample MAEs) are not appropriate tests of such bias. Out-of-sample testing strictly controlling the data generation process, simulating the self-selection of choice outcomes from a population where evoked sets are not an issue, and predicting to a holdout sample is required. Additionally, explicit modeling of the two processes accounting for the correlated error between them is required.

Regrettably the Belakov paper does not directly test for selection bias, nor was it intended to test this. Rather, the paper discusses ways to include out-of-set SKUs in estimation or simulation using rescaled utilities and/or data augmentation and comparing in-sample MAEs and hit rates. None of which model selection bias.

While the papers by Feinberg et al. (forthcoming) and Carson and Louviere (2104) are rigorous academic treatises on selection bias, we can find earlier practitioner efforts to examine the impact of using evoked sets. A previous Sawtooth Software Conference paper by York and Hall (2000) more closely resembles the type of research that can begin to model selection bias. In the York and Hall paper the authors examined two separate samples:

1. A sample who saw a more traditional CBC task where all possible combinations of brand and package attributes were shown (a total of 14 brands and 12 packages = 168 potential SKUs). These are essentially decomposed SKUs by brand and package.
2. A sample who were allowed to select a subset of brand and packages and then saw customized choice tasks using only those selected brands and packages.
3. Each respondent in both samples saw 15 tasks of 3 SKUs each.

The authors saw marginal improvements in 1st choice hit rates to internal holdout tasks by the sample who developed evoked sets; while, in terms of MAE, the full consideration set proved marginally better. While the York and Hall paper is not quite the same as the Belakov paper, there is some suggestion that the selection bias may be marginal. York and Hall also used a data augmentation approach to capture the effects of out-of-set SKUs. Interestingly, in his discussion of the York and Hall paper, Orme (2000) states that the customized tasks improved individual-level hit rates, but that aggregate share predictions were worse than the sample who used the full consideration set. This result could very well be the result of selection bias. Nevertheless, no attempt was made to model both components of the process.

Another underlying question is whether all adaptive methods we use in adaptive choice based conjoint are subject to such selection bias. Any adaptive process that eliminates alternatives as a function of attribute levels is affecting the sampling of choice outcomes. However, other aspects of adaptive methods designed to focus attention on some attribute levels more than others may not be subject to selection bias. My "gut-feel" is that not all adaptive methods used on the independent attributes of the utility function of choice models are necessarily affecting the sampling of choice outcomes (e.g., Liu, Otter, and Allenby, 2007). Feinberg et al. (forthcoming) suggest that sampling on the "X" values (e.g., attributes in the utility function) is not a source of selection bias. I doubt if the likelihood principle or "being Bayes" suffices to resolve whether customized evoked set selection bias exists in HB MNL models

How do we handle projects with large numbers of SKUs? We build experimental designs manipulating the presence and absence of the 100+ SKUs. Designs where alternatives are randomly drawn from the full set of SKUs is one approach to take. Every SKU has an equal probability of being seen by each respondent. Instead, we control the number of these SKUs being shown in each task; we systematically control the number of times each SKU is shown; and we control the pairwise appearance of the SKUs in a carefully constructed choice design. We also control the numbers of similar and competing SKUs to make the task resemble more of a true store shelf. New SKUs are introduced and removed via the same presence and absence design. The tasks will have anywhere from 20 to 30+ SKUs appearing at one time. Because the presence and absence design is systematically controlling for the SKUs, we have reduced any selection bias in the design or in the modeling. The full model is fit with the 100+ SKU constants and their associated price effects (effects often made generic across several individual SKUs—see below). The modeling process is the same as that used when using self-selected evoked sets of SKUs, only there is no selection bias by design.

ELASTICITY EFFECTS

There are two issues raised by the author I would like to briefly discuss regarding elasticity, or price, measurement in models with large numbers of SKUs.

First, do consumers have price thresholds? I suspect many consumers/respondents do have thresholds. If they do exist, does every respondent have the same threshold? If respondents do not have the same threshold, are the thresholds normally distributed across the sample/population (or distributed using any other probability distribution)? If respondents do have thresholds, and they differ across respondents according to some unimodal (multi-modal?) distribution, then we should be able to find that threshold depending upon the probability distribution we assume. If, however, the thresholds are more uniformly distributed across respondents, then can thresholds be said to exist? A purely uniform distribution of thresholds across the respondents, even if they are measured accurately at the individual level, would be reflected by a linear price effect in the aggregate and in the upper level of a hierarchical model. It is my experience in trying to test for thresholds in hundreds, if not thousands, of pricing models I have fit in over 30 years of choice modeling that they are extremely hard to find. This is especially true once we aggregate up from individual-level models.

I like the approach espoused by the author of using aggregate level modeling to try and pinpoint if and where a price threshold might exist before applying said results to the individual-level models we typically fit. It is an approach I have used for many years, even back in the days before HB MNL model existed. Using piecewise price coding is an approach I frequently use in modeling price thresholds.

Second, I wish to comment on alternative specific pricing. The author wisely suggests that fitting alternative specific price parameters across 100+ SKUs is impractical. In consumer goods, especially FMCGs that do not have clear price tiers, I cannot imagine respondents having alternative specific price sensitivities.

In the old days when we were restricted to fitting aggregate level models, alternative specific price parameters worked really well. However, most of what they were capturing was heterogeneity across respondents and across the alternatives. I have found in most FMCG studies the vast majority of alternative price effects (again, usually at the request of the client) I have modeled using HB methods have indistinguishable parameters across alternatives. This can be examined by closely examining the confidence intervals across alternatives at the aggregate level of our HB MNL models, or across draws for any single individual. Why do these effects seem to overlap/disappear upon examination? They disappear because we now have the ability to model heterogeneity much better than ever before.

This is reassuring to me because I strongly believe that the value of a dollar is a dollar for any single individual within a single category of FMCGs. Why would my value of a dollar be different for a diet soda and a regular soda when I have captured my preference for diet sodas over regular sodas with appropriate parameters? Economic theory would certainly suggest the value of a dollar does not vary within an individual.

The author describes a good approach to collapsing alternative specific price parameters into more generic effects across groups of alternatives. Using aggregate modeling and segmentation methods to collapse the 100+ SKU price parameters to a more meaningful manner are good ways to make our models more parsimonious. I have not tried using such approaches. I will start doing so. Critical *a priori* thought of what SKUs could have generic price parameters is better than fitting alternative (SKU) specific price effects. This is the approach I have typically undertaken. I would wager a dollar that after estimation of the model, careful examination of the confidence intervals in the upper level of the MNL model would suggest very few of these generic price parameters are significantly different from one another.

Thomas C. Eagle

REFERENCES

Boehmke, Frederick. (2004). "Selection bias." In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *Encyclopedia of social science research methods.* (pp. 1011–1012). Thousand Oaks, CA: SAGE Publications, Inc.

Carson, Richard T. and Jordan J. Louviere. (2014) "Statistical properties of consideration sets." *The Journal of Choice Modeling*. 13: 37–48.

Dubin, Jeffrey A. and Douglas Rivers. (1989) "Selection Bias in linear Regression, Logit, and Probit Models." *Sociological Methods and Research* 18 (2): 360–390.

Fader, Peter.S and Bruce Hardie. (1996) "Modeling Consumer Choice Among SKUs." *Journal of Marketing Research* 33 (November): 442–452.

Feinberg, Fred, Linda C. Salisbury, and Yuanping Ying. (forthcoming) "When Random Assignment is Not Enough: Accounting for Item Selectivity in Experimental Research." *Marketing Science*.

Greene, William. (2011) *Econometric Analysis*. Prentice-Hall: Upper Saddle River, NJ.

Heckman, James. (1979) "Sample selection bias as a specification error." *Econometrica* 47 (1): 153–61.

Liu, Qing, Thomas Otter, and Greg Allenby. (2007) "Endogeneity Bias—Fact or Fiction." *Proceedings of the Sawtooth Software Conference*. pp 345–350.

Orme, Bryan. (2000) "Comment on York and Hall." *Proceedings of the Sawtooth Software Conference*. p. 111.

Terui, Nobuhiko, Masataka Ban, and Greg Allenby. (2011) "The Effect of Media Advertising on Brand Consideration and Choice." *Marketing Science* 30 (1): 74–91.

York, Sue and Geoff Hall. (2000) "Using Evoked Set Conjoint Designs to Enhance Conjoint Data." *Proceedings of the Sawtooth Software Conference*. pp 101–109.

Zhou, M. and R. Lyles. (2014) "Self-Selection Bias in Driver Performance Studies." *Transportation Research Record* (1573): 86–90.

DEFINING THE EMPLOYEE VALUE PROPOSITION

TIM GLOWA
GARRY SPINKS
ALLYSON KUPER
BUG INSIGHTS

ABSTRACT

While most applications of conjoint focus on solving marketing problems, human resources (HR) also is increasingly using this tool. This paper discusses how conjoint can be used to design, deliver, and position employee benefits and rewards in a more effective and efficient manner, specifically around the subject of the aging workforce. The paper outlines how conjoint can be leveraged in order to position total rewards programs as key elements of the employee value proposition, resulting in the ability to more effectively retain Baby Boomers on the brink of retirement while also attracting talented Millennials who can fill talent gaps and step into leadership roles. Finally, this paper addresses some of the specific challenges of applying this analytical tool in the HR space.

INTRODUCTION

In an ever-evolving marketplace, the competition for talented employees is at an all-time high. Baby Boomers are beginning to retire, resulting in a widespread forecast of talent shortages. Coincidentally, as Baby Boomers leave the workforce, Millennials who lack professional experience move into more senior level positions. A joint survey by The Society for Human Resources Management (SHRM) and American Association of Retired Persons (AARP) shows that US companies are already implementing training programs to prepare for the gap in talent expected when Baby Boomers retire. The same study reports that organizations are making efforts to improve employee benefits in order to attract and retain older employees. With current attraction, retention and engagement trends rapidly evolving, so too must the department of Human Resources (HR).

Despite all the attention or activity in HR, there is very little attempt to fully understand preferences in this space. Our team recognizes HR as a widely underserved market. Since 1987, there have been a total of 473 articles published in the Sawtooth Software proceedings. Of those 473 articles, only three of them have focused on the HR space. Just as marketing aims to design products that satisfy the needs of customers so they buy more and more often, HR desires to understand and satisfy the needs of employees so organizations can attract, engage and retain workers. Taking into account the understated importance of HR, our article will focus on leveraging conjoint analysis for HR purposes. This article will outline an approach for organizations, instructing them on how to see from the perspective of an employee or key stakeholder. The remainder of this paper will focus on employees as consumers of benefits and rewards offered by organizations.

Background

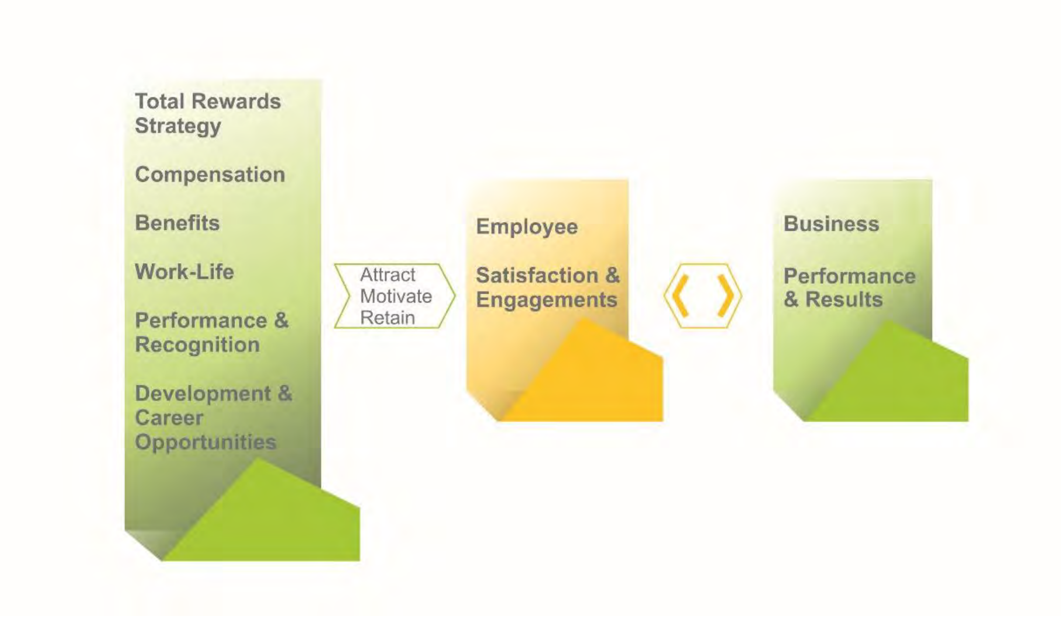
Although the job market is evolving rapidly, 44 percent of organizations have not updated their total rewards strategies for four or more years (Mercer). Because of these outdated total rewards strategies, an average of $1,500 a year (per employee) is wasted by offering benefits that employees do not value or appreciate (Bug Insights). Even companies as large as Microsoft experienced adversity during the downfall of the dot-com era, struggling to attract and retain key talent amidst lay-offs. To try and fix the problem, they conducted a large-scale conjoint analysis to identify which rewards were most valued by their employee population (Slade, Davenport, Roberts, & Shah, 2002). Slade published a follow-up article about total rewards in 2009, which again used conjoint and focused on a case study of rewards and employee turnover at Microsoft. In the article, Slade highlighted the difference between customer research and employee research. Because benefits are typically the third largest expense organizations face, allocating those dollars effectively is critical for maximizing return on investment. Regardless of the dire financial implications of not using quantitative data to make decisions in the HR space, only 32 percent of organizations actually do (Mercer).

When it comes to making decisions on benefits, employees can become frustrated or confused, especially when it comes to medical or dental options. The University of Iowa took this into consideration when they tested whether or not offering dental benefits would be appreciated by their faculty and staff (Cunningham, Gaeth, Juang, & Chakraborty, 1999). They too used conjoint to weigh the risks and benefits of various dental insurance plans.

Through the use of conjoint analysis in HR, employees are more likely to feel like they have a voice in choosing their benefits plans. Most importantly, understanding total rewards preferences is extremely instrumental in the attraction, retention and engagement of employees.

In a February 2015 report, Bug Insights found that 46 percent of the United States workforce is engaged in the workplace. This is problematic in that low employee engagement is detrimental to individuals and organizations. The United Kingdom government released a study in 2013 quantifiably highlighting the benefits of employee engagement. Not only do companies with engaged employees grow at 3 times the rate of companies with disengaged employees, but productivity and customer loyalty are 2 times higher as well. Similarly, organizations with high employee engagement have half the amount of turnover, as engaged employees are 87 percent less likely to leave an organization. This is a soft ROI, but a very real one. Successful organizations reap the rewards of the high correlation between employee engagement and productivity. Additionally, they have higher profitability and, in a retail environment, higher profit and sales per square foot. We also see a drop in employee turnover, which works well with an organization’s bottom line, shareholders and employee bonus pools.

What Is Total Rewards?



Source: WorldatWork

The term *total rewards* is defined as everything an employee receives from his or her employer that he/she perceives as valuable. Total rewards elements can be bucketed into three broad categories: compensation/benefits, work/life-balance, and performance and recognition. Compensation/benefits, the first category, is the most commonly thought of when total rewards is mentioned. It may include, but is not limited to, rewards typical across organizations such as pay, retirement packages, short and long term disability, healthcare benefits and wellness programs. Although the details of these programs may vary, they are often generalized to everyone in the organization, from customer-facing employees to executives.

The second total rewards category, work-life balance, includes often understated but extremely vital benefits. Reward elements that fall into this category focus on how work life and life outside of work are integrated. Organizations can (and often do) fail here in several ways. Employees may feel like they cannot take the time off they want or need, for fear that the work will not get done. Employees, feeling like they are required to work long hours, may feel pressured to constantly check work e-mails due to managers habitually sending e-mails well into the night. In these cases, employees feel like they have to respond even at midnight, from their kid’s sporting events or family dinners on Sundays. They feel unable to set personal boundaries and sacrifice personal time at home, with family, as well as the pursuit of interests outside of the workplace. Work-life balance also applies to the time employees spend at work. Do employees feel like they have time during work hours to take care of personal events they need to attend? Can they attend children’s theater performances, dentist appointments, or muffins-with-mom events? Unfortunately, work-life balance is consistently one of the biggest stated problems that employees face; not surprisingly, it is one of the biggest drivers impacting employee engagement as well.

The final category of total rewards program elements consists of benefits that reward employees for performance and good work. Do employees feel like their work is valued? Do they feel recognized? Can their experiences contribute to development of their career paths? This portion of total rewards is challenging, but not impossible, to measure. Nobody really *likes* to do performance reviews, yet most people want to be recognized and rewarded for the work that they do. If an employee is doing an outstanding job, he/she wants to be recognized, acknowledged, promoted, challenged, etc.

These three total rewards categories make up corporate benefits and rewards. As consultants, we work with clients to examine the total rewards programs that they offer employees in order to identify the most effective combination of benefits which meets the needs of both the employee and the organization. One of the biggest and most common gaps in a total rewards strategy is the ability to meet the needs of the employees; having this information can help an organization gain a competitive advantage for attracting and retaining talent.

Rationale

Traditional Conjoint

Most frequently, conjoint analysis is utilized to solve marketing problems. Taking a look at the last several Sawtooth Software conference proceedings will highlight that virtually all topics focus on marketing. Although conjoint analysis has mostly been used for marketing purposes, there are other spaces where conjoint can positively contribute—one of them being Human Resources. Conjoint can be used to design, deliver and position employee benefits and rewards in a more effective, efficient manner.

Typically, employee benefits are the third largest expense that an organization faces (the first is compensation, the second, cost of goods sold). This expense recurs year after year. Further, attracting, retaining and engaging human capital is the number one priority of most HR departments. Regardless, little rigor or analysis is given to see if this money is spent as effectively as possible. Baum and Kabst (2013) used conjoint to test which employer characteristics potential employees find most valuable when looking for a job. They tested the impact of a variety of organizational and job characteristics as well as the moderating role of involvement. The results (suggesting that involvement increases soft factors of Total Rewards, which decreases the effect of payment of job choice) may surprise many HR departments. That example examined how employers can better attract potential employees, but what about keeping employees engaged? One study used conjoint analysis to better understand Singaporean managers' trade-off attributes of training programs when making executive training decisions. Because of conjoint, this study was able to identify three main important attributes of training programs: word of mouth, trainers' practical experience, and institutional reputation (Gan, Lee, & Soutar, 2009). Without conjoint, the researchers may have been able to identify employee perceptions, but not employee preferences in managerial training.

We estimate that in the United States, employers spend about $25,000 on employee total reward packages annually (benefits in this instance include a conservative amount of paid time off, and do not include compensation, bonuses, or equity. Healthcare, as a benefit, accounts for the largest share of this spending—typically close to $10,000 per employee per year). Interestingly, our findings show that employees estimate that their employers spend a mere $10,080 on total rewards annually. If an organization employs 5,000 people, and is spending $25,000 per employee annually on total rewards, that organization is spending about $125 million that goes largely unchecked on an annual basis. This, arguably, would never happen in the IT or marketing space. Think for a moment about someone from the IT or marketing department coming into the budget planning process asking for an unlimited budget without much, if any, knowledge about whether or not the program would be successful and achieve the desired results.

Despite many organizations having little or no regard for how much their HR departments successfully spend, some organizations (Darden Restaurants, Best Buy, Delta Airlines, Wal-Mart, Google and others) are harnessing the power of conjoint to understand employee preferences for total rewards, using the information to design better, more efficient reward programs; ones that meet the needs of the organization and its most critical stakeholders.

Now that employees are being thought of as consumers of benefits and rewards, it is pivotal to consider the key components of the employee value proposition (*Four Cs).* The Four Cs are the key considerations for defining an employee value proposition.

1. Cost Impact

A certain level of investment must be planned for in order to provide employees with a total rewards package. An understanding of that budget and how much financial backing is available to spend on employees is critical. Typically for companies based in the U.S., benefits are the third largest expense that Fortune 500 companies face, behind compensation and the cost of goods sold. In spite of this, very little rigor goes into evaluating whether or not these massive annual expenses are being spent in an efficient and effective manner. It is disappointing to see that, in many cases, HR leaders are unable to point to empirical data that demonstrates a return on their investment.

2. Competition

Many organizations make decisions primarily based on benefits, but sometimes other rewards as well. Organizations aim to match their competitors (peer groups) in what they are providing. In this case, peer groups could mean a variety of different things, including: companies competing for similar hires (restaurants/retail stores), competitors in the same industry or people in a similar geographic area. Most companies will look at a variety of benchmarks and see how they match up against this set of peer group competitors. They will often position themselves to be at the 50th percentile. Essentially, organizations do not want to be too far above or below the standard of benefits in their industry, which is the exact statistical average of every organization in that group.

For many organizations, the total rewards strategy frequently ends here, and many decisions are made off cost and competitiveness alone. Organizations often want to manage costs while being near the average. This is likely an insufficient total rewards tactic, because this does not give employees any reasons to join the company, stay with the company or be engaged with their work, especially when they can go across the street and attain nearly the same rewards package from a competitor looking to hire.

Consultants often ask HR leaders, “What makes you different than company XYZ across the street?” Many of them, even the chief people officers, have difficulty articulating a difference between their company’s rewards program, and the rewards that a peer group competitor offers. That simply occurs because there often is not a big difference to speak of.

3. Core People Strategy

Some organizations will look at their people strategy and consider the way they want to be viewed in the marketplace. This includes the type of employee they want to attract and retain, who their workforce is today versus who it may be in the future and how they need to adapt their rewards programs to target Millennials vs. Baby Boomers. A certain group of people may have different skill sets. In order to accommodate for this difference, an alteration of benefits programs should be considered.

4. Consumer Preferences and Needs

This aspect of the total rewards framework, understanding the needs and preferences of your employees, is universally overlooked. Organizations must first shift their perspective of employees by viewing their employees as consumers of benefits. This is the first step of understanding consumer preferences.

It should be noted, of course, that reward programs are not exactly like the free consumer market. If employees are not satisfied with their organization's reward plan offerings, they cannot go elsewhere to buy programs that have the same advantages as the ones their employer offers. For instance, an employee cannot get a better vacation package, retirement plan or health benefits without first leaving his or her employer. For this reason, understanding employee preferences is of utmost importance.

The term *preferences (not perceptions)* is utilized with intention here. Measuring employee perceptions around their rewards is a descriptive but not prescriptive measure; it does not, for example, tell the organization much about what needs to be changed, what that change can look like, or what results could be achieved if that change was made. The true value comes from measuring employee preferences—that is, measuring employee needs regarding what total rewards packages provide.

A Better Approach: Taking the Aforementioned 4Cs into Consideration

Aligning all of those key dimensions into consideration will, ideally, lead to the best rewards package—a package that addresses employee needs, makes financial sense to the company and places the organization in a place where they want to be competitively in a thoughtful and strategic way. In terms of competitiveness, for example, a firm might be superior in some reward aspects but at benchmark in others. An organization may simply deem it unnecessary to exceed the industry benchmark in a certain benefit. For example, offering a top-notch life insurance policy may not be prioritized by an organization because that organization does not have a desire to be known as "the best life insurance provider." Taking that into consideration, the organization may decide to reduce the value offered by that benefit, or in some cases rid of it completely.

Aligning these four dimensions is key in order to have a successful employee value proposition. Because HR is not accustomed to making decisions based on data, having a clear outline is very important. As a means to achieve a clear outline, consultants may ask, "How can we design, deliver, and communicate those programs in a way that makes sense to employees and at the organization and provides a win for both sides?" The following preliminary considerations address those questions.

Preliminary Considerations

In order to measure preferences and conduct a significant conjoint study, a number of inputs from the survey design process must be considered:

- **Employee Demographics:** Understanding the demographic makeup of the employee population is important. This informs not only survey design, but also the logistics of communication and survey delivery. A population of CEOs should be surveyed very differently than a population of factory workers, or a population of doctors. Content must be structured to appropriately serve the targeted population; it must be relevant and structured at a level that takes the audience into consideration. Employee turnover is also essential information during this point in the process, and provides insight into how the population has changed over time. Finally, survey data can be analyzed by different population segments, so an understanding of the population and the key people groups is important. These key demographic groups should be identified early on in the survey design process.

- **Current Program Design:** This information primes the entire process, so it must be collected as one of the first inputs. Information from plan documents as well as plan administrators aids in understanding what to test for in both current plan importance and performance, as well as what makes sense to test relative to current plan design. A baseline level (the current state) is always tested as a part of the study. In order to do this, the current program design must be available. Program utilization information must also be collected at this point, including trending information over time (i.e., how and why has utilization for programs changed?).

- **Business Objectives:** Conceptualizing both short- and long-term business goals, as well as how the total rewards program is currently supporting these, is important. These will often inform which programs should (and should not) be tested. They provide an understanding of what is potentially off-limits and what is going to be relevant given the direction of the business in the coming years. Alignment with business objectives is a practice that is often forgotten in the planning stage.

- **Financial Goals:** Budgeting constraints and financial goals are also considered as parts of survey design. The study should not test anything that is impossible to implement, so understanding this information is essential in ensuring that that the study does not test something that is not financially feasible for the organization. A comprehensive understanding of the desired financial outcomes is necessary, whether to find cost savings, remain cost neutral or identify areas of investment.

- **People Strategy:** In both the long and short term, people strategy is an essential input when designing a Total Rewards Optimization study. It provides a framework for what organizations want to be known for, as well as the type of employee that organizations want to attract, retain and engage. The people strategy, and how the total rewards strategy supports it, should be considered when identifying programs to test.

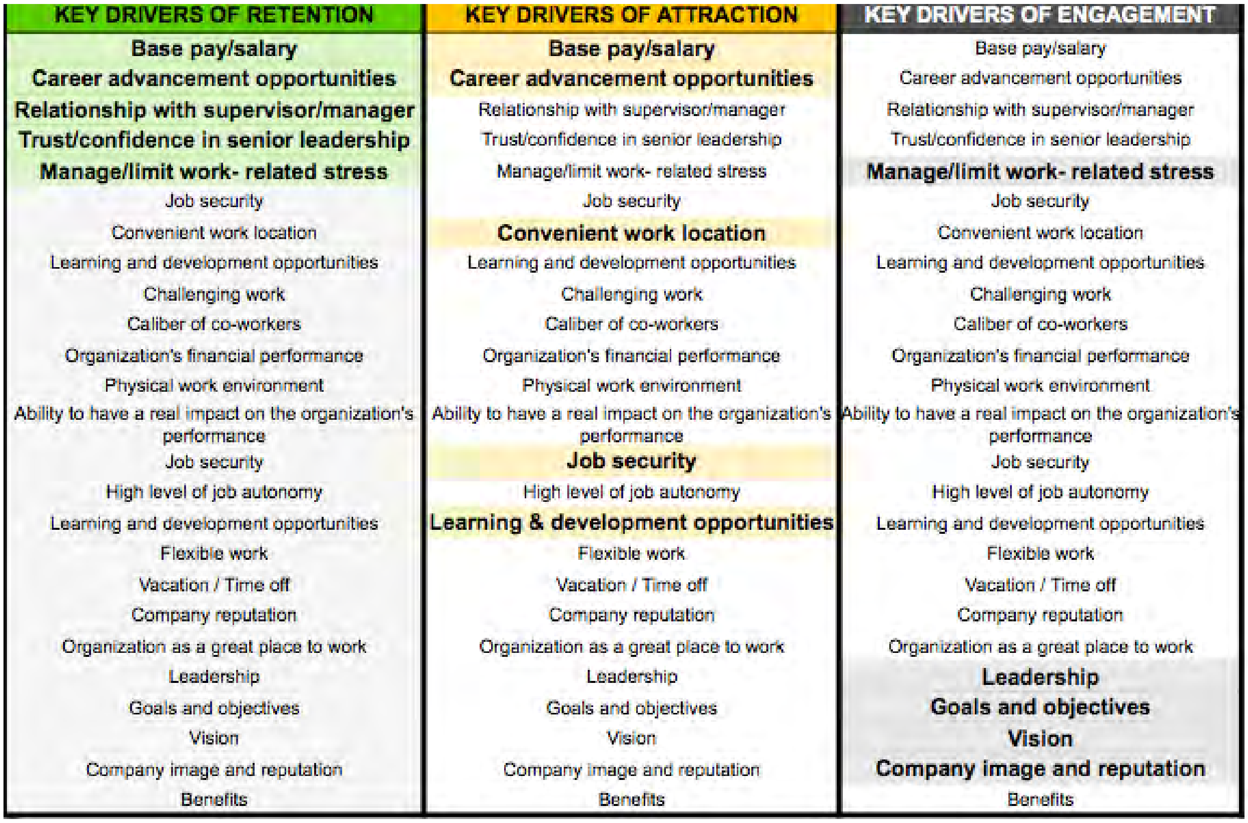
- **Competitive Positioning:** This allows for a grasp of how an organization compares to competitors of the same size, industry, geographic location, etc. It includes norms and benchmarking data and provides insights into how the rewards program compares to competitors. Understanding not only how an organization currently compares to the market but how it wants to be positioned relative to the market will inform the program design elements.

- **Engagement Data:** As a part of the study design, an organization should be able to provide input about current employee engagement levels, preferably segmented by key demographics. Understanding what the engagement data has looked like over time (whether it is trending up or down) and the key drivers of engagement can also help to provide input for what should be tested.

- **Geography:** The geographic location of an organization will impact how the study is designed. Rewards programs often differ geographically, especially from a global perspective, as do business goals and employee demographics. Recognizing and understanding these geographical differences will determine how many survey versions are needed, the languages that the survey should be conducted in, etc.

Attraction vs. Retention vs. Engagement

In addition to those critical inputs, the key drivers of attraction, retention, and engagement should be considered as well for the best results.



KEY DRIVERS OF RETENTION	KEY DRIVERS OF ATTRACTION	KEY DRIVERS OF ENGAGEMENT
Base pay/salary	**Base pay/salary**	Base pay/salary
Career advancement opportunities	**Career advancement opportunities**	Career advancement opportunities
Relationship with supervisor/manager	Relationship with supervisor/manager	Relationship with supervisor/manager
Trust/confidence in senior leadership	Trust/confidence in senior leadership	Trust/confidence in senior leadership
Manage/limit work- related stress	Manage/limit work- related stress	**Manage/limit work- related stress**
Job security	Job security	Job security
Convenient work location	**Convenient work location**	Convenient work location
Learning and development opportunities	Learning and development opportunities	Learning and development opportunities
Challenging work	Challenging work	Challenging work
Caliber of co-workers	Caliber of co-workers	Caliber of co-workers
Organization's financial performance	Organization's financial performance	Organization's financial performance
Physical work environment	Physical work environment	Physical work environment
Ability to have a real impact on the organization's performance	Ability to have a real impact on the organization's performance	Ability to have a real impact on the organization's performance
Job security	**Job security**	Job security
High level of job autonomy	High level of job autonomy	High level of job autonomy
Learning and development opportunities	**Learning & development opportunities**	Learning and development opportunities
Flexible work	Flexible work	Flexible work
Vacation / Time off	Vacation / Time off	Vacation / Time off
Company reputation	Company reputation	Company reputation
Organization as a great place to work	Organization as a great place to work	Organization as a great place to work
Leadership	Leadership	**Leadership**
Goals and objectives	Goals and objectives	**Goals and objectives**
Vision	Vision	**Vision**
Company image and reputation	Company image and reputation	**Company image and reputation**
Benefits	Benefits	Benefits

Source: Towers Watson

Often, a gap exists between factors that drive an employee to join an organization, factors that drive an employee to stay at that organization and factors that keep an employee engaged and motivated to do their best work. Those stages in an employment cycle are all very different. Salary (the biggest driver by far), job security, career/learning development opportunities, benefits packages and retirement are often key considerations when joining an organization. However, when considering employee retention, the drivers are altered. Benefits like salary and career development are both still important, yet other benefits like career development, relationship with manager, trust and confidence in leadership and ability to manage workload become the major factors. In order to maintain employee engagement, considerations like career path, career development, teamwork, involvement in decision-making, availability of tools and resources to get the required job done are all important drivers. Because the differences in employee life-cycle are vast, so should the approach to total rewards be. It is imperative to target different employees at different places in the life cycle instead of using a one-size fits all approach. One-size fits all rarely, if ever, truly fits all.

Method

We opted to conduct a MaxDiff study in order to collect employee preferences for this specific survey. MaxDiff conjoint was used for two main reasons:

1. Difference in Total Rewards Programs Across Companies and Employees

The HR space is very different than marketing, where product optimization typically requires just one study version that applies universally. HR is different because the specific details of total rewards programs vary both between companies and within companies. While many organizations include the same high level attributes as a part of their total rewards programs (i.e., paid time off, retirement, life insurance, etc.), the specific details of these programs will vary. One company may offer a pension program for their retirement, but only two weeks vacation, while another company may offer a richer vacation policy with three weeks off, but only a match for a 401(k) program rather than a pension.

In order to look across organizations and understand industry norms, attributes must be tested at a high level, meaning MaxDiff is a better option. This same concept applies internally for organizations as well, because total rewards programs vary by geography and employee level. For example, a total rewards program in the US will focus heavily on health care, whereas in Canada, that is not the case due to the government-supported health care. For employees in positions of leadership, compensation may include stock options and other bonus opportunities that are not available for non-managers. This variation in program offerings makes applying a very specific Adaptive Conjoint Analysis (ACA) or Choice Based Conjoint (CBC) study challenging, as it does not make sense to ask employees about programs that they do not have. Instead, a better approach is to leverage MaxDiff in this situation. MaxDiff allows the study to be focused at a higher attribute level, rather than drilling down to program specifics.

2. Implications of the Threat of Changing Rewards for Employees

HR leaders are very protective of their employees, for good reason. They typically try to control both the content of information that employees are exposed to and the method by which they are exposed. HR leaders must be careful to manage employees' expectations. They do not want to expose employees to potential benefits that there is no chance of them receiving (for example, employees not eligible for stock options with little likelihood of that changing should not see that as an option), and they do not want to create panic with employees inferring that the take-aways in a study mean that HR is only going to cut benefits. Especially in the last few years as the cost of healthcare has risen and the economy has slowed down, organizations have had to tighten up program offerings, and many have had to slim down rewards offerings (and sometimes even cut programs completely). This makes employees extremely sensitive to mention of changes perceived as losses in a program. A MaxDiff study avoids this issue altogether. The study is less threatening since instead of asking about specific level changes, it simply asks about the best and the worst of high-level programs. Asking an employee which program is most important to them and least important to them is significantly less threatening than asking about taking away three days of paid time off, or increasing the deductible of a healthcare benefit. For these reasons, MaxDiff was chosen for our study.

Our decision to use MaxDiff is not to say that a CBC or ACA study does not have its place in employee research; however, for this purpose MaxDiff made the most sense. In specific cases within organizations where rewards programs are the same across the population tested, communications are carefully managed and HR leadership teams are able to clearly articulate potential changes to reward programs, ACA and CBC studies can be very effective.

Study Detail

For this study, data was collected from Feb. 11, 2015–Feb. 13, 2015 to gather information about employee perceptions of total rewards. A total of 1,316 responses were recorded from participants who are currently employed and are eligible to receive benefits from their employer. The study was self-funded by Bug Insights and the sample was sourced through Survey Sampling, Inc.

Study Purpose

The primary purpose of this study was to collect Best/Worst Conjoint data to be used for this Sawtooth Software 2015 conference. Additionally, we hoped to leverage an opportunity to collect new insights surrounding employee perceptions of total rewards.

Study Design

The study was designed to apply to a broad population, given the survey was not organization, industry or geographically specific. A total of 15 attributes were included: base salary, bonus opportunity, retirement match, paid time off, health care benefits, work-life balance, development allowance, fitness allowance, dental coverage, vision coverage, wine of the month club membership, hobby allowance, tuition reimbursement, car allowance and finally, student loan forgiveness. Each respondent was asked to complete 10 Best/Worst questions as a part of the MaxDiff exercise. Each question included 5 attributes. Two survey versions were used, and respondents were randomly assigned to each version. Half of the sample received Version 1, and the other half received Version 2.

For Version 1, respondents were asked to identify which element of their total rewards program was most important and which was least important to them in order to determine the relative importance of each attribute. In Version 2, respondents were asked which attribute was highest performing and which was lowest performing in terms of keeping them engaged and motivated at their jobs. The purpose of collecting this information was to be able to compare across 2-dimentions of data, both importance and performance.

After the data was collected, an HB analysis was run, and results were analyzed in the aggregate and by several demographics, including: geographic location, age, gender, salary bracket, tenure with company, education and title.

DISCUSSION/RESULTS

Overall Study Outcomes

The chart below outlines the overall results of the two study versions. Variation between the two results indicates either an area where an attribute is important but underperforming, or an area where performance is actually outpacing importance. In the case that performance outweighs importance, this indicates that the program is meeting the needs of employees; however, when performance is falling behind importance, this misalignment indicates an opportunity for improvement for the organization. Whether it is enhancing the program or simply reevaluating the communications strategy, the organization should take notice of the discrepancy.

MAXDIFF STUDY RESULTS:	**Importance**	**Performance**
Base Salary	14.54	14.43
Bonus Opportunity	10.43	8.61
Retirement Contribution	12.04	11.90
Paid Time Off	12.84	12.32
Health Care Benefit	13.75	13.63
Work Life Balance	6.83	7.82
Development Opportunities	3.64	4.01
Fitness Allowance	1.69	2.07
Dental Coverage	6.46	7.62
Vision Coverage	4.42	4.96
Tuition Reimbursement (for You and Your Kids)	4.06	4.19
Car Allowance	2.98	2.35
Student Loan Forgiveness	3.13	3.25
Wine of the Month Club	1.46	1.42
Hobby Allowance	1.73	1.43

While aggregate results are important, some of the most valuable information is found when the results are segmented by demographics.

The Challenge of the Aging Workforce

For this specific study, the segmentation focus was comparing the importance of reward attributes for the Millennial versus Baby Boomer populations. The aging workforce presents a significant challenge for companies today, who must determine the most effective way to retain Baby Boomers while also attracting and engaging Millennials. As the global economy recovers, competition for critical and skilled talent will only get more challenging. Unemployment rates in general are falling, meaning that the rate for skilled workers—which is already much lower—is falling as well. In March 2015, unemployment rate for those with a bachelors degree or more was 2.5 percent compared to 5.5 percent overall, and the labor shortage is projected to last for a quarter century (US Department of Labor, 2015).

The talent issue has become enough of a problem that it has caught the attention of the executive suite. Talent is a top priority for CEOs, even more so than managing risk. World-at-Work (2011) stated it clearly: "As businesses move out of the downturn, CEOs are putting the focus firmly on their people." Leadership understands that "Employees' intentions to leave their current organization are on the rise, climbing back to pre-recession levels" (World-at-Work, 2011).

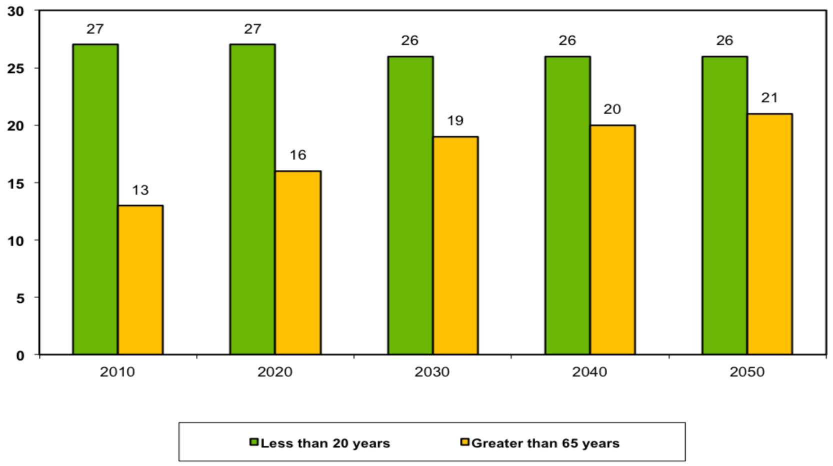
The study reinforces this point, with 2 out of 3 employees agreeing that they would consider leaving their organization for a mere 10 percent increase in salary. The study also found that 50 percent of Millennials believe that they can get a better total rewards package elsewhere. This finding underscores the point that workers are not convinced that their employer is offering the best package and they could likely move elsewhere for a seemingly more attractive program.

Leaders can see the writing on the wall with an understanding that the loss of Baby Boomers could be devastating to organizations, especially in industries like energy, for example, where the average age skews older. For many organizations, "training has increased, succession planning started, and flexible scheduling has been added at organizations that have started preparations for the retirement of Baby Boomers*"* (World at Work, 2010).

The challenge for HR leadership is twofold, including the pending retirement of the Baby Boomer generation and the need for the attraction and development of Millennials. Between 2005 and 2025 the number of people in U.S. ages 55–64 will grow by 11 million, whereas the number of people ages 25–54 will grow by only 5 million (U.S. Congressional Research Service). With the labor force participation historically declining at age 55, there is a critical challenge for organizations. According to the University of North Carolina (2011), 80 million Baby Boomers will exit the workforce in the next 20 years, with 8,000 Americans turning 65 each day. These numbers are daunting, considering what is at stake.

Companies are faced with the potential for worker shortages, talent/skill shortages, loss of critical business knowledge, loss of crucial technical skills, loss of key business relationships, lower economic growth and productivity (defined as workforce, capital employed, and change in productivity), making the market for 20- to 30-year-olds increasingly competitive. Peter Drucker (2001) summed up the crisis accurately, reflecting that "the confluence of a bulging aged population and a shrinking supply of youth is unlike anything that has happened since the dying centuries of the Roman Empire" (The Economist).

Distribution of Total US Population By Age



Source: US Census Bureau

Employers clearly recognize the issue, as 40 percent of organizations worry that the aging workforce will have a negative impact on their business, and only 14 percent of managers think they can cope with an aging workforce (The Economist, 2011). Despite this evidence, few organizations are ready to retain an older workforce, and instead recruit a younger one. Recognizing the problem is only the first step, and unfortunately HR leaders are failing to execute a solution successfully. Making talent a strategic priority requires alignment across three dimensions:

1. Identifying Changing Demographics

Demographics around the world are changing. Workers are becoming older. The percentage of Hispanics in the U.S. is rising. There are fewer workers in the western world, but an excess supply in developing nations.

2. Recognizing the Increasing Competition for Talent

Coming out of the great recession (which started in December 2007), talent is ready to move. Average tenure in the US is only five years among Millennials.

3. Understanding that Not All Employees Are Equal

The competition for skilled knowledge workers is especially intense; they produce three times the profit as other employees (McKinsey Quarterly, 2008). Qualified leaders are in short supply in China.

Companies are clearly facing a trade-off, as they must retain older workers while attracting younger ones. However, most organizations are currently offering a value proposition that satisfies neither end of the spectrum. Promoting a one-size fits all value proposition is the equivalent of offering lukewarm tea. Just because one person likes hot tea and one person likes iced-tea, offering lukewarm tea will not, of course, satisfy either preference. Statistically the solution makes sense, but practically nobody enjoys drinking lukewarm tea, just as offering a one-size fits all value proposition satisfies neither Millennials or Baby Boomers.

Organizations would be in trouble if all of their Baby Boomers decided to retire today, yet most companies have not explored ways to position value proposition to retain them. Millennials are expected to be next generation of leaders, yet the value propositions needed in order to attract and retain them often are often left out of business strategy discourse. Attribute importance differs by generational categories, so value propositions must be adapted in order to support these differences. This survey reinforces this. While many features are similar, there are some key differences between the preferences of the two groups, suggesting a need for flexibility and targeted messaging.

The table below outlines the different preferences for both Millennials and Baby Boomers, highlighting the importance nuances between the two groups. For Millennials, the most important attribute was base salary, followed by healthcare and time off. While these benefits were important for Baby Boomers, their priority was healthcare, followed by base salary and retirement. Awareness of valued benefits informs organizations and allows them to tailor retention strategies for Baby Boomers around healthcare and retirement benefits, while focusing attraction and retention strategies for Millennials around salary and paid time off. Results show that emphasizing traditional compensation and benefits programs will be impactful for Baby Boomers. Conversely, for Millennials, organizations should consider broadcasting the non-traditional, environmental and cultural aspects of the employee value proposition. The results below underscore this point that programs such as work-life balance, tuition reimbursement, development opportunities, student loan forgiveness, fitness allowance and hobby allowance are considerably more important to Millennials than to Baby Boomers.

Generational Differences for Attribute Importance	Millennials	Baby Boomers
Base Salary	13.7	16
Health Care Benefit	12.4	16.2
Paid Time Off	11.7	13.5
Retirement Contribution	10.8	14.8
Bonus Opportunity	8.5	9.1
Work Life Balance	8.3	6.9
Dental Coverage	7	8.9
Vision Coverage	5	5.1
Tuition Reimbursement (for You & Your Kids)	5	2
Development Opportunities	4.6	3.2
Student Loan Forgiveness	4.1	1.1
Car Allowance	2.6	1.2
Fitness Allowance	2.5	1.1
Hobby Allowance	1.9	0.5
Wine of the Month Club	1.7	0.4

These results are critical when addressing the aging workforce challenge. Understanding these employee preferences arms organizations with rich insights into the best ways to retain critical aging talent while also attracting the next generation of leadership. Conjoint analysis is an important tool in solving this issue in particular, but in reality, this is just one of the many employee challenges that can be addressed by leveraging insights from a conjoint study.

Implications and Recommendations

Unique Challenges in the HR Space

Conducting conjoint in the HR arena poses some interesting challenges relative to the typical Marketing study. Those challenges include, but are not limited to:

- *Global:* Many studies conducted in the HR space are conducted globally, posing much greater logistical challenges in addition to the need for translations.
- *Large Sampling Size:* It is not unusual to see a sample size of 20,000+ when conducting a TRO—significantly larger populations than a typical marketing sample.
- *Tie-Back Data:* HR surveys typically provide the ability to tie-back HRIS data files to survey results. This removes (or significantly reduces) the need for demographic questions to be included.
- *Paternalistic:* Often HR tends to be very protective of employees, taking an almost paternalistic role in controlling the messaging and exposure to employees.
- *Need for Education:* Buyers in the HR space require additional education due to the lack of familiarity with conjoint and the fact that data-driven decisions are less common in HR than other areas.
- *Emotional Response:* Because of the subject matter, employees are more likely to have stronger emotional responses to the questions. Therefore, questions and communication must be worded so that employee expectations are carefully managed.

Conclusion

As competition for human capital intensifies, it is increasingly important to understand the needs of employees. By leveraging conjoint techniques typically used for marketing, we are able to better attract, retain and engage employees in the HR space. This article offers the Four Cs (cost impact, competition, core people strategy and consumer preferences and needs) as key considerations in defining the employee value proposition.

Tim Glowa

Garry Spinks

Allyson Kuper

REFERENCES:

(2011). Age shall not wither them*:* Companies should start seeing older workers as assets rather than liabilities. *The Economist Print Edition.* Retrieved from: http://www.economist.com/node/18527063

(2012). SHRM-AARP poll shows organizations are concerned about boomer retirements and skills gaps. *The Society for Human Resource Management.* Retrieved from: http://www.shrm.org/about/pressroom/pressreleases/pages/shrmaarppressreleasepollretiringboomers.aspx

(2014). Few organizations' total rewards and business strategies fully align, according to Mercer survey. *Mercer.* Retrieved from: http://www.mercer.com/content/mercer/global/all/en/newsroom/few-reward-and-business-strategies-align-say-mercer-survey.html

Baum, M., & Kabst, R. (2013). Conjoint implications on job preferences: The moderating role of involvement. *The International Journal of Human Resource Management*, *24*(7), 1393–1417.

Brack, J. (2012). Maximizing Millennials in the workplace. *UNC Executive Development.* Retrieved from: http://www.kenan-flagler.unc.edu/executive-development/customprograms/~/media/DF1C11C056874DDA8097271A1ED48662.ashx

Drucker, P. (2001). The next society. *The Economist Special Report.* Retrieved from: http://www.economist.com/node/770819

Gan, C., Lee, J., & Soutar, G. (2009). Preferences for training options: A conjoint analysis. *Human Resource Development Quarterly*, *20*, 307–330.

Guthridge, M., Komm, A. B., & Lawson, E. (2008). Making talent a strategic priority, *McKinsey Quarterly*, *1,* 49–59. Retrieved from: http://www.americasdiversityleader.com/Downloads/McKinsey%20Report,%202008.pdf

Slade, L. A., Davenport, T. O., Roberts, D. R. and Shah, S. (2002). How Microsoft optimized its investment in people after the dot-com era. *Journal of Organizational Excellence, 22,* 43–52. doi: 10.1002/npr.10052

Slade, L.A. (2009). Minimizing promises and fears: Defining the decision space for conjoint research for employees versus customers. *Sawtooth Software Proceedings,* 51–58.

The Global Workforce Study. (2012). Drivers of attraction, retention and engagement chart. *Towers Watson.*

US Department of Labor. (2015). Economic news release: Employment situation summary. *Bureau of Labor Statistics.* Retrieved from: http://www.bls.gov/news.release/empsit.t04.htm

Vincent, G. K., and Velkoff, V.A. (2008). The next four decades: The older population in the United States: 2010 to 2050. *US Census Bureau*. Retrieved from: http://www.census.gov/prod/2010pubs/p25–1138.pdf

WorldatWork. (2006). WorldatWork Total Rewards Model. Retrieved from: http://www.worldatwork.org/pub/total_rewards_model.pdf

WorldatWork (2010). Beyond compensation: How employees prioritize total rewards at various life stages. Retrieved from: http://www.worldatwork.org/waw/adimLink?id=37007

WorldatWork. (2011). Bonus programs and practices. *WorldatWork: The Total Rewards Association.* Retrieved from: http://www.worldatwork.org/adimLink?id=50454

For Further Information

For additional information, please visit www.BugInsights.com or via Twitter @BugInsights. The authors can also be contacted as follows:

- Tim can be reached via e-mail at tim.glowa@buginsights.com or on Twitter @TimGlowa.
- Garry can be reached via e-mail at garry.spinks@buginsights.com or on Twitter @GarrySpinks.
- Allyson can be reached via e-mail at Allyson.kuper@buginsights.com or on Twitter @AllysonKuper.

MENU-BASED CHOICE: PROBIT AS AN ALTERNATIVE TO LOGIT?

CHRISTIAN NEUERBURG
GFK MARKETING & DATA SCIENCES

MOTIVATION

The choice situation in choice-based conjoint (CBC) experiments is fundamentally different from those in a menu-based choice experiment (MBC). In the traditional CBC-case the respondent's task is to pick one alternative out of a set of pre-defined alternatives—all alternatives presented in the choice tasks are substitutes per se. In the MBC-case the respondent is confronted with an experimental choice menu from which multiple alternatives can be selected. In that case, alternatives on the choice menu can be either substitutes or complements. This situation poses a challenge to the analyst responsible for selecting a suitable model.

Since the launch of *Sawtooth Software's* Menu-Based Choice software (MBC), logit-based estimation approaches (e.g., pooled MNL, latent class MNL, and especially HB-MNL) have become the workhorse for model building in the area of menu-based choice experiments. For a researcher who is familiar with specifying CBC-type Logit models it is easy to adapt that idea to the context of MBC. In addition, high performance estimation routines are available and the required estimation time is relatively low.

On the other hand there is the family of Probit models which seem to provide a natural alternative to Logit models in the context of MBC. Liechty, Ramaswamy and Cohen (2001) present in their seminal paper a multivariate probit formulation that can be applied in the context of MBC and showed promising results. Orme acknowledges in the MBC software documentation that "multivariate probit seems to provide a more theoretically complete model, directly incorporating the idea that items on the menu can be substitutes or complements" (Orme 2012, p. 6). Despite this, *Sawtooth Software* did not implement Probit in their standard software so far as they state to be "more familiar with logit analysis" and to "worry about the scalability of multivariate probit to the more complex types of menus" (Orme 2012, p. 6).

The overall impression is that Probit models seem to have merits but there is the need for further research in order to learn more about the properties of Probit in the context of MBC. In addition, a systematic comparison of Probit and Logit is required to answer the question if and under which conditions Probit is able to outperform Logit in the context of MBC.

PROBIT VS. LOGIT

Table 1 summarizes the most important similarities and differences between Logit and Probit in the context of MBC.

Both model types belong to the family of random utility models, assuming that the utility of an alternative consists of a deterministic part ($X\beta^T$) covering all observable aspects (e.g., price) and a random part representing all unobservable or omitted aspects (ε). The fundamental difference lies in the assumptions regarding the error term distribution. While the Logit model works on the assumption that all error terms follow a Gumbel distribution (independent and identically distributed) the Probit model assumes a multivariate normal distribution that allows for error term correlations between the different alternatives. If Logit models are applied in the context of MBC, the most widespread approach is to use separate multinomial or binary models for the different menu areas (e.g., burgers or fries) and to connect the area-specific models via selected cross-price effects (Orme 2010). A possible advantage of Probit is that there is no need to build separate models for different menu areas as all alternatives can be covered within a single overall model. The relationships between the different menu alternatives are represented by the error term correlations which are estimated along with the familiar beta estimates. Another important difference between both model types is the formulation of the likelihood function. Due to the error term assumptions, the Logit model has a closed-form expression for its likelihood. This is a very convenient property as it means that choice probabilities can be calculated without having to evaluate high-dimensional integrals. For the likelihood of the Probit model no closed-form expression exists, leading to further complications in estimation and the calculation of choice probabilities as it takes simulation-based approaches to evaluate the likelihood (Train 2003).

Table 1
Side-by-Side Comparison of Probit and Logit in the Context of MBC

	Multinomial Logit	**Multivariate Probit**
Utility Function	$U=X\beta^T+\varepsilon$	$U=X\beta^T+\varepsilon$
Error Terms	i.i.d. Gumbel	Multivariate normal (correlations allowed)
Model Structure	Separate models for each menu area	One single model
Interdependencies	Cross-price effects	Error term correlations
Likelihood	Closed-form expression	No closed-form expression

To illustrate the rationale behind the error term correlation matrix estimated for Probit models, Table 2 shows an example of a correlation matrix estimated in the context of an empirical MBC study conducted in the area of gaming consoles and accessories. In that particular case, the choice menu consisted of six binary menu areas (one console and two accessories for two brands each). Looking at the resulting error term correlations, one can see that within a certain brand family (e.g., Xbox) the console and the accessories are perceived as complements (positive correlation) and therefore frequently selected jointly. Between the two brands, the tested menu items are perceived as substitutes (negative correlations) and therefore likely to replace each other. Insights about which menu items can be considered substitutes or complements are certainly interesting for clients. But does this ability of the Probit model also give it an edge over Logit models in terms of predictive validity? Before reviewing the design of the comparative study it is necessary to introduce the type of Probit model analyzed in this piece of research.

Table 2
Posterior Means for Correlation Matrix of the Error Term

		Xbox ®			Playstation ®		
		Console	Camera	Wheel	Console	Camera	Wheel
Xbox ®	Console	1	0.30	0.50	-0.27	-0.32	-0.15
	Camera	0.30	1	0.23	-0.38	-0.18	-0.34
	Wheel	0.50	0.23	1	-0.19	-0.15	-0.06
Playstation ®	Console	-0.27	-0.38	-0.19	1	0.47	0.43
	Camera	-0.32	-0.18	-0.15	0.47	1	0.45
	Wheel	-0.15	-0.34	-0.06	0.43	0.45	1

The Multivariate Multinomial Probit Model (MVMNP)

In most real-world applications of MBC experiments there is a requirement to include a mixture of binary menu areas (where the decision maker can only decide to select or not to select a particular menu item) and multinomial menu areas (in which there is a choice between more than two mutually exclusive options). Therefore, the typical Multivariate Probit model will not be universally applicable as it is in general restricted to menus consisting of binary menu areas only and is therefore not used within this comparative study[1]. The Probit formulation used here is borrowed from the area of biostatistics and is called Multivariate Multinomial Probit model (MVMNP) (Zhang, Boscardin and Belin 2008). In the following section the basic assumptions of the MVMNP are described. Readers not interested in the technical details of the MVMNP estimation may skip this section which is based on Neuerburg and Koschate-Fischer (2015). In what follows we define

$i = 1, \ldots, n$	individuals;
$t = 1, \ldots, n.t$	tasks;
$q = 1, \ldots, g$	areas of the choice menu;
$j = 1, \ldots, pq$	alternatives (menu items) within a certain menu area.

In the context of a multivariate multinomial choice situation, an individual decision vector d_{it} $(1 \times g)$ is observed whose elements indicate the number of the selected menu item for each menu area and task:

[1] Liechty, Ramaswamy and Cohen (2001) present a Multivariate Probit formulation which can be applied to choice menus that include multinomial menu areas. Due to lack of documentation, it was not possible to replicate their approach within this study.

$$\mathbf{d}_{it}=\{d_{it1},\ldots,d_{itg}\}.$$

If respondent i in task t selects the first item in menu area 1 and the second item in menu area 2 the decision vector $\mathbf{d}_{it}=\{1,2\}$ is observed. When making decisions it is assumed that respondents maximize the latent utility of their selection within each menu area. The latent utility vector for a specific respondent and task is defined as:

$$\mathbf{z}_{it} = \mathbf{X}_{it} \cdot \boldsymbol{\beta}_i^T + \boldsymbol{\varepsilon}_{it}$$

where $\mathbf{z}_{it}$ is a vector containing the respective latent utilities for all modules qj and individual i in task t. To overcome the issue of additive redundancy which is present in all kinds of Probit formulations, the utility of one option within each menu area must be normalized to zero (Keane 1992, McCulloch and Rossi 1994). In our example, we define the "none" alternative within each menu area as the reference alternative. The error term follows a multivariate normal distribution, turning $\mathbf{z}_{it}$ into a multivariate Gaussian latent variable:

$$\boldsymbol{\varepsilon}_{it} \sim Normal(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\mathbf{0}$ is an n_{alt}-dimensional vector of zeros and $\boldsymbol{\Sigma}$ is the normalized covariance matrix of the error term of the dimensionality $n_{alt} \times n_{alt}$ (n_{alt} is the number of module alternatives across all functionalities). In the context of the MVMNP, a second identification issue exists: multiplicative redundancy. Zhang, Boscardin and Belin (2008) constrain the covariance matrix by normalizing g diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ to one to ensure identification. The following example illustrates the basic character of the normalized covariance matrix $\boldsymbol{\Sigma}$ and the nature of the MVMNP. For a choice menu consisting of two menu areas and three alternatives per area, two of which are menu items and one a "none" alternative, the following normalized covariance matrix results:

$$\mathbf{\hat{\imath}} = \begin{pmatrix} 1 & \sigma_{11\_12} & \sigma_{11\_21} & \sigma_{11\_22} \\ \cdot & \sigma_{12\_12} & \sigma_{12\_21} & \sigma_{12\_22} \\ \cdot & \cdot & 1 & \sigma_{21\_22} \\ \cdot & \cdot & \cdot & \sigma_{22\_22} \end{pmatrix}$$

The boxed areas describe the covariance structures of the error term for menu items of a certain menu area (as known from the Multinomial Probit). The other areas represent the covariance structures between items of different menu areas (as known from the Multivariate Probit). Zhang, Boscardin and Belin (2008) ensure in their proposed estimation procedure that this identification requirement is met.

Given the latent Gaussian structure of the model, individuals make their choices based on the following decision rule:

$$d_{itq} = \begin{cases} 0, & \text{if } \max\limits_{1 \le l \le p_q - 1} z_{itql} < 0 \\ j, & \text{if } \max\limits_{1 \le l \le p_q - 1} z_{itql} = z_{itqj} > 0 \end{cases}$$

With the utility of the "none" alternative for each menu area normalized to zero, a certain menu item is selected if it exhibits the highest latent utility of all menu items for the respective menu area and has a larger utility than the "none" option (which has an utility of zero in this example).

In order to run a HB-estimation of the MVMNP, a five-step Gibbs sampling procedure is necessary. The proposed estimation algorithm is based on work of Boscardin and Zhang (2004), Zhang, Boscardin and Belin (2006), and Zhang, Boscardin and Belin (2008). Several modifications had to be made to their original Gibbs sampler to allow for individual-level estimation. These modifications are based mainly on the work of McCulloch and Rossi (1994) and Zeithammer and Lenk (2006). The sampling procedure can be summarized as follows ("|~" means conditional on the other elements of the parameter space):

Step 1: Draw individual beta vectors | ~
Multivariate Normal Distribution

Step 2: Draw mean beta vector of the population distribution | ~
Multivariate Normal Distribution

Step 3: Draw covariance matrix of the population distribution | ~
Inverse Wishart Distribution

Step 4: Draw latent utilities of menu items | ~
Gibbs-within-Gibbs approach—truncated univariate Normal distributions

Step 5: Draw (normalized) covariance matrix of the error term | ~
Metropolis-Hastings-Step

Details on the estimation procedure are available from the author upon request. The described Gibbs sampler is more complex than for a traditional HB-MNL (which consists only of the first three steps in a slightly different implementation). The two additional steps (related to the latent utilities and the covariance matrix of the error terms) are time consuming iterative steps, which leads overall to significantly longer estimation times required for the MVMNP.

DESIGN OF THE MONTE CARLO STUDY

Table 3 summarizes the three models compared in this comparative study. The model of primary interest is the MVMNP as it represents the family of Probit models. For the MVMNP, only one integrated model covering all menu areas is estimated. As a benchmark, two Logit variants are added to the study design which are relevant from the perspective of a market research practitioner. The least complex version is the Independent MNL (IL). In that case, a separate MNL model is set up for each menu area. The deterministic part of the utility function consists only of alternative-specific constants and one linear price parameter. The area-specific models are entirely independent. As a more elaborate version of this "naïve" model formulation, the Serial Cross-Effects MNL (SCL) was added as a benchmark model[2]. This model connects the different area-specific models via selected cross-price effects (Orme 2010). Both tested logit variants do not incorporate error term correlations.

[2] For the selection of the cross-price effects to be incorporated into the model specification an aggregate Chi-Square test was used (Orme 2012).

Table 3
Characteristics of the Compared Models

	Independent MNL (IL)	Serial Cross-Effects MNL (SCL)	MVMNP
Model Type	Logit	Logit	Probit
Approach	Separate Models	Separate Models	Single Model
Beta Vectors	Individually (HB)	Individually (HB)	Individually (HB)
Cross-Effects	No	Yes	No
Error Term Correlations	No	No	Yes

For all three models respondent-level betas were estimated based on an HB-approach. All estimations were conducted in R 2.13.1 (32 Bit) using partly elements of the R-package *bayesm* with the computationally intensive parts outsourced to C. The Gibbs-Sampler used to estimate the HB-MNL formulations is equivalent to the algorithm implemented in *Sawtooth Software's CBC/HB* prior to version 5 (Train and Sonnier 2005). In order to utilize the available CPU-capacities in the best possible way, the number of iterations conducted for the different modeling alternatives were determined individually for each model type. Key factor for selecting the total number of iterations was the complexity of the respective Gibbs sampler, the expected level of autocorrelation and the average number of parameters that had to be estimated. Therefore, 200,000 iterations were conducted for the MVMNP, 100,000 iterations for SCL and 50,000 iterations for IL. 50% of the draws were discarded as burn-in iterations. 2,500 posterior draws were saved for each model and used later on for analysis. For all compared approaches the second-stage priors were set in a way to be non-informative.

To systematically analyze the performance of the three modeling approaches, a Monte Carlo experiment was designed (Neuerburg 2013). Monte Carlo experiments are based on synthetic responses and do not require "real" respondents. One of the advantages of this approach is that the analyzed datasets can be created under controlled conditions. In the present study, the artificial datasets were created based on a systematic variation of five simulation factors (see Table 4 for further details). The description of the simulation factors is based on Neuerburg and Koschate-Fischer (2015).

Table 4
Characteristics of the Synthetic Datasets

#	Simulation Factor	Level 1	Level 2	Level 3	Level 4
1	**Respondent Heterogeneity** ("Het")	Low	High		
2	**Menu Complexity** ("Complex")	C1 (10 areas / 2 items)	C2 (15 areas / 3 items)	C3 (10 areas / 6 items)	
3	**Sample Size** ("Sample")	100	250	500	
4	**Number of Tasks** ("Tasks")	5	10		
5	**Behavioral Model** ("Behavior")	Combinatorial MNL[3]	Serial Cross-Effects MNL	Independent MNL	MVMNP

Respondent heterogeneity describes the degree to which preference structures that build the foundation for decisions in the experimental menu differ between respondents in the sample. Heterogeneity is a relevant aspect from a marketing perspective because it gives rise to differentiated product offerings or segment-specific communication strategies. From a technical perspective, heterogeneity influences the variance of the population distribution in hierarchical models and therefore indirectly determines the Bayesian mixture of individual and population data. In this study two levels of respondent heterogeneity were included. If the level of heterogeneity is low it is assumed that all respondents of the respective sample belong to the same population in terms of their preference structure. In case the level of heterogeneity is high it is assumed that respondents belong to a more fragmented market consisting of three different preference segments.

The *menu complexity* of an experimental choice menu is determined by the number of available menu areas and options within each menu area. It can be summarized by the number of possible configuration patterns that can be created from the experimental choice menu. The complexity of the experimental choice menu can influence the performance of the different models because it determines the number of parameters to be estimated. In the current study, three levels of complexity have been incorporated: C1, which includes 10 areas with each 1 item alternative and 1 "none" alternative (1024 possible configuration patterns); C2, which includes

[3] The "Combinatorial MNL" resembles an approach presented by Ben-Akiva and Gershenfeld (1998). The results for this model are not reported here.

15 areas with each 2 item alternatives and 1 "none" alternative (~14 million possible configuration patterns); and C3, which includes 10 areas with each 5 item alternatives and 1 "none" alternative (~60 million possible configuration patterns).

The factor *sample size* indicates the number of respondents available for model calibration and is one of the most important aspects in the design of a menu-based choice experiment because of the potentially enormous cost implications associated with it. Therefore, the researcher strives to know in advance what implications different sample sizes may have for the performance of the modeling approaches used. The available sample size is important from a technical perspective because it can influence the implicit mixture of individual- and population-based information in the context of hierarchical Bayesian modeling. In the current research design, three different levels of sample sizes are surveyed: n = 100 represents a relatively small sample size, such as in the area of health care research or personal interviews in the context of car clinics; n = 250 represents a medium-sized sample, such as in the area of personal interviews; and n = 500 represents a medium-sized web-based sample.

The *number of tasks* represents the number of individual repetitions of the experiment to be completed by each respondent. As with the available sample size, the number of tasks determines the data available for model estimation. The number of tasks also has potentially relevant implications for the cost of the menu-based choice experiment because the number of available repetitions increases the required time for the experiment and, thus, the average length of the interview. The factor "tasks" potentially influences the implicit weighting of individual- and population-based information in the context of hierarchical Bayesian model estimation. In the current simulation study, 5 and 10 tasks per respondent are tested.

To generate synthetic data sets, it is necessary to make explicit assumptions on the *behavioral model* applied by the respondents. Because the models (implicitly) applied in reality are unknown, it seems reasonable to use the models compared in the study also as behavioral assumptions. An advantage of this approach is that none of the compared models are systematically preferred. In addition, based on this approach the robustness of the different modeling alternatives against a "wrong" underlying model can be tested.

Based on an exhaustive enumeration of all simulation factors, 144 datasets were created. In order to increase the generalizability, the simulation experiment was replicated once. Therefore, 288 synthetic datasets were available for analysis and have been evaluated for all three models of interest[4]. For details on the data generation process please refer to Neuerburg (2013).

Evaluation of Predictive Validity

In order to evaluate the predictive validity of the different approaches, responses for five in-sample holdout tasks were created (the holdout responses were not used for model calibration). The degree to which the different models are able to predict the holdout responses delivers a measure of the internal validity of the compared approaches.

Three different performance measures were used:

[4] All computations were conducted on a Windows HPC 2008 Compute Cluster (operated by University of Erlangen-Nuremberg). At time of the experiment, 16 computer nodes with 2 Hexacore AMD Opteron Istanbul processors were available (2.6 GHz, 32 GB RAM). Depending on the utilization of the cluster, the computations could use up to 192 cores in parallel. The estimations were conducted between January and July 2012.

1. **Mean Absolute Error (MAE)**
   Mean absolute error between the observed and predicted choice shares across all items and holdout tasks—the lower the better.
2. **Combinatorial Hit Rate**
   Percentage of correctly predicted choice patterns ("combinations") across all respondents and holdout tasks—the higher the better.
3. **Item Hit Rate**
   Average share of correctly predicted item choices across all respondents and holdout tasks—the higher the better.

While combinatorial and item hit rate measure the model's ability to predict individual decisions, the mean absolute error is a measure of the aggregate precision that can be reached by using the different models. Note that point estimates were used to make predictions for the two MNL models. For the MVMNP a simulation-based approach (based on the stored posterior draws) is necessary as the complex likelihood function cannot be evaluated directly.

Results

Table 5 summarizes the overall results for the three compared models and all performance measures. In addition, the average estimation time for the different approaches is reported (the average estimation time for IL was normalized to 1).

The overall results clearly show that the MVMNP performs worst for all performance measures. The results of SCL and IL are very similar with the SCL having a slight edge over IL. Looking at these overall results, one would conclude that using the MVMNP cannot be recommended—especially when taking the average required estimation times into account.

The reported estimation times have been transformed so that the average estimation time for IL (the least complex model) is normalized to 1. As for the two logit models IL and SCL separate logit models are estimated for the different menu areas, the reported estimation times are summed up over all required partial models. The average estimation time required for the MVMNP is roughly 4 times higher than the (cumulative) estimation time for IL (and still roughly 3 times higher than for the more complex SCL). Taking into account the possibility of running several MNL estimations in parallel on different cores and to use faster estimation routines like implemented in *Sawtooth Software's CBC/HB* or the R-package *ChoiceModelR*, the resulting differences in estimation time between the logit approaches and the MVMNP will be even more dramatic.

Table 5
Overall Results

	Independent MNL (IL)	Serial Cross-Effects MNL (SCL)	Multivariate Multinomial Probit (MVMNP)
Mean Absolute Error	5.29	5.09 ★	6.71 ⬇
Combinatorial Hit Rate	10.37%	10.77% ★	7.81% ⬇
Item Hit Rate	59.49%	59.62% ★	54.59% ⬇
Estimation Time[5] (IL=1)	1.00 ★	1.43	4.24 ⬇

★ = best; ⬇= worst; 288 observations available for each model

As the reported overall results are averages over 288 datasets with different characteristics, some particular data conditions might exist under which the MVMNP is able to outperform the logit approaches. Therefore, Tables 6–8 summarize the average results for the three performance measures split by data condition.

[5] All estimations have been conducted in an identical system environment which allows a direct comparison. For modeling approaches IL and SCL, for which separate models are estimated for each functional area, the reported estimation times represent the sum of CPU-time that is needed to estimate all required model components. While interpreting these figures, one has to take into account that for each modeling approach a typical number of iterations was defined and therefore the resulting CPU-times also reflect these differences.

Table 6
Mean Absolute Error (by Data Condition)

		IL	SCL	MVMNP
Het	high	4.16 ★	4.26	5.46 ⬇
	low	6.43	5.91 ★	7.96 ⬇
Complex	C1	6.16	5.58 ★	6.38 ⬇
	C2	5.64	5.41 ★	7.56 ⬇
	C3	4.08 ★	4.27	6.19 ⬇
Sample	100	6.21	6.19 ★	7.47 ⬇
	250	5.13	4.86 ★	6.59 ⬇
	500	4.54	4.21 ★	6.06 ⬇
Tasks	5	5.45	5.32 ★	6.82 ⬇
	10	5.14	4.85 ★	6.59 ⬇
Behavior	IL	1.87 ★	2.09	4.11 ⬇
	SCL	4.15	2.53 ★	7.85 ⬇
	MVMNP	12.39	12.52 ⬇	11.48 ★

★ = best; ⬇= worst[6]

The key take-aways from the detailed analysis of the MAE results are:

- MVMNP exhibits the worst results under almost all data conditions.
- Under most conditions SCL exhibits the lowest MAE.
- IL has a lower MAE than SCL for very complex menus (higher probability of erroneous selection of cross-effects).
- All models benefit from a larger sample size and a larger number of tasks.
- Given that cross-effects are present, erroneously using IL will lead to an increase of MAE (4.15 vs. 2.53).

[6] Observations per data condition: Het: n=144/ Complex: n=96/ Sample: n=96/ Tasks: n=144/ Behavior: n=72.

Table 7
Combinatorial Hit Rate [%] (by Data Condition)

		IL	SCL	MVMNP
Het	high	8.93	8.96 ★	4.81 ⬇
	low	11.82	12.59 ★	10.82 ⬇
Complex	C1	16.62	17.25 ★	16.57 ⬇
	C2	7.91	8.32 ★	3.02 ⬇
	C3	6.58	6.75 ★	3.85 ⬇
Sample	100	10.27	10.63 ★	7.74 ⬇
	250	10.30	10.70 ★	7.86 ⬇
	500	10.55	10.99 ★	7.84 ⬇
Tasks	5	10.28	10.67 ★	7.85 ⬇
	10	10.46	10.87 ★	7.77 ⬇
Behavior	IL	24.23 ★	24.14	18.10 ⬇
	SCL	16.50	18.21 ★	10.68 ⬇
	MVMNP	0.15	0.14 ⬇	1.92 ★

★ = best; ⬇= worst[7]

The key take-aways from the detailed analysis of the Combinatorial Hit Rates are:

- MVMNP exhibits worst results for almost all data conditions.
- For small menus MVMNP is head to head to SCL and IL.
- Menu complexity is a main driver for combinatorial hit rates.
- MVMNP suffers most from an increase in menu complexity.
- Combinatorial hit rates are insensitive to sample size and number of tasks.

[7] Observations per data condition: Het: n=144/ Complex: n=96/ Sample: n=96/ Tasks: n=144/ Behavior: n=72.

Table 8
Item Hit Rate [%] (by Data Condition)

		IL	SCL	MVMNP
Het	high	51.69	51.65 ★	42.65 ↓
	low	67.29	67.58 ★	66.52 ↓
Complex	C1	72.76	72.99 ★	72.51 ↓
	C2	60.17	60.37 ★	52.82 ↓
	C3	45.55	45.49 ★	38.44 ↓
Sample	100	59.34	59.37 ★	54.36 ↓
	250	59.50	59.66 ★	54.59 ↓
	500	59.65	59.82 ★	54.81 ↓
Tasks	5	59.30	59.39 ★	54.25 ↓
	10	59.69	59.85 ★	54.93 ↓
Behavior	IL	77.78 ★	77.66	68.69 ↓
	SCL	74.91	75.94 ★	61.55 ↓
	MVMNP	40.74	40.67 ↓	44.18 ★

★ = best; ↓= worst[8]

The key take-aways from the detailed analysis of the Item Hit Rate are:

- MVMNP exhibits worst results for almost all data conditions.
- MVMNP shows strongest decrease in item hit rates when menu complexity increases.
- Menu complexity is a main driver of item hit rates.
- For low levels of menu complexity all three models perform similarly.
- Item hit rates are insensitive to sample size and number of tasks.

Summary and Recommendations

Although Probit models allow the researcher to identify substitutes and complements by analyzing the estimated correlations of the error terms, this ability does not offer Probit (in the tested MVMNP formulation) an advantage over the tested Logit-based approaches in terms of predictive validity (also not for predicting combinatorial patterns).

[8] Observations per data condition: Het: n=144/ Complex: n=96/ Sample: n=96/ Tasks: n=144/ Behavior: n=72.

Compared to Logit models, Probit models are less parsimonious and their estimation requires a more complex Gibbs-Sampler which results in enormous estimation times. Therefore, Logit-based approaches are by far better scalable to choice menus of larger complexity which makes them the preferred alternative for commercial applications.

Based on this research, some general recommendations for the use of Logit models in the context of MBC can be derived:

1. **Valid aggregate predictions require sufficient sample sizes**
   IL and SCL seem to benefit from larger sample sizes as far as the mean absolute error of the holdout prediction is concerned. Individual hit rates are relatively unaffected by an increase in sample size, which indicates that in the context of the tested number of tasks (5, 10) sufficient individual information is available to predict individual choices.
2. **The number of tasks can be kept relatively low**[9]
   An increase of the number of tasks leads to no substantial improvement of the predictive validity of IL and SCL. Therefore, one can conclude that MBC models can be estimated at a sufficient level of quality using a relatively low number of tasks (e.g., five individual repetitions).
3. **When using logit, always check for cross-price effects**
   Not incorporating cross-price effects when actually present will negatively affect predictive validity. The effect of erroneously incorporating cross-price effects when not present in the underlying dataset has a lower potential to harm predictive validity as the estimated effects tend to be small anyway.
4. **Try to keep menu complexity as low as possible**
   The results clearly show that complexity of the choice menus is a key driver for model performance. While hit rates for the prediction of single item choices (Item Hit Rate) are quite acceptable throughout all levels of complexity, hit rates for combinatorial patterns (Combinatorial Hit Rate) deteriorate as complexity increases.

Christian Neuerburg

[9] Cautionary note: the required number of tasks should always be determined based on design tests. Models with very complex specifications may require a higher number of tasks to deliver precise individual predictions.

References

Ben-Akiva, Moshe and Shari Gershenfeld (1998), "Multi-featured Products and Services: Analysing Pricing and Bundling Strategies," *Journal of Forecasting*, 17, 175–196.

Boscardin, W. J. and Xiao Zhang (2004), "Modeling the Covariance and Correlation Matrix of Repeated Measures," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley series in probability and statistics*, Andrew Gelman and Xiao-Li Meng, eds. Chichester: Wiley, 215–226.

Keane, Michael P. (1992), "A note on identification in the multinomial probit model," *Journal of Business & Economic Statistics*, 10 (2), 193–200.

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*, 38, 183–196.

McCulloch, Robert and Peter E. Rossi (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64 (1–2), 207–240.

Neuerburg, Christian (2013), Modellierung von Wahlverhalten in modularen Auswahlsituationen. Ein simulationsbasierter Vergleich verschiedener Modellvarianten unter Berücksichtigung der Zahlungsbereitschaft. Nuremberg: GfK-Verein.

Neuerburg, Christian and Nicole Koschate-Fischer (2015), "Menu-Based Choice Models: A Comparison of Reservation Price Recoverability, Model Fit and Predictive Validity under Varying Data Conditions." Working Paper University Erlangen-Nuremberg.

Orme, Bryan K. (2010), "Menu-Based Choice Modeling Using Traditional Tools," in *Proceedings of the Sawtooth Software Conference 2010*, Sawtooth Software Inc, ed. Sequim, WA, 37–57.

——— (2012), "Menu-Based Choice (MBC) for Multi-Check Choice Experiments," (accessed May 1, 2015), [available at http://www.sawtoothsoftware.com/download/mbcbooklet.pdf].

Train, Kenneth (2003), *Discrete Choice Methods with Simulation.* Cambridge: Cambridge Univ. Press.

Train, Kenneth and Garrett Sonnier (2005), "Mixed Logit with Bounded Distributions of Correlated Partworths," in *Applications of Simulation Methods in Environmental and Resource Economics*. *The Economics of Non-Market Goods and Resources*, Vol. 6, Riccardo Scarpa and Anna Alberini, eds. Dordrecht: Springer, 117–134.

Zeithammer, Robert and Peter Lenk (2006), "Bayesian Estimation of Multivariate-Normal Models When Dimensions are Absent," *Quantitative Marketing & Economics*, 4 (3), 241–265.

Zhang, Xiao, W. J. Boscardin and Thomas R. Belin (2006), "Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables," *Journal of Computational and Graphical Statistics*, 15 (4), 880–896.

——— (2008), "Bayesian Analysis of Multivariate Nominal Measures Using Multivariate Multinomial Probit Models," *Computational Statistics & Data Analysis*, 52 (7), 3697–3708.

COMBINING LATENT-CLASS CHOICE, CART AND CBC/HB TO IDENTIFY SIGNIFICANT COVARIATES IN MODEL ESTIMATION

GEORGE BOOMER
STATWIZARDS LLC
KILEY AUSTIN-YOUNG
COMCAST CORP.

ABSTRACT

Covariates are often important, for example, gender in the handbag market, income in the exotic car market, age in the market for geriatric medicine. How then, can we identify key covariates and incorporate them into a CBC simulation within a time frame that comports with practitioners' schedules?

Our approach makes use of three techniques applied to a common data set. First, CBC/HB is employed to produce a set of individual-level utilities. Second, a latent-class choice (LGC) estimation identifies groups of respondents who share a common set of utilities. Third, CART is used to improve upon LGC's covariate classification. Finally, the latent classes and significant covariates from modern data mining techniques are brought together in a common market simulator. We use both a simulated data set and a disguised, real-world example from the telecommunications industry to illustrate this approach.

This paper is not an attempt to use the covariates in the CBC/HB upper model, nor is it a direct comparison of the above methods. Rather, it is an attempt to show an alternative approach for identifying significant covariates in a choice-modeling exercise.

1. THE PROBLEM

Hierarchical Bayes (CBC/HB) combines individual-level estimates of utility with excellent fit of holdout samples and allows covariates to be entered in the upper-level model.

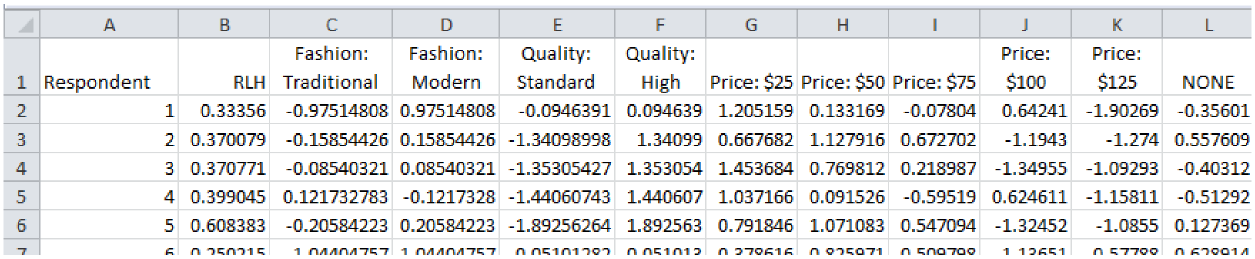
To illustrate the problem, we borrowed a hypothetical data set for a fictional shoe market from Statistical Innovations, Inc. The product in this data set contains three attributes—fashion, quality and price—and four covariates—age, gender, eye color and hair color. Each variable has a number of levels, as shown in the following table:

Figure 1. List of Attributes and Covariates

Attributes			Covariates			
Fashion	Quality	Price	Sex	Age	Eye color	Hair color
Traditional	Standard	$25	Male	Under 25	Blue	Blond
Modern	High	$50	Female	25-39	Brown	Brown
		$75		40+	Black	Black
		$100			Green	Red
		$125			Hazel	
					Gray	
					Other	

Using an experimental design, we estimated a CBC/HB model using the attributes on the left. The result was a standard set of individual-level utilities which we read into Excel.

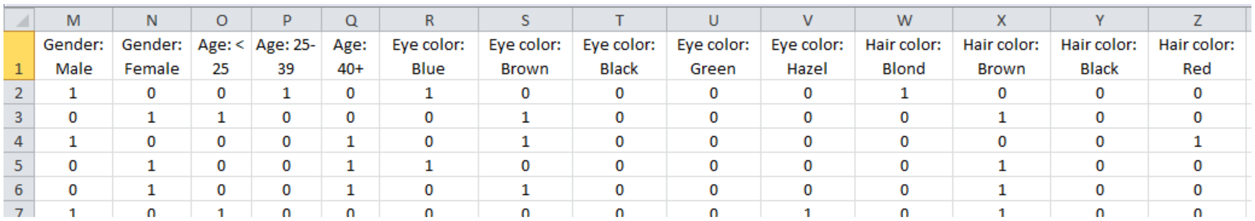
Figure 2. CBC/HB Utilities File Imported to Excel



	A	B	C	D	E	F	G	H	I	J	K	L
1	Respondent	RLH	Fashion: Traditional	Fashion: Modern	Quality: Standard	Quality: High	Price: $25	Price: $50	Price: $75	Price: $100	Price: $125	NONE
2	1	0.33356	-0.97514808	0.97514808	-0.0946391	0.094639	1.205159	0.133169	-0.07804	0.64241	-1.90269	-0.35601
3	2	0.370079	-0.15854426	0.15854426	-1.34098998	1.34099	0.667682	1.127916	0.672702	-1.1943	-1.274	0.557609
4	3	0.370771	-0.08540321	0.08540321	-1.35305427	1.353054	1.453684	0.769812	0.218987	-1.34955	-1.09293	-0.40312
5	4	0.399045	0.121732783	-0.1217328	-1.44060743	1.440607	1.037166	0.091526	-0.59519	0.624611	-1.15811	-0.51292
6	5	0.608383	-0.20584223	0.20584223	-1.89256264	1.892563	0.791846	1.071083	0.547094	-1.32452	-1.0855	0.127369
7	6	0.250215	1.04404757	1.04404757	0.05101282	0.051013	0.378616	0.825971	0.509798	1.13651	0.57788	0.628914

To this file we appended a separate worksheet containing for each respondent dummy-coded covariates.

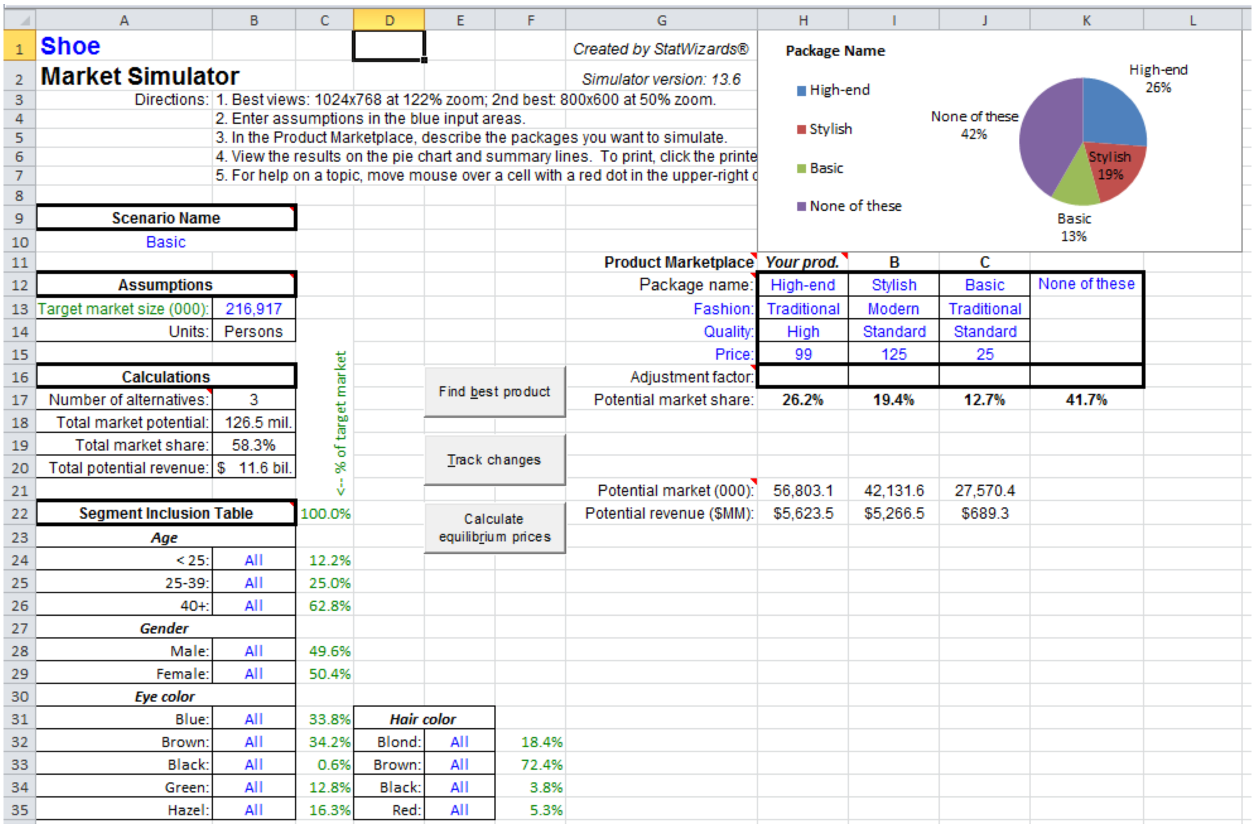
Figure 3. Covariates Appended to Utilities File



	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Gender: Male	Gender: Female	Age: < 25	Age: 25-39	Age: 40+	Eye color: Blue	Eye color: Brown	Eye color: Black	Eye color: Green	Eye color: Hazel	Hair color: Blond	Hair color: Brown	Hair color: Black	Hair color: Red
2	1	0	0	1	0	1	0	0	0	0	1	0	0	0
3	0	1	1	0	0	0	1	0	0	0	0	1	0	0
4	1	0	0	0	1	0	1	0	0	0	0	0	0	1
5	0	1	0	0	1	1	0	0	0	0	0	1	0	0
6	0	1	0	0	1	0	1	0	0	0	0	1	0	0
7	1	0	1	0	0	0	0	0	0	1	0	1	0	0

Using this combined file, we built a simulator. Like many simulators, this one supports analysis by subgroups. Starting with an arbitrary scenario in Figure 4,

Figure 4. Excel-Based Simulator with All Covariates



Shoe
Market Simulator

Created by StatWizards®
Simulator version: 13.6

Directions: 1. Best views: 1024x768 at 122% zoom; 2nd best: 800x600 at 50% zoom.
2. Enter assumptions in the blue input areas.
3. In the Product Marketplace, describe the packages you want to simulate.
4. View the results on the pie chart and summary lines. To print, click the printe
5. For help on a topic, move mouse over a cell with a red dot in the upper-right c

Scenario Name
Basic

Assumptions	
Target market size (000):	216,917
Units:	Persons

Calculations	
Number of alternatives:	3
Total market potential:	126.5 mil.
Total market share:	58.3%
Total potential revenue:	$ 11.6 bil.

<-- % of target market

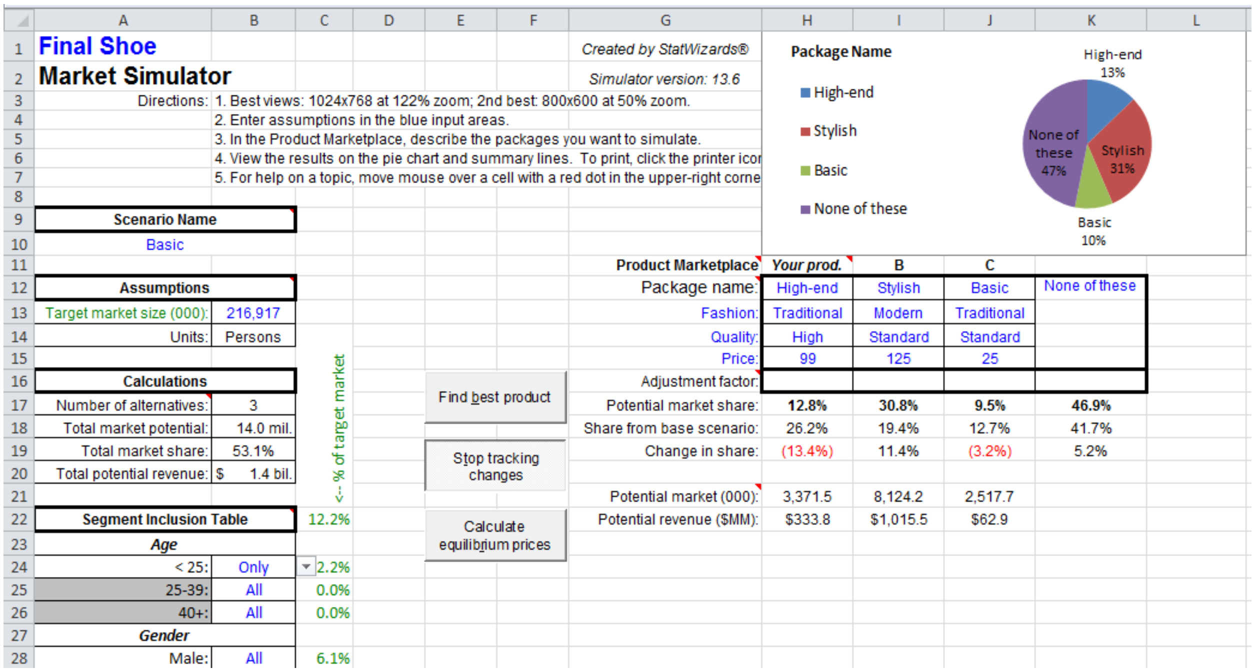
Segment Inclusion Table		100.0%
Age		
< 25:	All	12.2%
25-39:	All	25.0%
40+:	All	62.8%
Gender		
Male:	All	49.6%
Female:	All	50.4%
Eye color		
Blue:	All	33.8%
Brown:	All	34.2%
Black:	All	0.6%
Green:	All	12.8%
Hazel:	All	16.3%

Hair color		
Blond:	All	18.4%
Brown:	All	72.4%
Black:	All	3.8%
Red:	All	5.3%

Product Marketplace	*Your prod.*	B	C	
Package name:	High-end	Stylish	Basic	None of these
Fashion:	Traditional	Modern	Traditional	
Quality:	High	Standard	Standard	
Price:	99	125	25	
Adjustment factor:				
Potential market share:	**26.2%**	**19.4%**	**12.7%**	**41.7%**
Potential market (000):	56,803.1	42,131.6	27,570.4	
Potential revenue ($MM):	$5,623.5	$5,266.5	$689.3	

we took a snapshot of this scenario and filtered the data to show only respondents under 25 years old.

Figure 5. Filtering on Respondents Age 25 and Under



Final Shoe
Market Simulator

Created by StatWizards®
Simulator version: 13.6

Directions: 1. Best views: 1024x768 at 122% zoom; 2nd best: 800x600 at 50% zoom.
2. Enter assumptions in the blue input areas.
3. In the Product Marketplace, describe the packages you want to simulate.
4. View the results on the pie chart and summary lines. To print, click the printer icor
5. For help on a topic, move mouse over a cell with a red dot in the upper-right corne

Scenario Name
Basic

Assumptions	
Target market size (000):	216,917
Units:	Persons

Calculations	
Number of alternatives:	3
Total market potential:	14.0 mil.
Total market share:	53.1%
Total potential revenue:	$ 1.4 bil.

<-- % of target market

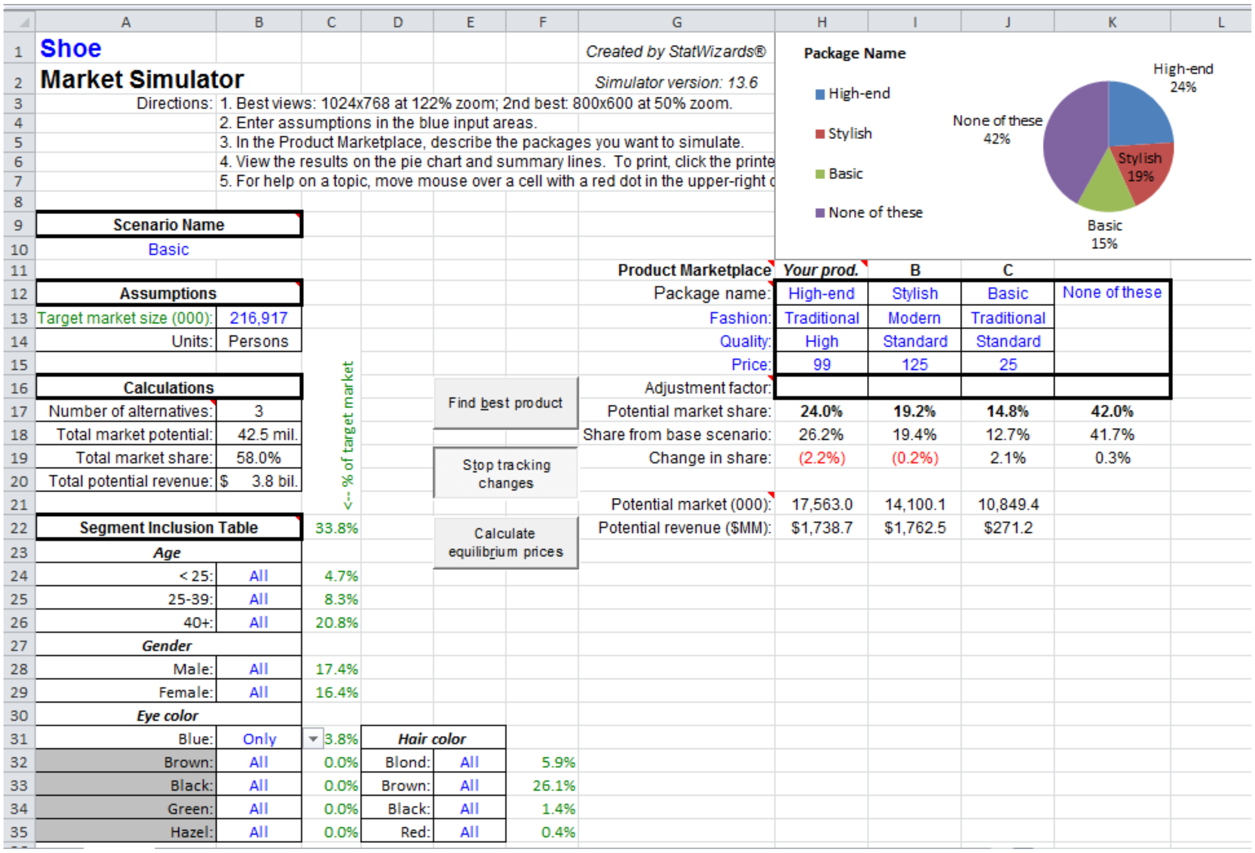
Segment Inclusion Table		12.2%
Age		
< 25:	Only	2.2%
25-39:	All	0.0%
40+:	All	0.0%
Gender		
Male:	All	6.1%

Product Marketplace	*Your prod.*	B	C	
Package name:	High-end	Stylish	Basic	None of these
Fashion:	Traditional	Modern	Traditional	
Quality:	High	Standard	Standard	
Price:	99	125	25	
Adjustment factor:				
Potential market share:	**12.8%**	**30.8%**	**9.5%**	**46.9%**
Share from base scenario:	26.2%	19.4%	12.7%	41.7%
Change in share:	(13.4%)	11.4%	(3.2%)	5.2%
Potential market (000):	3,371.5	8,124.2	2,517.7	
Potential revenue ($MM):	$333.8	$1,015.5	$62.9	

Cell I19 shows that young respondents show a greater preference for Stylish shoes, product B, by an amount more than 11 percentage points greater than the base scenario.

Resetting the filter, we now select people with blue eyes by changing the value in cell B31.

Figure 6. Filtering on Respondents with Blue Eyes



	A	B	C	D	E	F	G	H	I	J	K	L
1	Shoe						Created by StatWizards®					
2	Market Simulator						Simulator version: 13.6					
3	Directions:	1. Best views: 1024x768 at 122% zoom; 2nd best: 800x600 at 50% zoom.										
4		2. Enter assumptions in the blue input areas.										
5		3. In the Product Marketplace, describe the packages you want to simulate.										
6		4. View the results on the pie chart and summary lines. To print, click the printe										
7		5. For help on a topic, move mouse over a cell with a red dot in the upper-right c										
8												
9	Scenario Name											
10	Basic											
11							Product Marketplace	Your prod.	B	C		
12	Assumptions						Package name:	High-end	Stylish	Basic	None of these	
13	Target market size (000):	216,917					Fashion:	Traditional	Modern	Traditional		
14	Units:	Persons					Quality:	High	Standard	Standard		
15			<-- % of target market				Price:	99	125	25		
16	Calculations						Adjustment factor:					
17	Number of alternatives:	3			Find best product		Potential market share:	24.0%	19.2%	14.8%	42.0%	
18	Total market potential:	42.5 mil.					Share from base scenario:	26.2%	19.4%	12.7%	41.7%	
19	Total market share:	58.0%			Stop tracking changes		Change in share:	(2.2%)	(0.2%)	2.1%	0.3%	
20	Total potential revenue:	$ 3.8 bil.										
21							Potential market (000):	17,563.0	14,100.1	10,849.4		
22	Segment Inclusion Table		33.8%		Calculate equilibrium prices		Potential revenue ($MM):	$1,738.7	$1,762.5	$271.2		
23	Age											
24	< 25:	All	4.7%									
25	25-39:	All	8.3%									
26	40+:	All	20.8%									
27	Gender											
28	Male:	All	17.4%									
29	Female:	All	16.4%									
30	Eye color											
31	Blue:	Only	3.8%	Hair color								
32	Brown:	All	0.0%	Blond:	All	5.9%						
33	Black:	All	0.0%	Brown:	All	26.1%						
34	Green:	All	0.0%	Black:	All	1.4%						
35	Hazel:	All	0.0%	Red:	All	0.4%						

We see that blue-eyed people show a slightly increased (2.1 percentage points) preference for product C, a basic shoe.

The question is whether either of these scenarios reveals a significant difference in preferences between filtered and unfiltered groups. While using CBCHB, we want to know which covariates show significantly different preferences for products in this market. How can we combine the two objectives?

2. A Solution

We propose the use of alternate methodologies that permit significance tests for covariates. Here's how the process works:

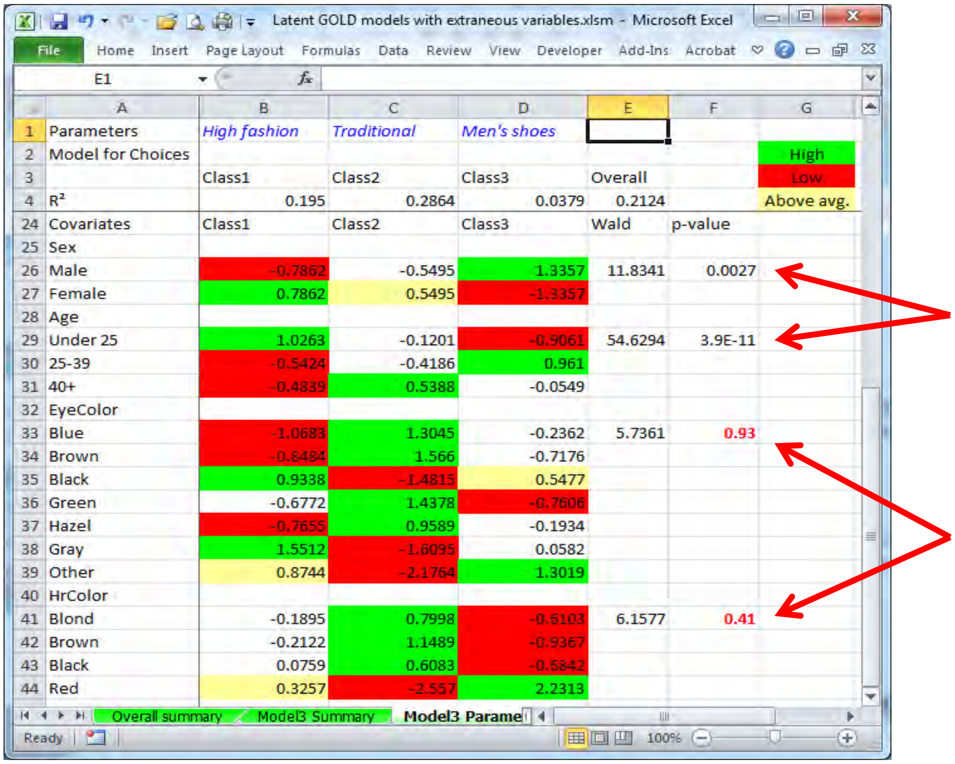
1. Estimate a model using the same dataset but an alternate methodology such as Latent GOLD® Choice (LGC).
2. Include all subgroups as covariates.
3. Perform significance tests on covariates.
4. Eliminate insignificant covariates.
5. Append significant covariates to the CBC-HB utilities file.

6. Optional, if LGC is used: Append segments from the LGC model.

Because results from CBC/HB are not affected, there is no confounding from using two techniques. We are just identifying which covariates to append to CBC/HB utilities.

Running an LGC model using the same data set and all covariates, we obtain the following results for a three-segment model:

Figure 7. Latent GOLD Choice Estimation with All Covariates



	A	B	C	D	E	F	G
1	Parameters	*High fashion*	*Traditional*	*Men's shoes*			
2	Model for Choices						High
3		Class1	Class2	Class3	Overall		Low
4	R^2	0.195	0.2864	0.0379	0.2124		Above avg.
24	Covariates	Class1	Class2	Class3	Wald	p-value	
25	Sex						
26	Male	-0.7862	-0.5495	1.3357	11.8341	0.0027	
27	Female	0.7862	0.5495	-1.3357			
28	Age						
29	Under 25	1.0263	-0.1201	-0.9061	54.6294	3.9E-11	
30	25-39	-0.5424	-0.4186	0.961			
31	40+	-0.4839	0.5388	-0.0549			
32	EyeColor						
33	Blue	-1.0683	1.3045	-0.2362	5.7361	0.93	
34	Brown	-0.8484	1.566	-0.7176			
35	Black	0.9338	-1.4815	0.5477			
36	Green	-0.6772	1.4378	-0.7606			
37	Hazel	-0.7655	0.9589	-0.1934			
38	Gray	1.5512	-1.6095	0.0582			
39	Other	0.8744	-2.1764	1.3019			
40	HrColor						
41	Blond	-0.1895	0.7998	-0.6103	6.1577	0.41	
42	Brown	-0.2122	1.1489	-0.9367			
43	Black	0.0759	0.6083	-0.5842			
44	Red	0.3257	-2.557	2.2313			

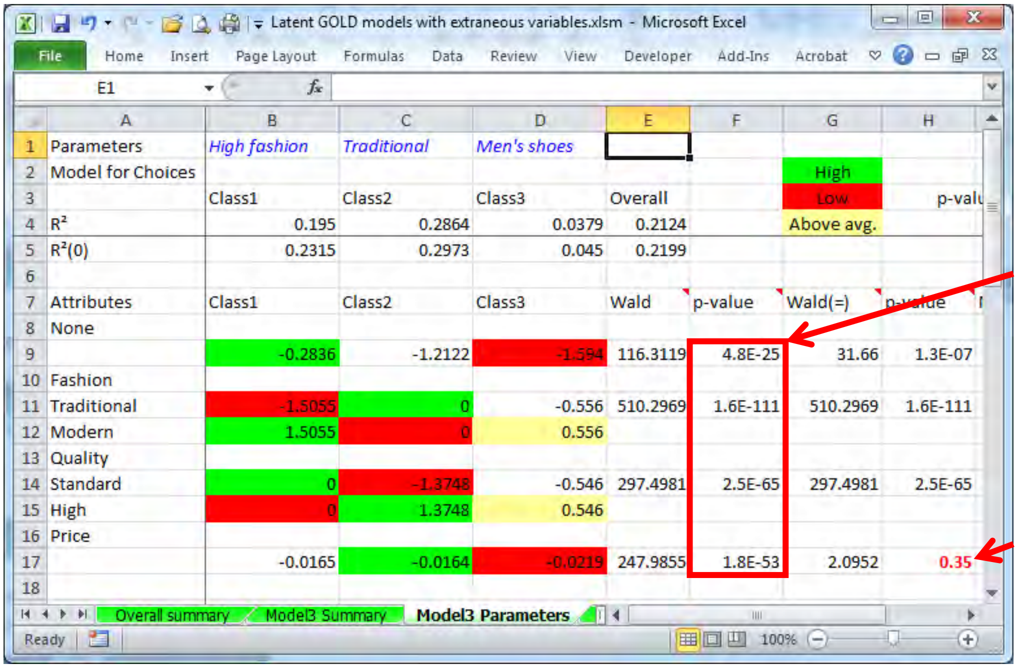
Sex and Age pass significance tests.

Eye color and hair color do not.

As the p-values in column F reveal, sex and age pass Wald tests of significance, whereas eye color and hair color fail. The presumption of independence of exogenous variables is not violated, as Latent GOLD employs covariates in separate logit models to predict membership in segments.

Turning for a moment to the model's attributes, we find that all of these coefficients pass Wald tests for significance,

Figure 8. Wald Tests on Attributes



Latent GOLD models with extraneous variables.xlsm - Microsoft Excel

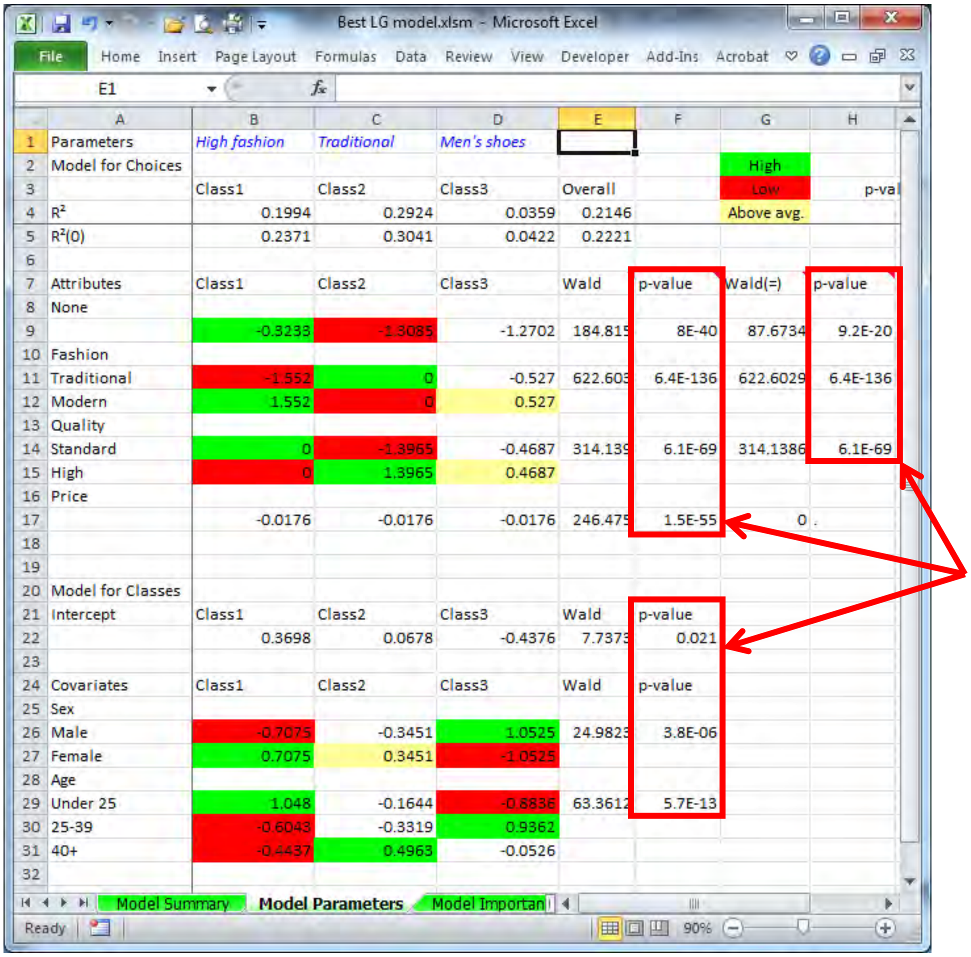
	A	B	C	D	E	F	G	H
1	Parameters	*High fashion*	*Traditional*	*Men's shoes*				
2	Model for Choices						High	
3		Class1	Class2	Class3	Overall		Low	p-valu
4	R^2	0.195	0.2864	0.0379	0.2124		Above avg.	
5	$R^2(0)$	0.2315	0.2973	0.045	0.2199			
6								
7	Attributes	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value
8	None							
9		-0.2836	-1.2122	-1.594	116.3119	4.8E-25	31.66	1.3E-07
10	Fashion							
11	Traditional	-1.5055	0	-0.556	510.2969	1.6E-111	510.2969	1.6E-111
12	Modern	1.5055	0	0.556				
13	Quality							
14	Standard	0	-1.3748	-0.546	297.4981	2.5E-65	297.4981	2.5E-65
15	High	0	1.3748	0.546				
16	Price							
17		-0.0165	-0.0164	-0.0219	247.9855	1.8E-53	2.0952	0.35
18								

All attributes are significant,

but no significant class differences exist among price coefficients

but the hypothesis that price coefficients are equal across segments cannot be rejected. We revise our model to eliminate insignificant covariates and impose class independence on the price coefficient. In the revised model, all significance tests pass.

Figure 9. LGC Model with Revised Specification

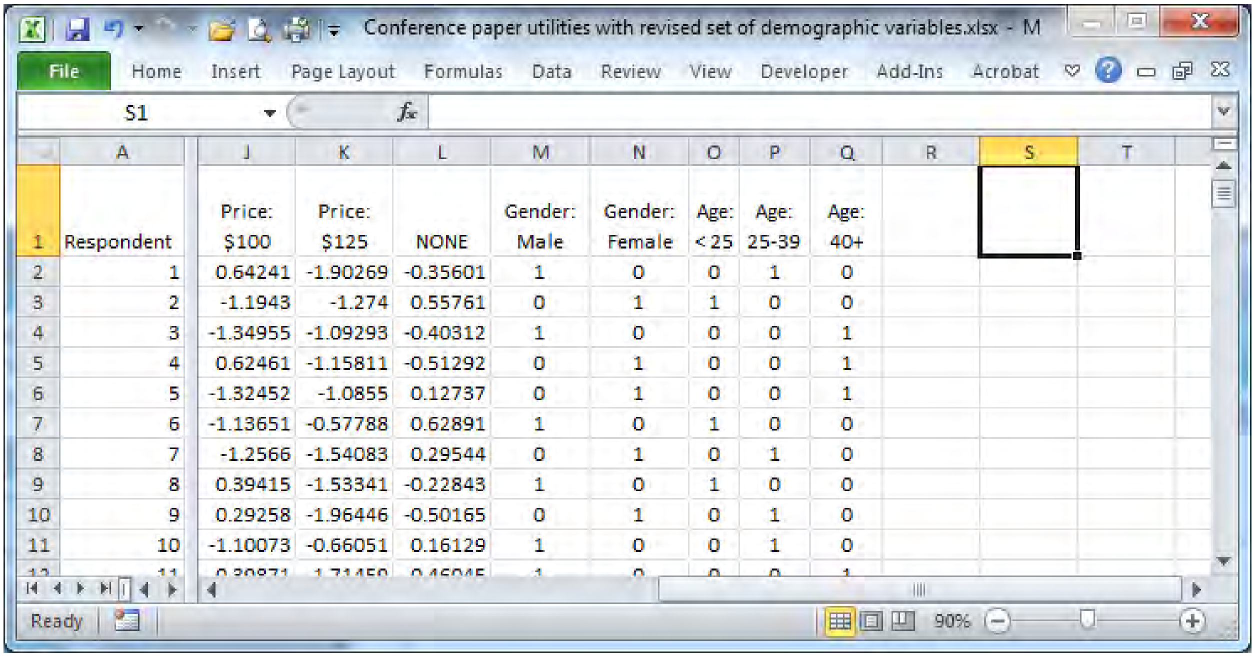


	A	B	C	D	E	F	G	H
1	Parameters	*High fashion*	*Traditional*	*Men's shoes*				
2	Model for Choices						High	
3		Class1	Class2	Class3	Overall		Low	p-val
4	R^2	0.1994	0.2924	0.0359	0.2146		Above avg.	
5	R^2(0)	0.2371	0.3041	0.0422	0.2221			
6								
7	Attributes	Class1	Class2	Class3	Wald	p-value	Wald(=)	p-value
8	None							
9		-0.3233	-1.3085	-1.2702	184.815	8E-40	87.6734	9.2E-20
10	Fashion							
11	Traditional	-1.552	0	-0.527	622.603	6.4E-136	622.6029	6.4E-136
12	Modern	1.552	0	0.527				
13	Quality							
14	Standard	0	-1.3965	-0.4687	314.139	6.1E-69	314.1386	6.1E-69
15	High	0	1.3965	0.4687				
16	Price							
17		-0.0176	-0.0176	-0.0176	246.475	1.5E-55	0	.
18								
19								
20	Model for Classes							
21	Intercept	Class1	Class2	Class3	Wald	p-value		
22		0.3698	0.0678	-0.4376	7.7373	0.021		
23								
24	Covariates	Class1	Class2	Class3	Wald	p-value		
25	Sex							
26	Male	-0.7075	-0.3451	1.0525	24.9823	3.8E-06		
27	Female	0.7075	0.3451	-1.0525				
28	Age							
29	Under 25	1.048	-0.1644	-0.8836	63.3612	5.7E-13		
30	25-39	-0.6043	-0.3319	0.9362				
31	40+	-0.4437	0.4963	-0.0526				
32								

All significance tests pass.

To the spreadsheet containing our original CBC-HB utilities, we append the significant covariates only.

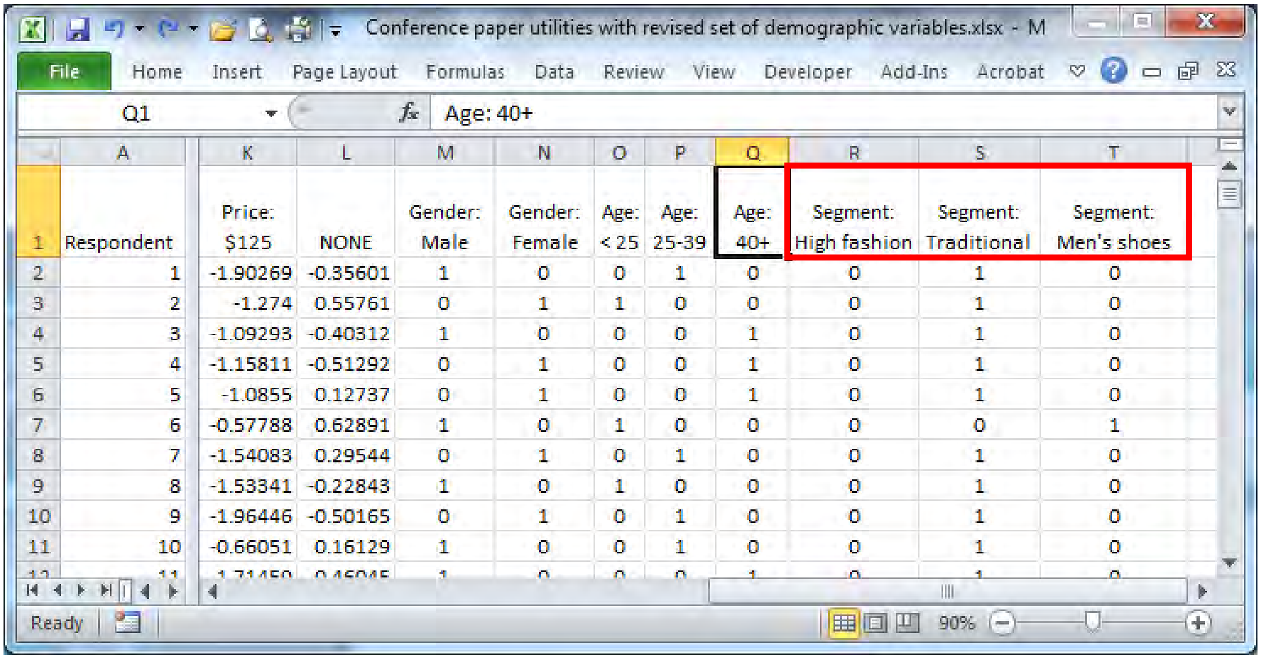
Figure 10. CBC/HB Utilities File with Significant Covariates Appended



	A	J	K	L	M	N	O	P	Q	R	S	T
1	Respondent	Price: $100	Price: $125	NONE	Gender: Male	Gender: Female	Age: < 25	Age: 25-39	Age: 40+			
2	1	0.64241	-1.90269	-0.35601	1	0	0	1	0			
3	2	-1.1943	-1.274	0.55761	0	1	1	0	0			
4	3	-1.34955	-1.09293	-0.40312	1	0	0	0	1			
5	4	0.62461	-1.15811	-0.51292	0	1	0	0	1			
6	5	-1.32452	-1.0855	0.12737	0	1	0	0	1			
7	6	-1.13651	-0.57788	0.62891	1	0	1	0	0			
8	7	-1.2566	-1.54083	0.29544	0	1	0	1	0			
9	8	0.39415	-1.53341	-0.22843	1	0	1	0	0			
10	9	0.29258	-1.96446	-0.50165	0	1	0	1	0			
11	10	-1.10073	-0.66051	0.16129	1	0	0	1	0			

Employing LGC as an alternative methodology yielded a bonus: classification of respondents into segments. As an option, we can append these classifications to the same file.

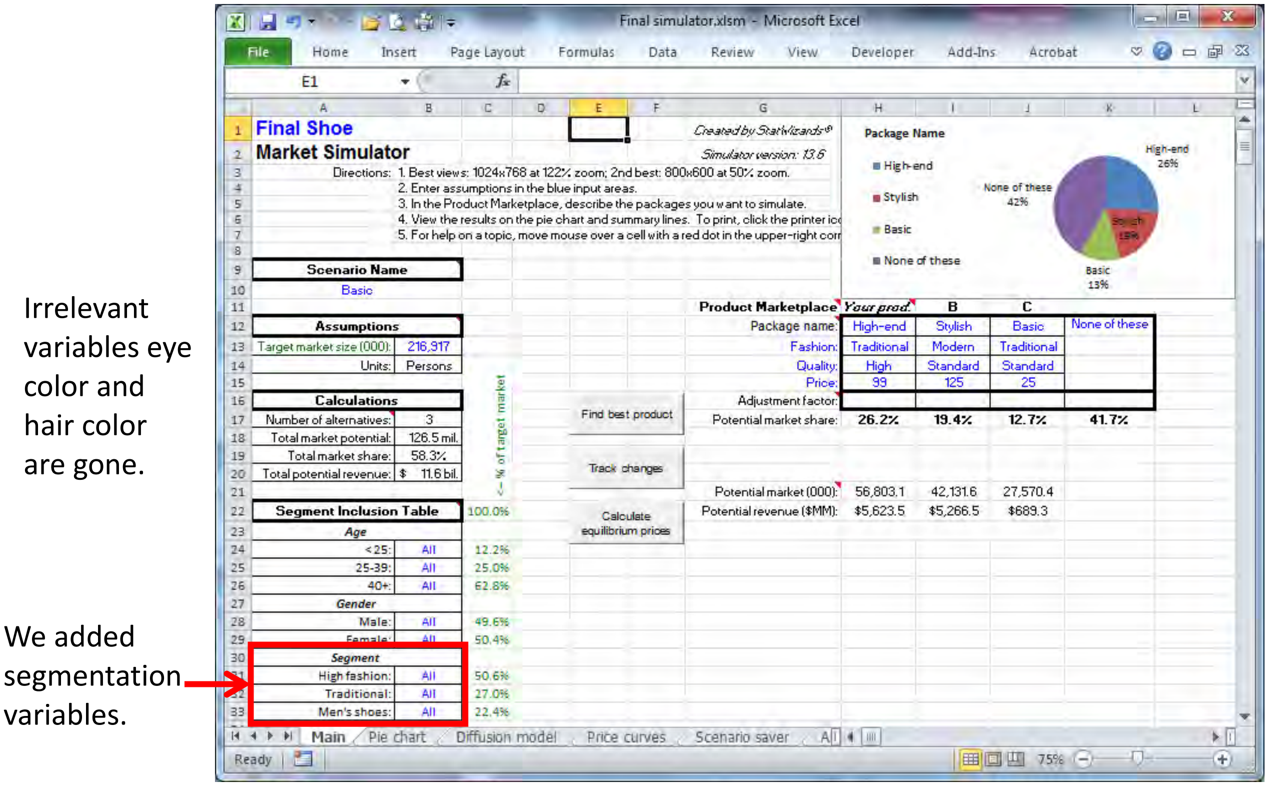
Figure 11. CBC/HB Utilities File with Latent Classes Appended



	A	K	L	M	N	O	P	Q	R	S	T
1	Respondent	Price: $125	NONE	Gender: Male	Gender: Female	Age: < 25	Age: 25-39	Age: 40+	Segment: High fashion	Segment: Traditional	Segment: Men's shoes
2	1	-1.90269	-0.35601	1	0	0	1	0	0	1	0
3	2	-1.274	0.55761	0	1	1	0	0	0	1	0
4	3	-1.09293	-0.40312	1	0	0	0	1	0	1	0
5	4	-1.15811	-0.51292	0	1	0	0	1	0	1	0
6	5	-1.0855	0.12737	0	1	0	0	1	0	1	0
7	6	-0.57788	0.62891	1	0	1	0	0	0	0	1
8	7	-1.54083	0.29544	0	1	0	1	0	0	1	0
9	8	-1.53341	-0.22843	1	0	1	0	0	0	1	0
10	9	-1.96446	-0.50165	0	1	0	1	0	0	1	0
11	10	-0.66051	0.16129	1	0	0	1	0	0	1	0

We can now use this file to build a new simulator, this time including only covariates that matter along with the option to filter on latent-class segments.

Figure 12. Revised Simulator



Irrelevant variables eye color and hair color are gone.

We added segmentation variables.

Final Shoe
Market Simulator

Created by StatWizards®
Simulator version: 13.6

Directions: 1. Best views: 1024x768 at 122% zoom; 2nd best: 800x600 at 50% zoom.
2. Enter assumptions in the blue input areas.
3. In the Product Marketplace, describe the packages you want to simulate.
4. View the results on the pie chart and summary lines. To print, click the printer ic
5. For help on a topic, move mouse over a cell with a red dot in the upper-right cor

Scenario Name
Basic

Assumptions	
Target market size (000):	216,917
Units:	Persons

Calculations	
Number of alternatives:	3
Total market potential:	126.5 mil.
Total market share:	58.3%
Total potential revenue:	$ 11.6 bil.

Segment Inclusion Table		100.0%
Age		
< 25:	All	12.2%
25-39:	All	25.0%
40+:	All	62.8%
Gender		
Male:	All	49.6%
Female:	All	50.4%
Segment		
High fashion:	All	50.6%
Traditional:	All	27.0%
Men's shoes:	All	22.4%

Product Marketplace	*Your prod.*	B	C	
Package name:	High-end	Stylish	Basic	None of these
Fashion:	Traditional	Modern	Traditional	
Quality:	High	Standard	Standard	
Price:	99	125	25	
Adjustment factor:				
Potential market share:	**26.2%**	**19.4%**	**12.7%**	**41.7%**
Potential market (000):	56,803.1	42,131.6	27,570.4	
Potential revenue ($MM):	$5,623.5	$5,266.5	$689.3	

3. Alternate Methodologies

You don't have to use LGC as an alternate methodology; any methodology that can identify significant covariates in the source data will do. Because it is closely related to hierarchical Bayes[1], mixed logit serves this purpose well and would be our second choice. Software packages for estimating mixed-logit models are available from a number of sources, including

- Limdep's NLOGIT[2]
- R library mlogit[3]
- Michel Bielaire's Biogeme[4]

Unlike the second and third programs in this list, NLOGIT produces individual-level utilities, much like CBC/HB.

Tree-based methods such as SI-CHAID[5], CART, random forests and stochastic gradient boosting[6] can also be used to identify covariates. All of these methods begin with a dependent

[1] Train, Kenneth (2001). "A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit," Department of Economics, University of California, Berkeley.
[2] Available for license from Econometric Software at http://www.limdep.com/.
[3] Can be downloaded from http://cran.r-project.org/web/packages/mlogit/index.html.
[4] Free download available at http://biogeme.epfl.ch/.
[5] SI-CHAID is a product from Statistical Innovations, Inc. For more information, see http://statisticalinnovations.com/products/sichaid.html.

variable and a set of independent variables. The dependent variable usually consists of discrete categories. Approaches employed by these methods vary, but all tree methods identify cutpoints within key independent variables and use them to create splits, such that the grouping of the data after splitting becomes more concentrated around one of the dependent variable's discrete categories. The process continues with additional branches (and in some cases additional trees) being built until the final nodes are as concentrated as possible.

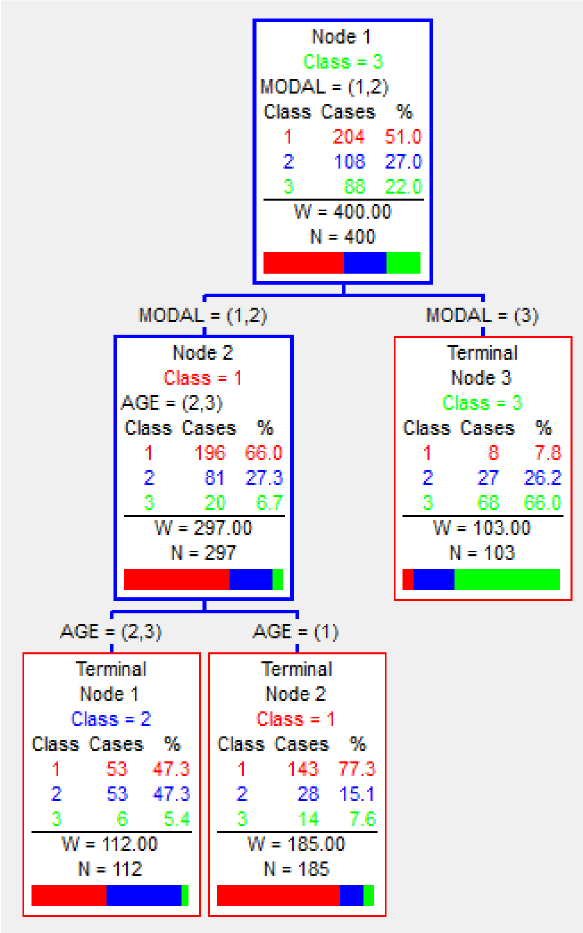
In the course of building trees, each method identifies the most important independent variables that contribute to splits. We can identify these variables and append them to the CBC/HB utilities file.

4. Another Use for Tree Methods

If you choose LGC as your alternate methodology, you have the option of appending latent-class (i.e., segment) memberships to your CBC/HB utilities file along with chosen covariates. If you do this, you can employ the same tree methods described above to assign out-of-sample subjects to latent classes. Returning to our example, here is such a tree built by CART.

[6] Salford Systems, provides software for all three tree methods. The company uses the trademark TreeNet® for its stochastic gradient boosting software. For more information, see http://www.salford-systems.com/products.

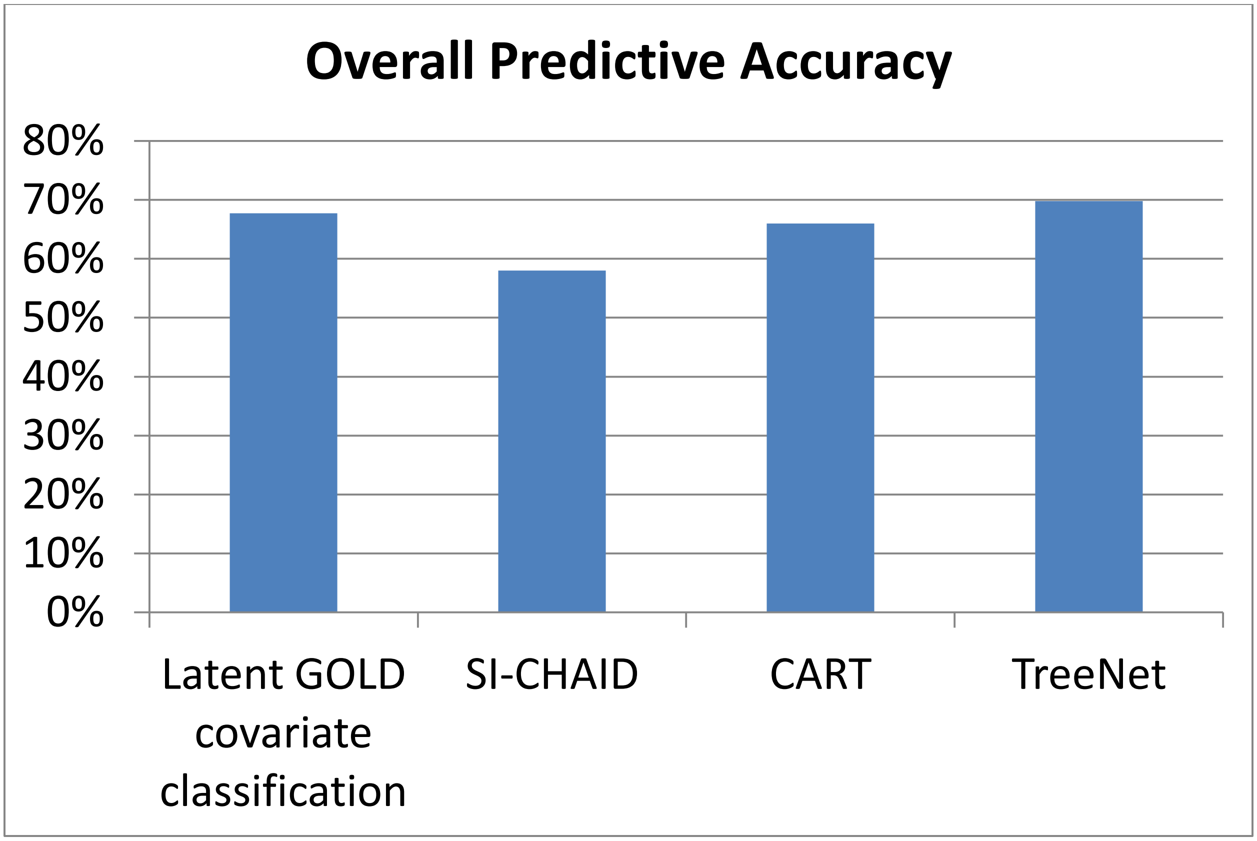
Figure 13. CART Tree Used to Predict Segment Membership



Covering this chart in detail lies beyond the scope of this paper, but suffice it to say that classes 1, 2 and 3 correspond to segments 1, 2 and 3 in Figure 7.

With all of these tree-based methods, a question arises about which to use. The answer depends partly on the pricing and availability of software and partly on predictive accuracy. Regarding the latter criterion, we applied various tree-based methods to our LGC model with the following results:

Figure 14. Comparison of Predictions Using Selected Tree-Based Methods



In this example, TreeNet scored best, followed closely by Latent GOLD Choice's internal covariate classification algorithm, though the differences between LG, CART and TreeNet are not statistically significant[7]. This finding is not surprising given TreeNet's consistent performance in a number of data mining competitions[8].

Our approach appears to work well on this synthetic data set, but how does it work in practice?

5. Comcast Example

Comcast's core services comprise a portfolio of voice, video, and data packages. The company's product pricing, packaging, and planning team was asked to consider different approaches to product pricing as well as the lineup of the package components and package constructs.

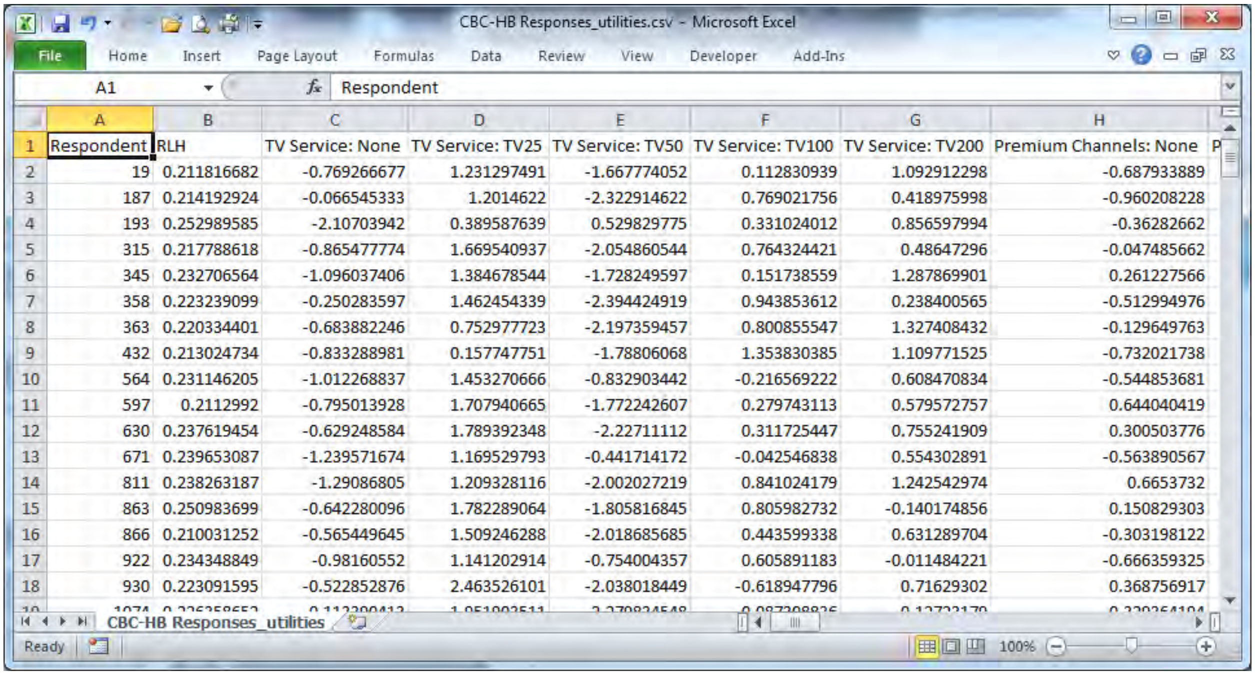
As part of this effort, Comcast employed a multi-product choice model to assess the value of different cable channels based on customer viewing habits and the potential upside from introducing new packaging portfolio approaches in market—for example, video-inclusive multi-product packages focused less on traditional product levers such as TV channels and more on a full suite of potential services.

[7] Using a proportions test at a 95% confidence interval.

[8] As one example, a TreeNet model placed first in the Duke/NCR Teradata Center for CRM Competition held in 2003.

Packages were anchored by services other than video, such as HSD, and supplemented by a rich set of emerging and differentiated services, including home security/control and IP-based solutions such as storage, were tested in order to frame an actionable recommendation for content packaging efforts. The choice model included the following attributes and covariates:

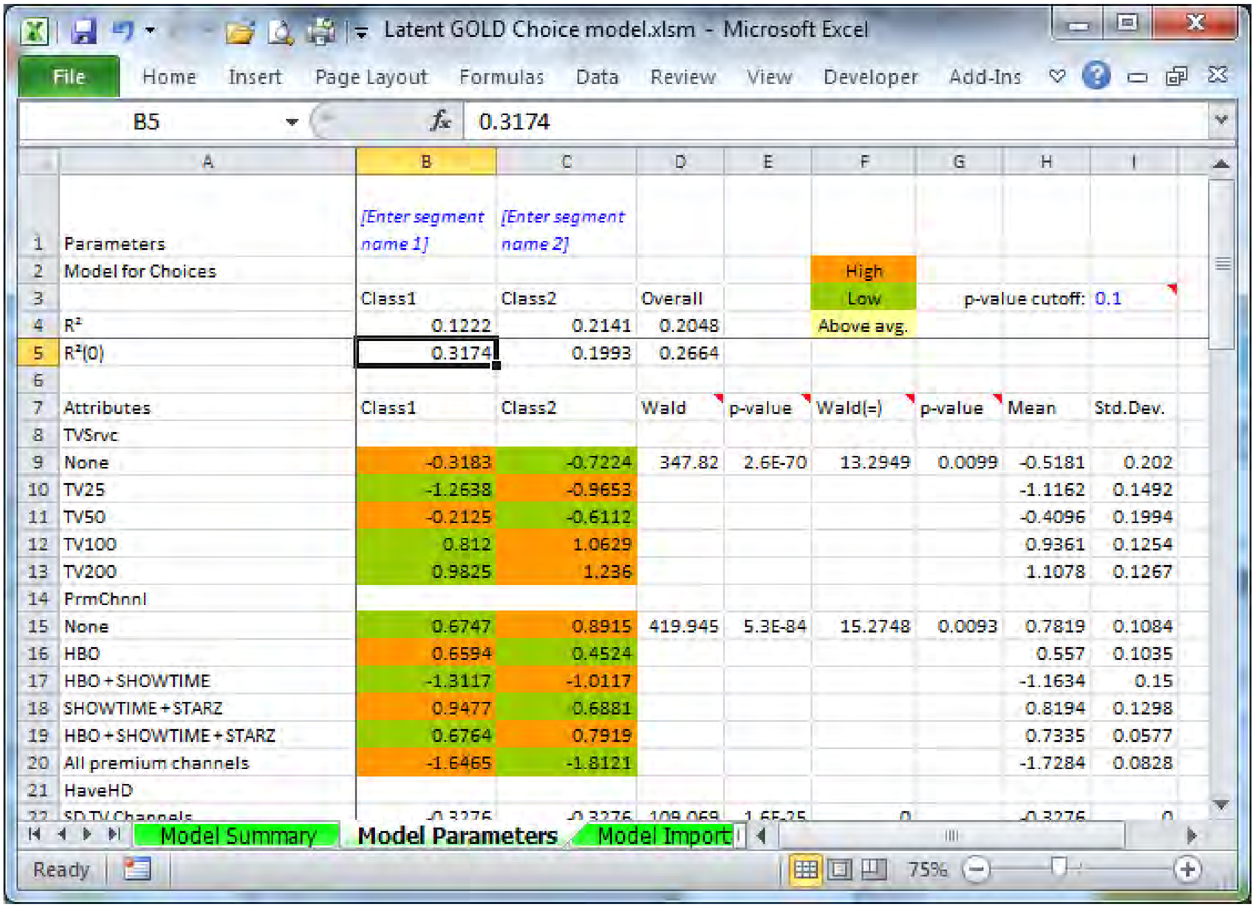
Figure 15. CBC/HB Utilities from Comcast Model



	A	B	C	D	E	F	G	H
1	Respondent	RLH	TV Service: None	TV Service: TV25	TV Service: TV50	TV Service: TV100	TV Service: TV200	Premium Channels: None
2	19	0.211816682	-0.769266677	1.231297491	-1.667774052	0.112830939	1.092912298	-0.687933889
3	187	0.214192924	-0.066545333	1.2014622	-2.322914622	0.769021756	0.418975998	-0.960208228
4	193	0.252989585	-2.10703942	0.389587639	0.529829775	0.331024012	0.856597994	-0.36282662
5	315	0.217788618	-0.865477774	1.669540937	-2.054860544	0.764324421	0.48647296	-0.047485662
6	345	0.232706564	-1.096037406	1.384678544	-1.728249597	0.151738559	1.287869901	0.261227566
7	358	0.223239099	-0.250283597	1.462454339	-2.394424919	0.943853612	0.238400565	-0.512994976
8	363	0.220334401	-0.683882246	0.752977723	-2.197359457	0.800855547	1.327408432	-0.129649763
9	432	0.213024734	-0.833288981	0.157747751	-1.78806068	1.353830385	1.109771525	-0.732021738
10	564	0.231146205	-1.012268837	1.453270666	-0.832903442	-0.216569222	0.608470834	-0.544853681
11	597	0.2112992	-0.795013928	1.707940665	-1.772242607	0.279743113	0.579572757	0.644040419
12	630	0.237619454	-0.629248584	1.789392348	-2.22711112	0.311725447	0.755241909	0.300503776
13	671	0.239653087	-1.239571674	1.169529793	-0.441714172	-0.042546838	0.554302891	-0.563890567
14	811	0.238263187	-1.29086805	1.209328116	-2.002027219	0.841024179	1.242542974	0.6653732
15	863	0.250983699	-0.642280096	1.782289064	-1.805816845	0.805982732	-0.140174856	0.150829303
16	866	0.210031252	-0.565449645	1.509246288	-2.018685685	0.443599338	0.631289704	-0.303198122
17	922	0.234348849	-0.98160552	1.141202914	-0.754004357	0.605891183	-0.011484221	-0.666359325
18	930	0.223091595	-0.522852876	2.463526101	-2.038018449	-0.618947796	0.71629302	0.368756917
19	1074	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

With this set of attributes we estimated an HB model and generated a CBC/HB utilities file. Next, using the same data set but incorporating covariates, we estimated an LGC model.

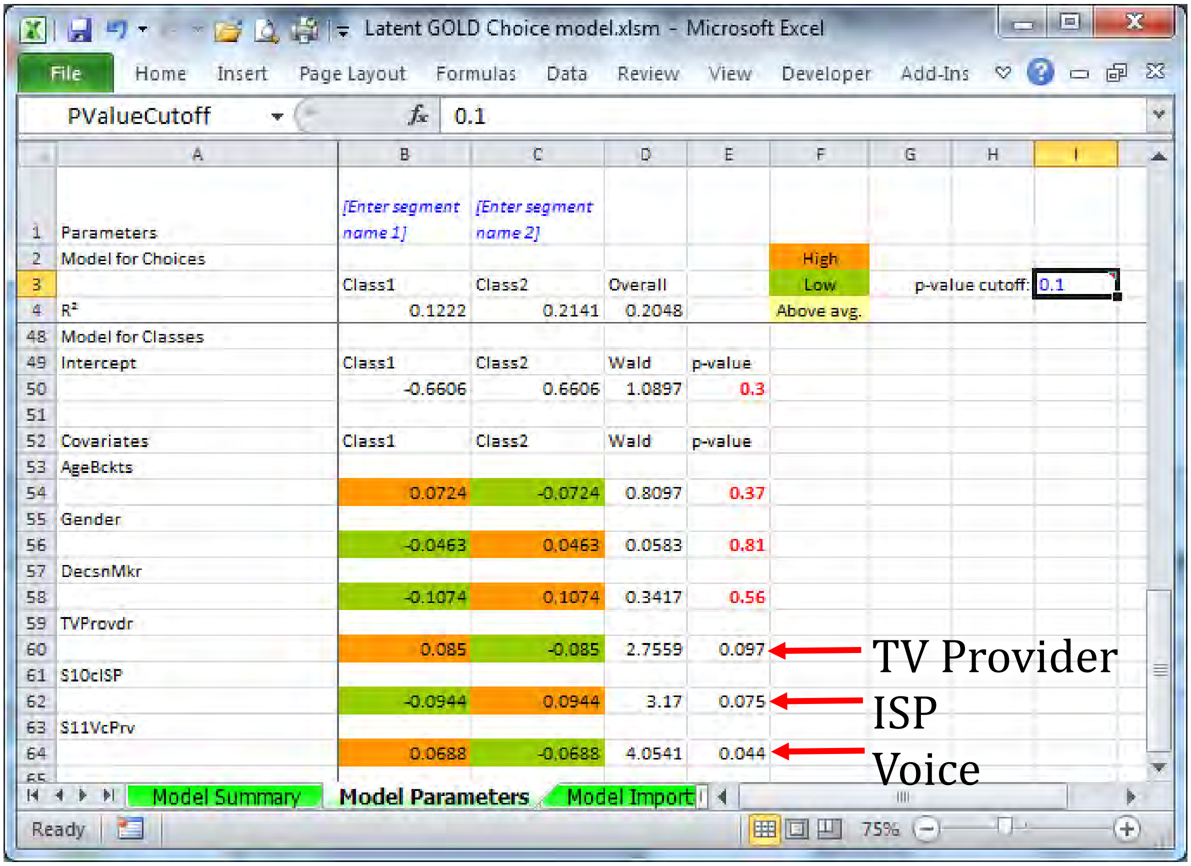
Figure 16. LGC Model Based on Comcast Data



	A	B	C	D	E	F	G	H	I
1	Parameters	[Enter segment name 1]	[Enter segment name 2]						
2	Model for Choices					High			
3		Class1	Class2	Overall		Low		p-value cutoff:	0.1
4	R²	0.1222	0.2141	0.2048		Above avg.			
5	R²(0)	0.3174	0.1993	0.2664					
6									
7	Attributes	Class1	Class2	Wald	p-value	Wald(=)	p-value	Mean	Std.Dev.
8	TVSrvc								
9	None	-0.3183	-0.7224	347.82	2.6E-70	13.2949	0.0099	-0.5181	0.202
10	TV25	-1.2638	-0.9653					-1.1162	0.1492
11	TV50	-0.2125	-0.6112					-0.4096	0.1994
12	TV100	0.812	1.0629					0.9361	0.1254
13	TV200	0.9825	1.236					1.1078	0.1267
14	PrmChnnl								
15	None	0.6747	0.8915	419.945	5.3E-84	15.2748	0.0093	0.7819	0.1084
16	HBO	0.6594	0.4524					0.557	0.1035
17	HBO + SHOWTIME	-1.3117	-1.0117					-1.1634	0.15
18	SHOWTIME + STARZ	0.9477	0.6881					0.8194	0.1298
19	HBO + SHOWTIME + STARZ	0.6764	0.7919					0.7335	0.0577
20	All premium channels	-1.6465	-1.8121					-1.7284	0.0828
21	HaveHD								
22	SD TV Channels	-0.3276	-0.3276	109.069	1.6E-25	0		-0.3276	0

The model allowed us to apply significance tests to isolate important covariates. In this case, personal covariates (age, gender and role in decision making) failed significance tests, whereas covariates describing content providers passed. We therefore selected provider variables to append to the CBC/HB utilities file.

Figure 17. Significance Tests on Covariates



Latent GOLD Choice model.xlsm - Microsoft Excel

PValueCutoff | 0.1

	A	B	C	D	E	F	G	H	I
1	Parameters	[Enter segment name 1]	[Enter segment name 2]						
2	Model for Choices					High			
3		Class1	Class2	Overall		Low		p-value cutoff:	0.1
4	R²	0.1222	0.2141	0.2048		Above avg.			
48	Model for Classes								
49	Intercept	Class1	Class2	Wald	p-value				
50		-0.6606	0.6606	1.0897	0.3				
51									
52	Covariates	Class1	Class2	Wald	p-value				
53	AgeBckts								
54		0.0724	-0.0724	0.8097	0.37				
55	Gender								
56		-0.0463	0.0463	0.0583	0.81				
57	DecsnMkr								
58		-0.1074	0.1074	0.3417	0.56				
59	TVProvdr								
60		0.085	-0.085	2.7559	0.097				
61	S10cISP								
62		-0.0944	0.0944	3.17	0.075				
63	S11VcPrv								
64		0.0688	-0.0688	4.0541	0.044				

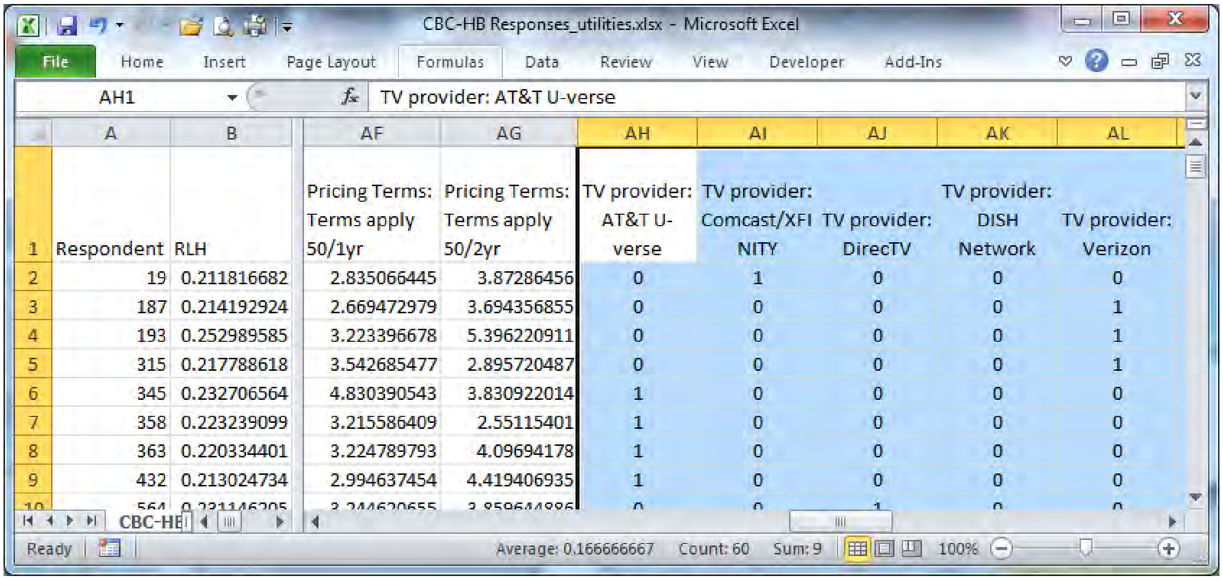
TV Provider

ISP

Voice

We appended the provider variables to the CBC/HB utilities file,

Figure 18. Significant Covariates Appended to Utilities File



CBC-HB Responses_utilities.xlsx - Microsoft Excel

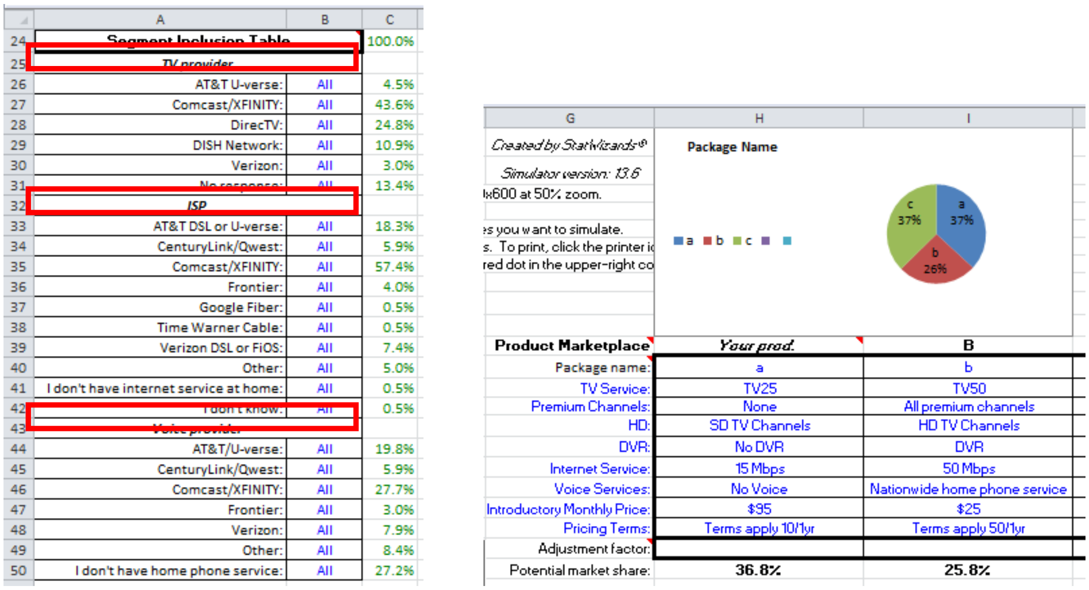
AH1 | TV provider: AT&T U-verse

	A	B	AF	AG	AH	AI	AJ	AK	AL
1	Respondent	RLH	Pricing Terms: Terms apply 50/1yr	Pricing Terms: Terms apply 50/2yr	TV provider: AT&T U-verse	TV provider: Comcast/XFINITY	TV provider: DirecTV	TV provider: DISH Network	TV provider: Verizon
2	19	0.211816682	2.835066445	3.87286456	0	1	0	0	0
3	187	0.214192924	2.669472979	3.694356855	0	0	0	0	1
4	193	0.252989585	3.223396678	5.396220911	0	0	0	0	1
5	315	0.217788618	3.542685477	2.895720487	0	0	0	0	1
6	345	0.232706564	4.830390543	3.830922014	1	0	0	0	0
7	358	0.223239099	3.215586409	2.55115401	1	0	0	0	0
8	363	0.220334401	3.224789793	4.09694178	1	0	0	0	0
9	432	0.213024734	2.994637454	4.419406935	1	0	0	0	0
10	564	0.231146205	3.244620655	3.859644886	0	0	1	0	0

Ready | Average: 0.166666667 | Count: 60 | Sum: 9

and we used the combined file to construct an Excel-based simulator.

Figure 19. Comcast Simulator with Covariates as Filters



	A	B	C
24	**Segment Inclusion Table**		100.0%
25	***TV provider***		
26	AT&T U-verse:	All	4.5%
27	Comcast/XFINITY:	All	43.6%
28	DirecTV:	All	24.8%
29	DISH Network:	All	10.9%
30	Verizon:	All	3.0%
31	No response:	All	13.4%
32	***ISP***		
33	AT&T DSL or U-verse:	All	18.3%
34	CenturyLink/Qwest:	All	5.9%
35	Comcast/XFINITY:	All	57.4%
36	Frontier:	All	4.0%
37	Google Fiber:	All	0.5%
38	Time Warner Cable:	All	0.5%
39	Verizon DSL or FiOS:	All	7.4%
40	Other:	All	5.0%
41	I don't have internet service at home:	All	0.5%
42	I don't know:	All	0.5%
43	***Voice provider***		
44	AT&T/U-verse:	All	19.8%
45	CenturyLink/Qwest:	All	5.9%
46	Comcast/XFINITY:	All	27.7%
47	Frontier:	All	3.0%
48	Verizon:	All	7.9%
49	Other:	All	8.4%
50	I don't have home phone service:	All	27.2%

Created by StatWizards®

Simulator version: 13.6

x600 at 50% zoom.

es you want to simulate.
s. To print, click the printer ic
red dot in the upper-right co

G	H	I
Product Marketplace	***Your prod.***	**B**
Package name:	a	b
TV Service:	TV25	TV50
Premium Channels:	None	All premium channels
HD:	SD TV Channels	HD TV Channels
DVR:	No DVR	DVR
Internet Service:	15 Mbps	50 Mbps
Voice Services:	No Voice	Nationwide home phone service
Introductory Monthly Price:	$95	$25
Pricing Terms:	Terms apply 10/1yr	Terms apply 50/1yr
Adjustment factor:		
Potential market share:	**36.8%**	**25.8%**

6. Summary

In most marketing situations, covariates matter. All other things being equal, younger people express greater demand for technology products than older people. Women have greater preference than men for manicures, and the list goes on. In modeling choices, it's important to identify which of many possible covariates are the ones associated with demand for a product. We can do that directly in Latent GOLD choice or mixed logit models. In those situations where CBC/HB is the technique of choice, one can employ alternate methodologies to select important variables to include in a market simulator.

George Boomer

Kiley Austin-Young

UNCOVERING CUSTOMER SEGMENTS BASED ON WHAT MATTERS MOST TO EACH

EWA NOWAKOWSKA[1]
GFK CUSTOM RESEARCH NORTH AMERICA
JOSEPH RETZER[2]
MARKET PROBE

I INTRODUCTION

The world is changing; the new digital consumer is no longer forced to choose between a limited number of available options; more is available and more is acceptable; markets are becoming more fragmented. Also, not everybody has access to everything. People use smartphones, tablets, computers; have access to different apps and different software; and can be reached via different media and channels. Joel Cadwell wrote on his blog that “the wanting and the means impose a structure” and this is the structure we need to uncover. However this is not something we can do within the classical segmentation framework. The classical image of separate clouds of points in a common space is based on the assumption that everything is described by everything, which is no longer the case. Cadwell continues, “Everyone does not own every device, nor do they use every feature. Instead, we discover recurrent patterns of specific device usage at different occasions with a limited group of others.” So we need to look at the challenge of segmentation from a different angle—instead of describing everyone by everything, we must search for interesting areas in the data, identifying subgroups of people and the attributes that matter to them. This is the rationale behind the approach we are presenting here.

The paper is split into two sections—the introduction gives an overview of the technique, talks about the benefits and goes into detail about the process. The example that follows starts with a description of the data we are employing and then discusses the ever-present challenge of determining the number of clusters. As we are talking about co-clusters here, the process is more involved. We need to specify both the number of row and column clusters, where row clusters are groups of respondents and column clusters are groups of variables. We present an approach for addressing this challenge and then discuss the results.

1.Overview

What is Co-Clustering? It is an emerging method that allows for analysis of dyadic data connecting two entities. The entities are typically the rows and columns of the data set. Co-clustering entails a simultaneous segmentation of rows and columns of the data matrix. It is essential that this segmentation explicitly utilizes relationships between the entities. This is a meaningful difference with respect to other seemingly similar approaches e.g., independent clustering of rows and columns of the data matrix. This approach produces blocks in the data that appear similar to co-clusters but in fact are not as it ignores dyadic (pairwise) relationships

[1] Director, Marketing & Data Sciences. 8401 Golden Valley Rd , Minneapolis, MN 55427, USA. T: 515 441 0066. ewa.m.nowakowska@gmail.com

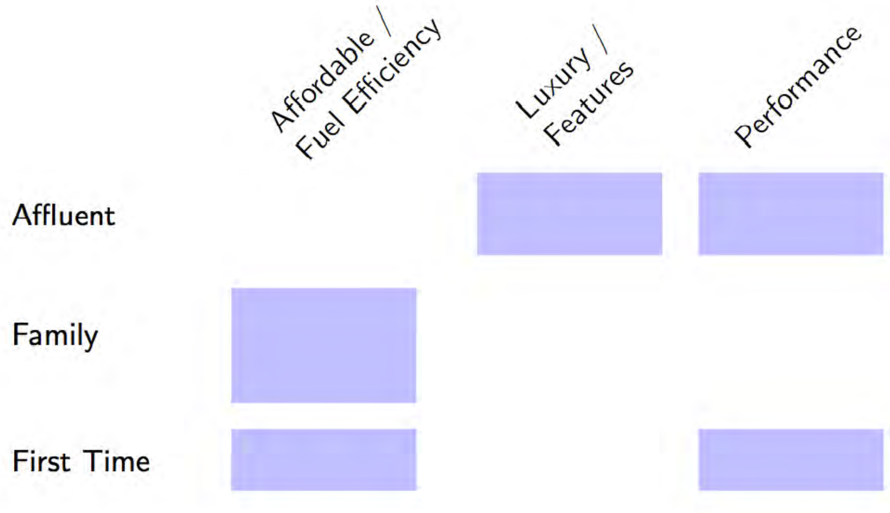
[2] CRO. 2655 N. Mayfair Road, Milwaukee, WI 53226, USA. T: 414 778 6000. retzerjj@gmail.com

between rows and columns. The process may be thought of as clustering *elements of the data matrix* according to certain rules rather than clustering rows and/or columns directly. Once we have the co-clusters we can project them on rows and columns but it is important to remember that this is not how the co-clusters are created. Finally, similarly to a classical segmentation, co-clusters may be profiled on covariates in order to facilitate understanding of the uncovered groups and their potential future targeting.

2. Key Benefit of Co-Clustering

So what can co-clustering do? A typical co-clustering task is to identify customer groups seeking specific sets of products or product features. Let us look at a hypothetical example that illustrates this application.

Figure 2.1 Co-Clustering Example

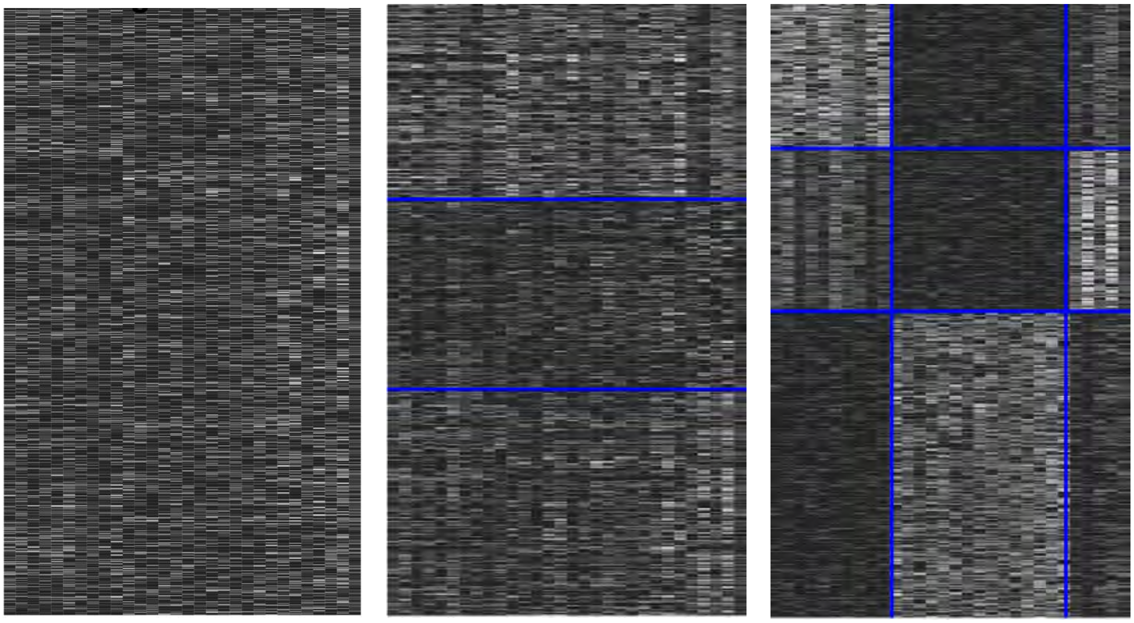


Imagine a data set containing information on how important certain car features are for respondents. The outcome from co-clustering performed on this kind of data set might look like what is shown in Figure 2.1. Here, the algorithm found 3 groups of car buyers identified as affluent, buying a car for a family and buying their first car. The algorithm simultaneously grouped features into 3 groups that the researcher interpreted as corresponding to efficiency, luxury and performance. The outcome in Figure 2.1 shows the relationship between groups of people and groups of variables by reflecting what is important for each group. In this example the affluent respondents focus on luxury & performance. Efficiency matters most to family buyers and for the first timers it is both efficiency and performance. Note also that in terms of group size family is the largest segment.

3. Co-Clustering versus Classical Clustering

How does co-clustering differ from classical segmentation? This example is based on real data. Figure 3.1 shows the heatmap of the original data along with classical clusters and co-clusters as returned by latent class and co-clustering algorithms respectively. The blue lines indicate partitions and the data points are re-arranged according to correspond to the clustering method partition.

Figure 3.1 Raw Data, Clustered Data, Co-Clustered Data



Theoretically, one could do a regular segmentation (the chart in the middle), divide the respondents into groups based on the similarity of their overall profiles and then rearrange variables to find blocks most relevant for each cluster, hoping to get an outcome similar to the block structure of the co-clustering solution (the chart on the right). But this approach would cluster rows and column independently, not taking into account the dyadic relationship. This approach would produce a suboptimal solution however. The difference in the outcome can be seen in the charts of Figure 3.1. The row cluster partitions differ across solutions, and independent re-arrangement of the columns for the classical solution would not change that. In this example, classical clustering finds three clusters of balanced sizes while co-clustering returns one larger and two smaller row clusters. Co-clustering explicitly utilizes what matters most—the relationship between rows and columns, e.g., between people and products/features or whatever might be represented by the columns of the data set. Hence when dyadic relationships underlie the structure we try to uncover, this approach tends to outperform classical segmentation attempts.

4. Other Notes and Benefits of Co-Clustering

Co-clustering was introduced by Hartigan 1972 however only recently have the algorithms improved at reaching global as opposed to local convergence (Bro et al., 2012) and became of interest to a wider group of practitioners. The algorithm used in this work—block cluster (Govaert and Nadif, 2003)—extends the standard latent class model. Similarly to classical latent class analysis, it produces probabilities of belonging allowing respondents to be in multiple row clusters and features in multiple column clusters at the same time. It works best when blocks can be detected in the data, which in practical terms typically means that certain level of sparsity in the data can be observed. Hence genom data, web traffic data or text mining are good examples of other areas where co-clustering may establish its position relatively early. In the case of text mining there are typically many words in the corpus and only its small subsample represented in each document, so the document-term matrix is sparse and co-clustering has the potential of being successful in finding the underlying structure. Online traffic is another example—with large numbers of websites and each person visiting only a small selection we obtain a sparse traffic matrix, which can efficiently be examined with co-clustering algorithms. Classical co-clustering takes into account market fragmentation and has the potential of providing more relevant insights into the structure of relationships and/or preferences but is still relatively easy to perform and interpret. There are also Bayesian approaches to co-clustering that offer certain extensions and individual level analysis. These methods can make the model and estimation much complex however (Shan and Banerjee, 2008; Ekina et al., 2013).

There are two more features of co-clustering that are considered beneficial when compared to classical segmentation approaches. First, noise in the data has relatively small impact on the ultimate partition. Due to the similarity of distributions and typically low correlation with meaningful features, noise variables are often grouped together in a single grouping, and hence do not have substantial impact on the uncovered column clusters. Also, since by definition noise is more or less evenly distributed over all the row clusters, it doesn't affect the row partition in any meaningful way either. Second—the clusters tend to be nested in a practical sense, meaning new clusters are typically formed by a split of existing clusters. Both phenomena are illustrated in Figures 4.1a and 4.1b.

Figure 4.1a Clustering-migrations

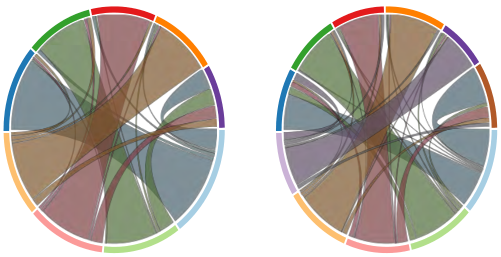
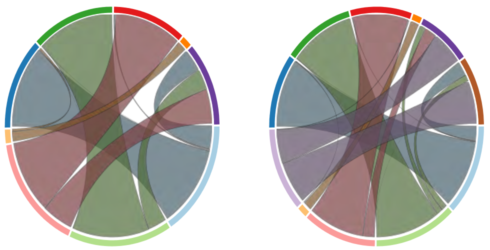


Figure 4.1b Co-clustering–migrations

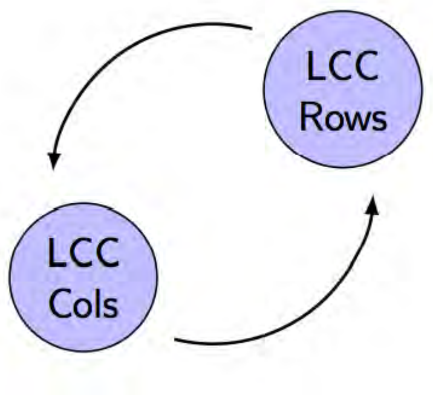


The graphs in Figure 4.1a and 4.1b depict a new form of a clustogram which offers a way of analyzing between-cluster migrations. For each circle the bottom half shows one solution and the top half shows another solution. The ribbons reflect the migrations from the clusters of the bottom to the clusters of the top. The first graph in each pair shows migrations between 4 and 5 segment solutions, while the second—from 5 to 6 segment solutions. The pair in Figure 4.1a shows between segment migration for standard clustering, while the pair in Figure 4.1b illustrate migrations for an increasing number of row clusters in the case of co-clustering. For standard clustering we see a fair amount of near to random migrations, typically due to the noise in the data. For co-clustering the segments seem much more stable in this respect. That is to say an increase in the number of segments triggers a split but the random migrations are minimal.

5. The Estimation

The co-clustering algorithm used in this work is referred to as "block clustering." Figure 5.1 graphically illustrates how it works.

Figure 5.1 Estimation Process



Being an extension of latent class clustering, it estimates the parameters for the rows, uses the estimates to repeat estimation for the columns iterating between the two solutions until it reaches convergence. The implementation used in this work comes from the R package "blockcluster."

II Showcase

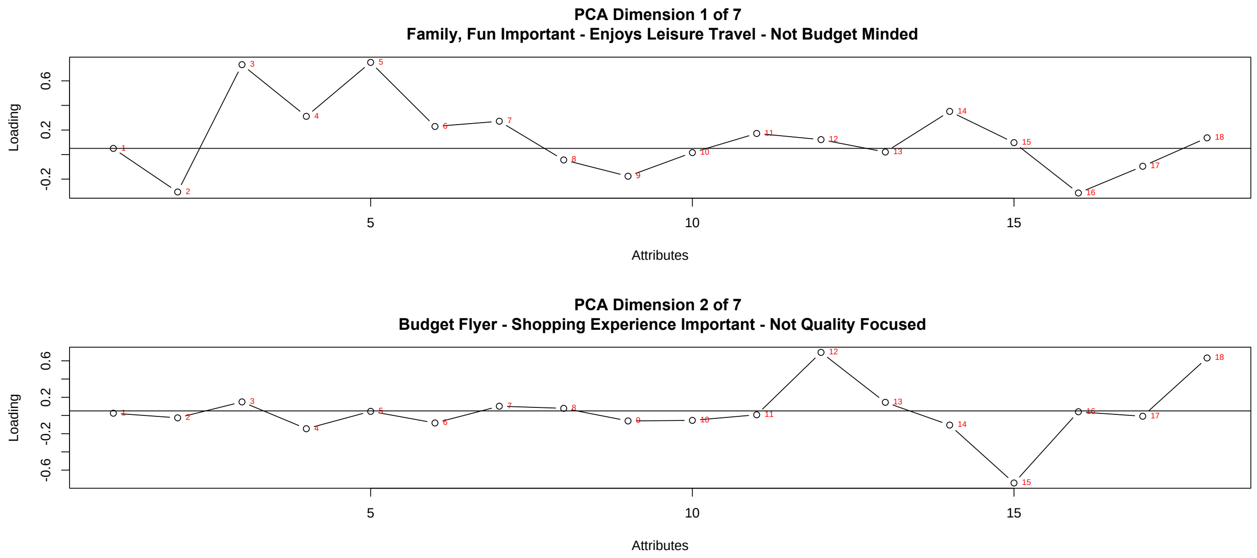
6. The Data

We illustrate co-clustering using disguised airline traveler data. The data consists of survey collected attitudinal variables, binary needs-based measures and database demographics. Examples of each are given below:

- Survey collected attitudinal data (Basis Variables), e.g.,
  - Prefer to travel mostly for fun/visit family,
  - Tend to embrace new technology,
  - Passionate about my job, etc.
- Survey collected, binary needs-based measures, e.g.,
  - Airline good for long distance travel,
  - Airline loses luggage too often,
  - Airline flies direct to my preferred destinations, etc.
- Data-base and demographic information, e.g.,
  - Business traveler,
  - Current customer,
  - Rewards member tenure, etc.

The challenge of identifying relevant groups of respondents based on, in total, 18 basis variables, can be seen in plots of the first two principle components as shown in Figure 6.1 below:

Figure 6.1 PCA Plots



PCA Dimension 1 of 7
Family, Fun Important - Enjoys Leisure Travel - Not Budget Minded

PCA Dimension 2 of 7
Budget Flyer - Shopping Experience Important - Not Quality Focused

The first principal component loads strongly on

- (3) Travel is fun
- (4) Embrace New Technology
- (5) Travel for Pleasure
- (6) Travel new Places
- (7) Plan ahead

while the second loads on

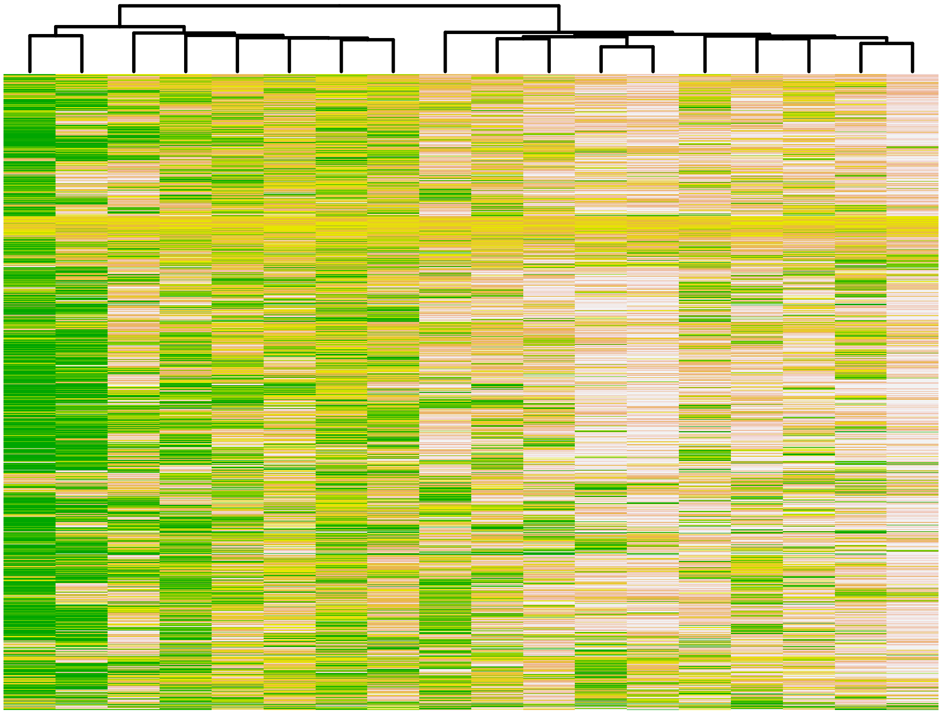
- (12) Least Expensive Flight
- (14) Like Different Brands
- (18) Ideal Shopping

Interpretation of the first seems clear, however, when moving to the second PC and beyond, it becomes more difficult. This may be due to the fact that the PCA is attempting to describe ALL respondents on ALL attributes simultaneously. Co-clustering on the other hand allows different respondent clusters to be defined by differing subsets of variables.

7. Selecting the Number of Co-Clusters

Selecting the number of co-clusters entails not only deciding on the number of groups of respondents (rows) but also the number of distinct sets of basis variables (columns). Cursory examination of the data may be performed using simple hierarchical clustering of both rows and columns independently using the “heatmap” function in R. It is important to note that as this approach does not take into consideration the dyadic relationship between rows and columns, it is not in fact co-clustering. An illustration of a heat map representation of the data with a dendrogram representing attribute clustering is shown in Figure 7.1 below. It would appear from this depiction of the data that 2 column clusters are most likely.

Figure 7.1 Data Heat Map



Proper pre-specification of the number of clusters in a co-clustering analysis can be a challenge. It is well known that numerous measures for comparative evaluation of standard (non co-cluster) partitions exist. These measures are primarily based on "cluster quality." Cluster quality simply put implies respondent groupings that are (1) similar within the group and (2) different between groups[3].

Measuring co-cluster quality however requires we condition on the subsets of attributes comprising the column clusters when evaluating the quality of the respondent groupings (or vice versa). A recent article by Charrad et al. (2010) suggests such an approach which is suitable for simultaneous specification of both the number of row and column clusters.

Three metrics of cluster quality will be considered in the analysis:

1. **Hubert's Gamma**: correlation between distances and a 0–1 vector where 0 means same cluster, 1 means different clusters.
2. **Dunn Index**: minimum separation between clusters/maximum cluster diameter.
3. **Calinski & Harabasz:**

[3] It is important to note that cluster quality must always be considered in conjunction with the partition's ability to "facilitate marketing strategy."

$$C(g) = \frac{tr(\boldsymbol{B})}{g-1} \bigg/ \frac{tr(\boldsymbol{W}\ )}{n-g} \quad \text{where}$$

$$\boldsymbol{W} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x}_m)(x_{ml} - \bar{x}_m)' \text{ (within group dispersion)},$$

$$\boldsymbol{B} = \sum_{m=1}^{g} n_m(\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})' \text{ (between group dispersion)}$$

Stacked bar charts depicting the relevant index value for row and column cluster combinations are shown in Figure 7.2.

Figure 7.2 Quality Measures

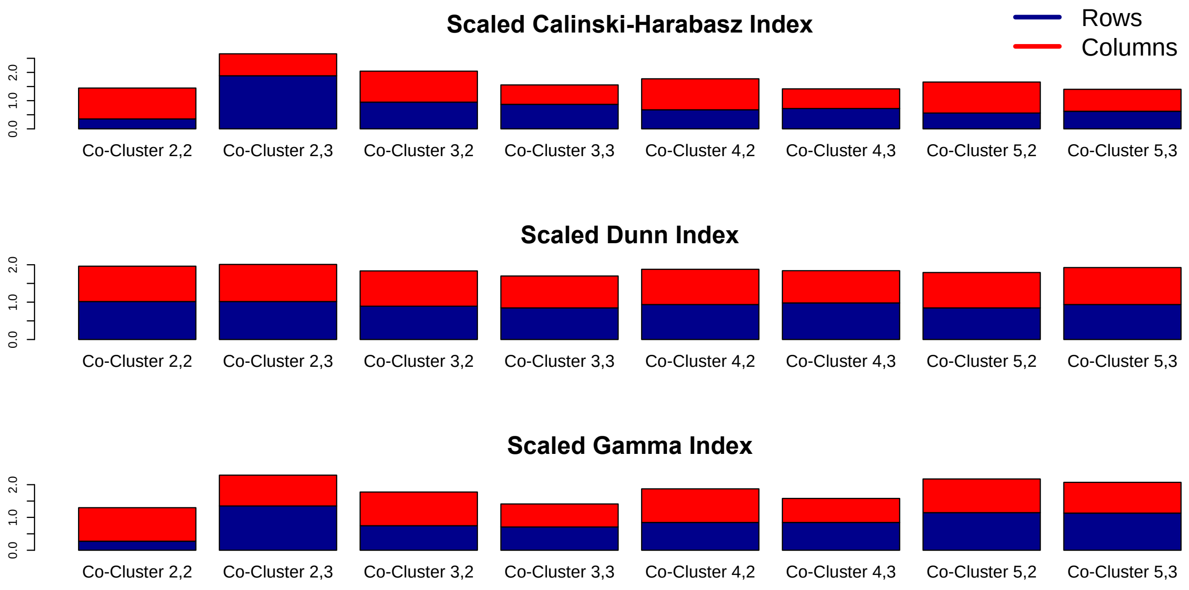


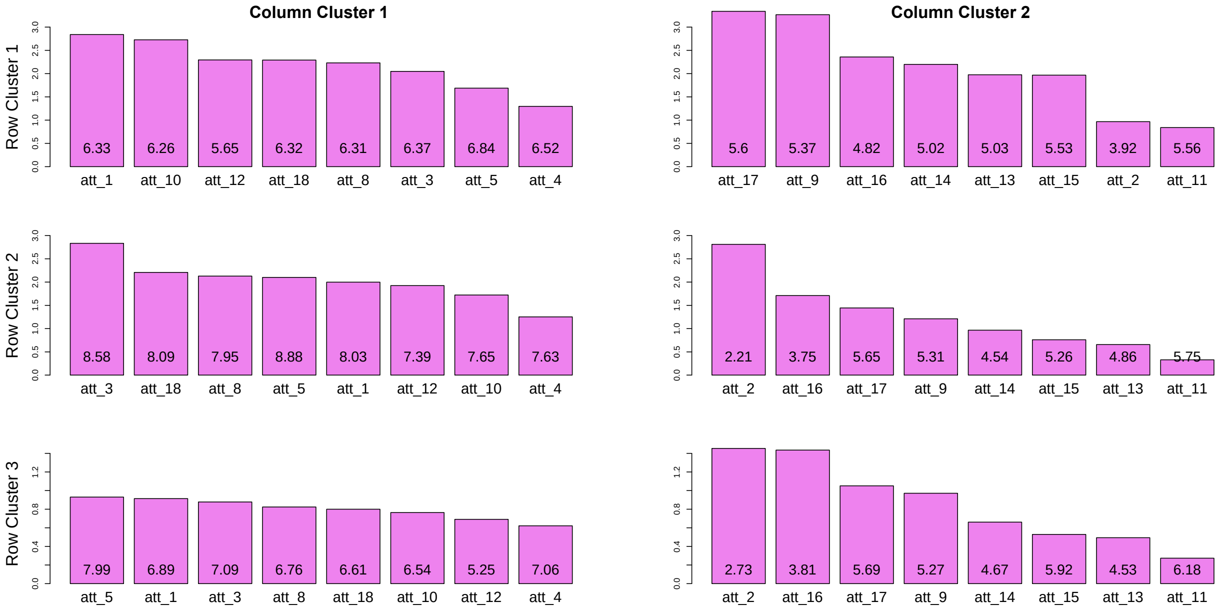
Figure 7.2 suggests no strongly dominant co-cluster solution based on quality exists. While co-clusters 2,3 for each index seem slightly better than the closely competing 3,2 solutions, the latter was chosen based on its ability to "facilitate marketing strategy" evident in the column cluster interpretations. Specifically, the column clusters can be described using strongly associated basis variables as:

- Column Cluster 1 Attributes:
    - Prefer to visit family or friends during holidays
    - Enjoy traveling for pleasure
    - Travel is fun
- Column Cluster 2 Attributes
    - Career importance
    - Passionate about work
    - Check bags when traveling

8. Identifying the Co-Clusters

In order to understand how each of the 3 row clusters align with the column clusters, each co-cluster's attribute importances were measured using relative variation within the cluster. Higher attribute importances within a given column cluster identifies the co-cluster (as defined by both the row and column cluster).

Figure 8.1 Importance's by Co-Cluster



Examination of the importances by co-cluster suggests that:

- Row cluster 1 seems relatively similar across the two column clusters
- Row cluster 2 appears more strongly related to column cluster 1
- Row cluster 3 appears more strongly related to column cluster 2

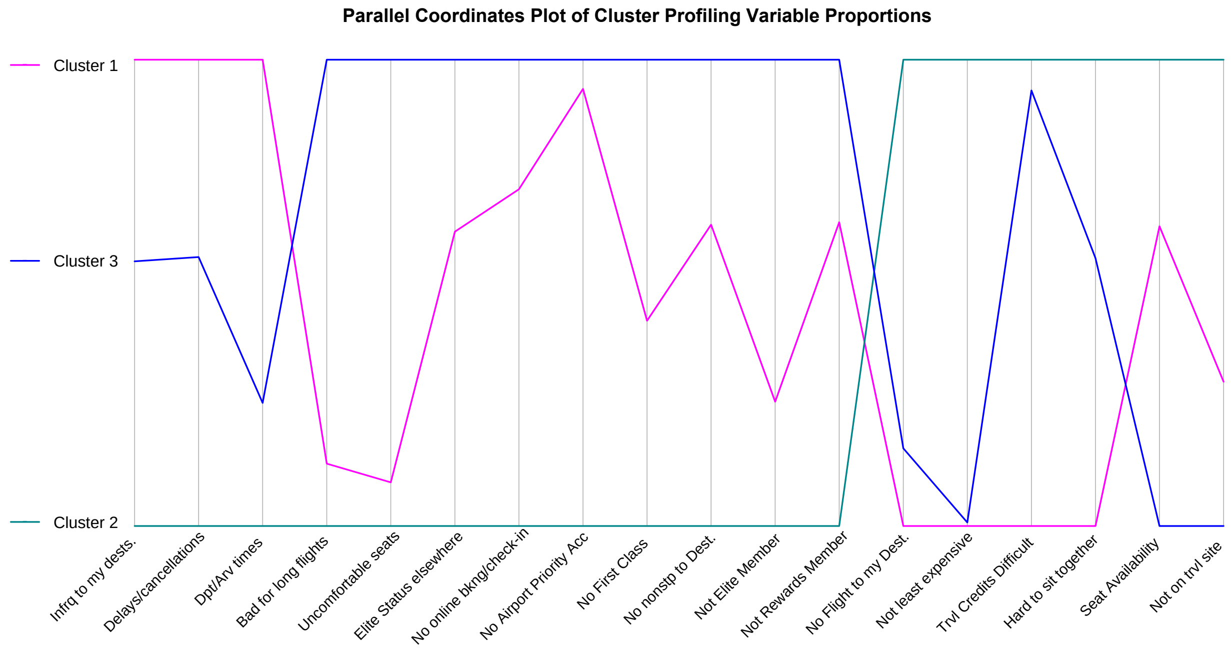
These relationships may be summarized in a conceptual block diagram such as that illustrated in Figure 8.2.

Figure 8.2 Co-Clusters

	Visit Family / Fun	Business Orientated
Row Cluster 1	■	■
Row Cluster 2	■	
Row Cluster 3		■

Additional insight into the nature of the row clusters is given in a parallel coordinates plot of the profiling variables as shown in Figure 8.3.

Figure 8.3 Profiling Variables Parallel Coordinates Plot



While row cluster 1 appears to be described equally well across both column clusters, row clusters 2 and 3 offer more unique insights when looking at the column cluster subsets. An overview of their description is give as follows:

- Row 2, Column 1 Co-Cluster:
  - Rewards/elite membership not barrier to choosing airline.
  - Not barriers to choosing airline:
    - First Class Seating

        - Airport Priority Access
        - Non-stop to Destination
        - Comfortable Seats
    - Looking for Cheapest Flight.
    - Online booking not a priority.
- Row 3, Column 2 Co-Cluster:
    - Rewards/elite membership may be barrier to choosing airline.
    - Likely barriers to choosing airline:
        - First Class Seating
        - Airport Priority Access
        - Non-stop to Destination
        - Comfortable Seats
    - Less concerned about getting Cheapest Flight.
    - Prefer online booking.

9. Conclusion

Modern data sets often contain large numbers of potential basis variables for clustering. This presents a challenge for traditional clustering algorithms as evidenced in resultant low quality partitions when applying standard cluster algorithms to high dimensional data. It is important to recognize that respondents may be very similar on subsets of basis variables rather than on all of them simultaneously. Co-clustering is one approach to identifying those variable subsets and in turn providing market researchers with greater insights into their data.

Ewa Nowakowska

Joseph Retzer

REFERENCES

Bro, R., Papalexakis, E. E., Acar, E. and Sidiropoulos, N. D. (2012). *Co-clustering—a useful tool for chemometrics*. Journal of Chemometrics, 26(6):256–263.

Charrad, M., Lechevallier, Y., Ben Ahmed, M. and Saporta, G. (2010). *On the Number of Clusters in Block Clustering Algorithms.* Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), 1–6.Ekina, T., Leva, F., Ruggeri, F. and Soyer, R. (2013). *Application of Bayesian Methods in Detection of Healthcare Fraud*. Chemical Engineering Transaction, 33.

Govaert, G. and Nadif, M. (2003). *Clustering with block mixture models*. Pattern Recognition, 36(2):463–473.

Govaert, G. and Nadif, M. (2007). *Clustering of contingency table and mixture model*. European Journal of Operational Research, 183(3):1055–1066.

Hartigan, J. A. (1972). *Direct clustering of a data matrix*. Journal of the American Statistical Association, 67(337):123-129.

Shan, H. and Banerjee, A. (2008). *Bayesian co-clustering*. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), 530–539.

Appendix:

R-code for blockcluster

```
 library(blockcluster) # Load R blockcluster package

 nRowClusts <- 2       # Select number of Row Clusters
 nColClusts <- 2       # Select number of Column Clusters

# Run blockcluster

 out <-cocluster(basisData,datatype="binary",collabels=colnames(basisData),
+                nbcocluster=c(nRowClusts,nColClusts))

 summary(out)        # Simple summarize
 plot(out)           # Heatmap plot
```

CLIMBING THE CONTENT LADDER: HOW PRODUCT PLATFORMS AND COMMONALITY METRICS LEAD TO INTUITIVE PRODUCT STRATEGIES

SCOTT FERGUSON[1]
NORTH CAROLINA STATE UNIVERSITY

1. MOTIVATION

It is generally accepted that content laddering occurs in practice. More expensive products often have higher-end features that meet or exceed those found on their cheaper counterparts. A higher-end sailboat, for example, is likely to have an improved auto-pilot, repeaters that provide information to multiple instrument panels, and a better mast material (RCR Yachts, 2015). Similarly, a higher-end drill may have a wider range of clutch settings, additional torque, or a longer run time (DeWalt, 2015). Satisfying different business goals in a heterogeneous market requires different price points to be targeted, and the product configurations at each price point likely are achieved by unique combinations of features and specifications.

To meet this challenge, assume that you have already used one of Sawtooth Software's products to survey thousands of respondents and have used the CBC/HB module (Sawtooth Software, 2009) to estimate part-worths. Thinking that optimization can help you search the large solution space (Mulhern, 2007), you launch SMRT with ASM (Sawtooth Software, 2003) and set up your problem. After selecting an objective and an algorithm to explore the space, you define the number of products to offer (Chapman and Alford, 2011) and establish the possible attribute levels for each product. The product search begins, and as shown in Figure 1, the results are not what you expected. This solution does not offer a product strategy that caters to low-end, mid-range and high-end users simultaneously. However, this solution is intriguing because there is a noticeable amount of commonality within each attribute.

Figure 1. Results of an Unconstrained Product Search

Product	Attribute 1	Attribute 2	Attribute 3	Price
Product A	1-8	1-6	1-7	0-600
Product B	1-8	1-6	1-7	0-600
Product C	1-8	1-6	1-7	0-600

Product	Attribute 1	Attribute 2	Attribute 3	Price
Product A	3	3	2	400
Product B	3	4	6	425
Product C	4	4	6	435

Manually placing bounds on product price can lead to increased solution diversity, as shown in Figure 2. Expected behavior of the final solution is that increased distinction between products along the price axis will occur by defining more unique product configurations. After re-running the optimization, you notice that product commonality has been reduced, but the combination of features across the product line would likely confuse a customer (or manager). A high-end feature has been included on *Product A* (Attribute 3) that does not appear on either of the other two products. Perhaps more alarming is that the features offered on the other two products are lower-end options.

[1] Associate Professor, North Carolina State University; scott_ferguson@ncsu.edu

Figure 2. Results of a Product Search with Manually Included Price Constraints

Product	Attribute 1	Attribute 2	Attribute 3	Price
Product A	1-8	1-6	1-7	0-200
Product B	1-8	1-6	1-7	200-400
Product C	1-8	1-6	1-7	400-600

⇨

Product	Attribute 1	Attribute 2	Attribute 3	Price
Product A	1	1	7	155
Product B	2	6	2	310
Product C	7	5	3	500

Understanding that this solution will be difficult to market, you embark on the process of finding a new solution. However, if you manually change the solution, you will likely find a local optimum. Making commonality decisions—how much, where, around which attribute level(s)—requires the user to have insights into the property of the final solution or accept the risk that the guess may result in decreased product line performance. A second strategy involves reformulating the optimization problem.

Previous work by the author has shown that product line solutions obtained from a product search will naturally gravitate toward having some structure (Ferguson and Foster, 2013; Turner et al., 2014). Cheaper products will often have lower-end features, especially when cost information is incorporated. Some commonality may naturally occur when one attribute level is strongly preferred by the entire market. In previous presentations at the Sawtooth Software Users Conference, the author demonstrated how respondent-level part-worth estimates could be used to create a more effective starting point for a genetic search (Turner et al., 2012). The increased solution quality and algorithm efficiency achieved by this approach allowed for multi-objective problem formulations to be considered where the tradeoffs between various business objectives could be more effectively explored. It was shown that using a commonality measure as an objective led to some inherent (and intuitive) content laddering in the final solutions (Ferguson and Foster, 2013).

In this paper, the Commonality Index (CI) will be introduced and used as an objective so that the tradeoff between product diversity and various business goals can be explored. A second approach discussed is reformulating the design string used to represent product configurations by enforcing commonality through a product platforming approach. The hypothesis explored in this work is that these approaches can be used to enforce a more intuitive solution structure while simultaneously finding the optimal configuration.

2. Initial Problem Formulation

To demonstrate the concepts discussed in this paper, consider the hypothetical design of an MP3 player product line. Product attributes are shown in Table 1 and the cost associated with each attribute level is shown in Table 2. Respondent part-worths are estimated using Sawtooth Software's CBC/HB module, ten products are to be designed, and market share is diverted by the "None" option and a set of competitor products. Product price is calculated by multiplying each attribute cost by a price markup variable and adding a constant base price of $52.

Table 1. MP3 Player Product Attributes Considered

Level	Photo/video/camera	Web/app/ped	Input	Screen size	Storage	Background color	Background overlay	Price
1	None	None	Dial	1.5 in diag	2 GB	Black	No pattern/graphic overlay	$49
2	Photo only	Web only	Touchpad	2.5 in diag	16 GB	White	Custom pattern overlay	$99
3	Video only	App only	Touchscreen	3.5 in diag	32 GB	Silver	Custom graphic overlay	$199
4	Photo and video only	Ped only	Buttons	4.5 in diag	64 GB	Red	Custom pattern and graphic overlay	$299
5	Photo and lo-res camera	Web and app only		5.5 in diag	160 GB	Orange		$399
6	Photo and hi-res camera	App and ped only		6.5 in diag	240 GB	Green		$499
7	Photo, video and lo-res camera	Web and ped only			500 GB	Blue		$599
8	Photo, video and hi-res camera	Web, app, and ped			750 GB	Custom		$699

Table 2. MP3 Player Product Attribute Cost

Level	Photo/video/camera	Web/app/ped	Input	Screen size	Storage	Background color	Background overlay
1	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00
2	$2.50	$10.00	$2.50	$12.50	$22.50	$5.00	$2.50
3	$5.00	$10.00	$20.00	$22.50	$60.00	$5.00	$5.00
4	$7.50	$5.00	$10.00	$30.00	$100.00	$5.00	$7.50
5	$8.50	$20.00		$35.00	$125.00	$5.00	
6	$15.00	$15.00		$40.00	$150.00	$5.00	
7	$16.00	$15.00			$175.00	$5.00	
8	$21.00	$25.00			$200.00	$10.00	

So that multiple solutions can be explored, the problem is formulated as a multi-objective product line optimization. To get started, two objective functions will be maximized—market share of preference and a surrogate measure of profit. Ten products are designed in the line, and the full multi-objective problem formulation is given by Equation 1.

Maximize: *Market share of preference*
Profit
by changing: *Feature content*
1 < Price markup variable for each attribute level < 2 (1)
with respect to: *$49 < Price for each product < $699*
No identical products in the product line
Lower and upper level bounds on each attribute

The optimization problem formulated in Equation 1 consists of 116 design variables—46 price markup variables and 70 product configuration variables, as shown in Figure 3. Market share of preference is given by Equation 2, where n_r is the number of survey respondents, and C1–C5 represent the configurations and price of five competitor products. Profit is approximated by Equation 3, using the contribution margin per person in the market (i.e., per capita). To combine the margin of the four products in the line, a weighting scheme must be constructed using the share of preference of each individual product. This ensures that a product with high margin and low sales does not artificially inflate the metric.

$$SOP = \sum_{i=1}^{10} \left(\frac{\sum_{j=1}^{n_r} \left(\frac{e^{V_{ji}}}{\sum_{k=1}^{10} \left(e^{V_{jk}} \right) + e^{V_{j(C1)}} + e^{V_{j(C2)}} + e^{V_{j(C3)}} + e^{V_{j(C4)}} + e^{V_{j(C5)}} + e^{V_{j(none)}}} \right)}{n_r} \right) \times 100\% \tag{2}$$

$$Profit \approx \sum_{\mathrm{i}=1}^{10} [(\mathrm{P_i} - \mathrm{C_i}) * \mathrm{SOP_i}] \tag{3}$$

To identify the non-dominated points for this problem formulation, a multi-objective genetic algorithm (MOGA) was fielded. The initial population was created using a subset of targeted population designs, and other relevant MOGA parameters are given in Table 3. The MOGA used in this paper was coded in Matlab (Matlab, 2014) and was an extension of the foundational theory presented in (Deb et al., 2002). Figure 4 depicts the location of the solution set when the stopping criterion of 500 generations was achieved.

Figure 3. Illustration of Design String

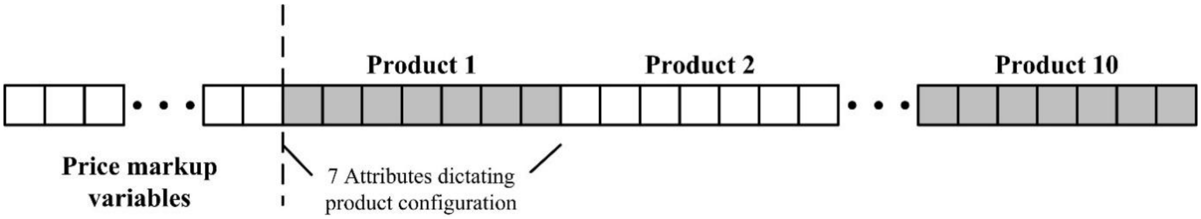
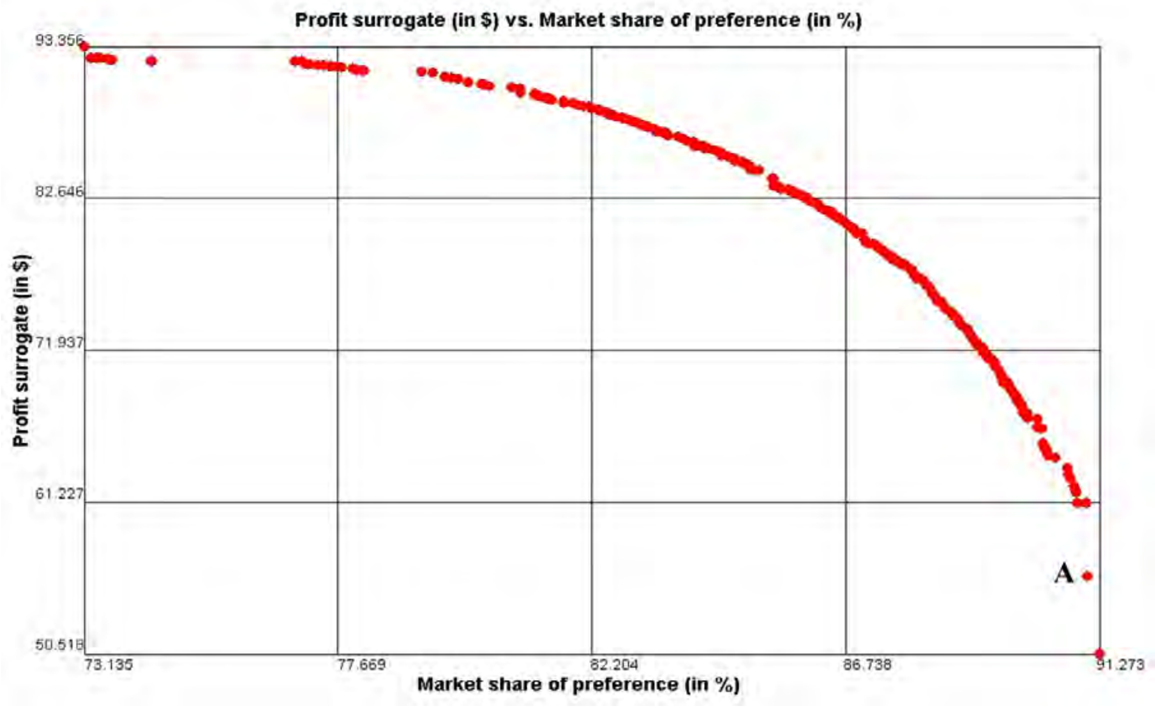


Table 3. Input Parameters for the MOGA

Criteria	Setting
Initial population size	232 (2 times the number of design variables)
Offspring created within a generation	232 (equal to original population size)
Selection	Tournament (4 candidates)
Crossover type	Scattered
Crossover rate	0.5
Mutation type	Adaptive
Mutation rate	5% per bit
Stop after	500 generations

Figure 4. Set of Non-Dominated Solutions after 500 Generations



The location in the solution space denoted by Point A in Figure 4 represents a product line configuration that achieves a market share of preference of 91.07% and a calculated "profit" value of $55.95. The design configuration of the ten products is shown in Table 4. To make these results easier to read:

- Background color and Background overlay have been omitted as these are relatively easy to change and do not represent significant sources of engineering design re-work;
- Product configurations are grouped by common values of the Display size attribute;
- The market share captured by each product is also included.

While an optimization algorithm will try to exploit the configuration and pricing of a product line to maximize each objective, the results in Table 4 show that there is inherently some structure to each solution. For example, low-end products (*P1, P2*) have very basic product configurations. There is also a natural dispersion of products along the price axis. *P1* and *P2* capture share from the low-end segment of the market, *P8* captures share from the high-end segment of the market, and product variety is used in the remaining products to capture share from the middle segment of the market.

Table 4. Design Configuration of the 10 Products Represented by Point A in Figure 4

	Photo/video/camera	Web/app/ped	Input type	Display size	Storage	Price (in $)	Share (in %)
P1	Photo, video and hi-res camera	App only	Dial	1.5 in diag	16 GB	$115.50	8.98%
P2	Photo, video and hi-res camera	Web and app only	Dial	1.5 in diag	16 GB	$128	5.1%
P3	Photo, video and hi-res camera	Web, app, and ped	Touchpad	3.5 in diag	16 GB	$162.15	5.09%
P4	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	3.5 in diag	16 GB	$172.75	18.86%
P5	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$175.50	10.47%
P6	Photo and video only	Web, app, and ped	Touchscreen	4.5 in diag	64 GB	$244.40	10.93%
P7	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	160 GB	$295.05	12.29%
P8	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	5.5 in diag	500 GB	$491.63	4.8%
P9	Photo and video only	Web, app, and ped	Touchscreen	6.5 in diag	64 GB	$269.60	6.85%
P10	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	160 GB	$288.40	7.7%

Looking closely at where variety is introduced in the product line, *P1* is unique in that it is the only product not able to access the web. Rather, apps must be downloaded to the device when it is connected to the computer. This design decision could be justified by arguing that it is a low-end product and removing Wi-Fi capability is achieved with minimal engineering design re-work or added manufacturing costs. However, the uniqueness in *P3* is a case where a lack of commonality may not be justified. This product captures only 5% of the market but is the only product to use a Touchpad as an input type. Otherwise, this product is functionally similar to *P4* which captures 18.86% of the market.

When multiple products are defined with the same display size (*P5, P6, P7* and *P9, P10*) variety generally is achieved by laddering the Storage size attribute. From an engineering perspective, this type of laddering is particularly attractive because it can likely be completed without changing the physical footprint of the device. However, this solution is not perfect and there are a few instances where products would need to be "repaired" so that the final configuration is more intuitive. These repairs to the product string may include:

- Changing the input type for P3—eliminating the single use of the touchpad;
- Changing the display size for P8—rather than introducing a single display size for a single product that captures only 4.8% of the market, it may be more effective to offer the product with a smaller screen size (4.5 in diag) or a larger screen (6.5 in diag);
- Adding a pedometer to P10.

While such repairs could be manually initiated, this example problem illustrates how an optimization algorithm will exploit the design space of a problem. Reformulating an optimization problem to incorporate commonality as a performance measure allows for a trade to be made between exploitation of the space and the "intuitiveness" of the solution. The next section explains how a commonality measure can be added to the multi-objective problem formulation originally posed in Equation 1.

3. Commonality as an Objective—The Commonality Index

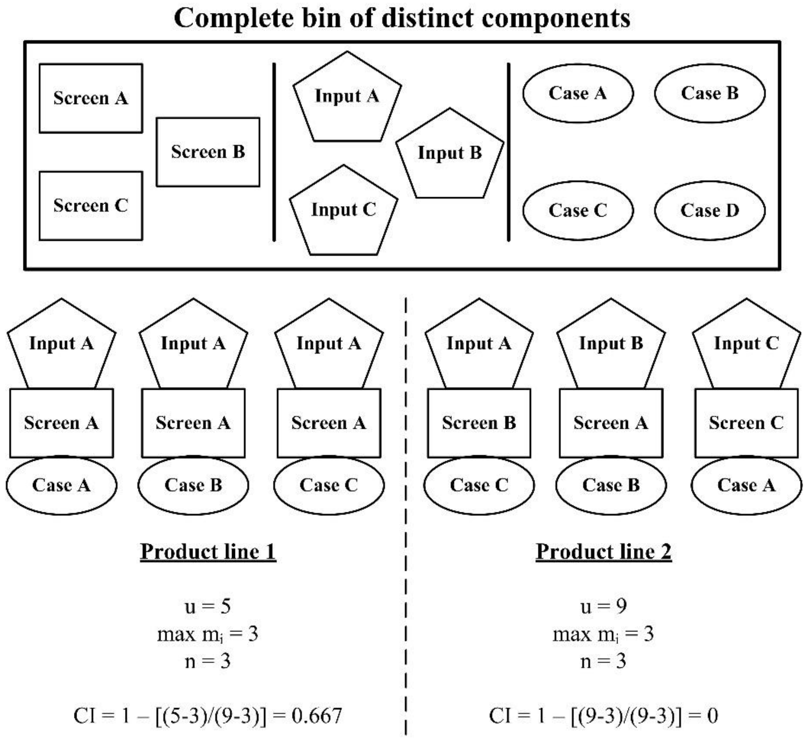
Commonality has been studied in the engineering design literature as a means toward satisfying heterogeneous customer needs while simultaneously driving down manufacturing costs (Thevenot and Simpson, 2006). This allows a firm to offer as much variety to the market as possible while having as little variation between the products themselves. Studies exploring the benefits of commonality have demonstrated that it can lead to a decreased risk during product development (Collier, 1981), reduced inventories and handling costs, and reductions in manufacturing line complexity and retooling times.

Several commonality indices have been developed in the literature so that commonality within a set of products can be measured. Such measures are often based on the number of common components, their costs, and their manufacturing processes. The presence of a quantifiable metric provides a starting point for benchmarking and comparisons between possible solutions. The Commonality Index (CI) was introduced by Martin and Ishii (1996, 1997) as a measure of unique parts in a product line solution. As shown in Equation 4, u is the total number of unique feature levels in the entire product line, m_i is the number of components used in variant i, and n is the number of variants in the product line. CI ranges from 0 to 1, where a smaller value indicates a greater number of unique parts used. While an advantage of this metric is that it is easy to compute, it only focuses on unique parts and not the costs of the components (or of manufacturing).

$$CI = 1 - \frac{u - \max m_i}{\sum_{i=1}^{n} m_i - \max m_i} \tag{4}$$

Example calculations of the CI metric are shown in Figure 5. Here, there are three screen options, three input types, and four case options that can be used to create a product line of three variants. In Product line 1, the three products are created using five unique options (one input, one screen and three cases). Each product consists of three components, and the CI measure for this solution is 0.667. Conversely, Product line 2 uses nine unique components and has a CI of 0.

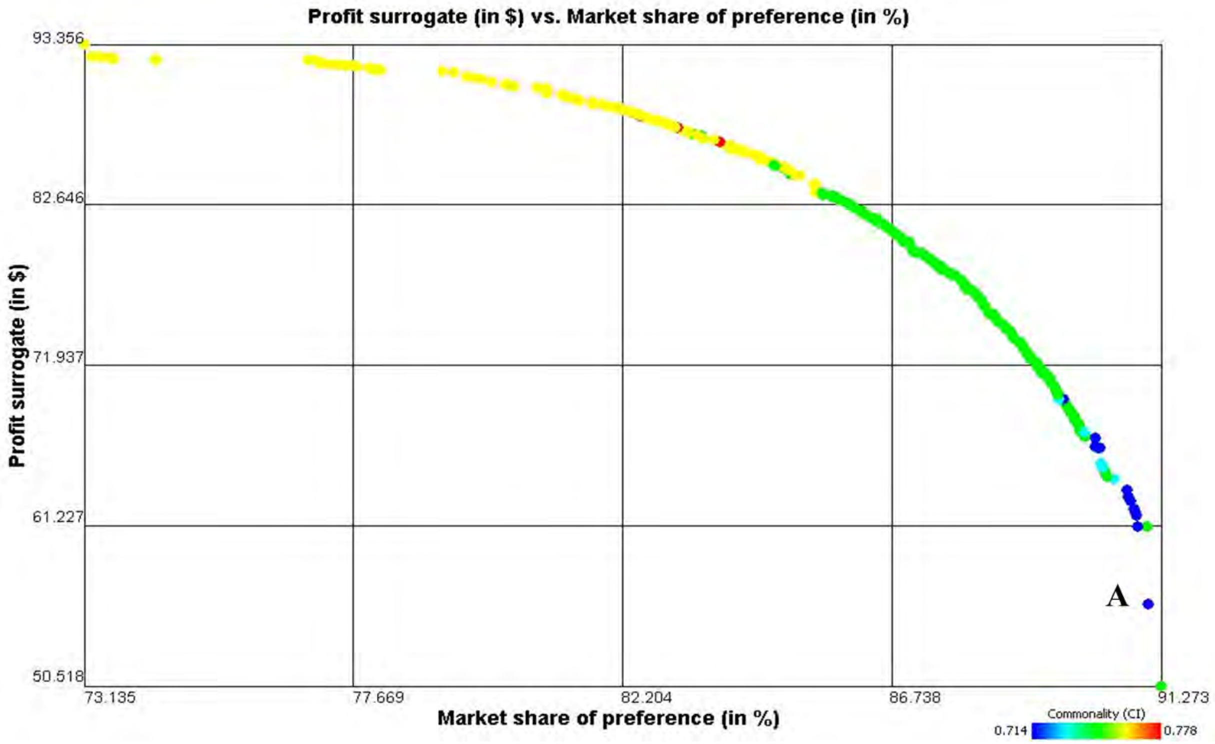
Figure 5. Example Calculations of the CI Metric



Returning to the solution set originally shown in Figure 4, the graph in Figure 6 shows the evaluation of each design string using the CI measure. While CI goes between 0 and 1, we only see a small envelope of that space in this frontier of non-dominated solutions. Additionally, Point A has one of the lowest levels of CI in the set of solutions. As commonality is sacrificed the market share of preference captured by a solution decreases but the estimate of profit increases.

Achieving a wider range of CI in the reported set of solutions requires the problem to be restructured so that the trades considered include more than just share of preference and profit. This is because as currently formulated the optimization problem does not explicitly consider the potential cost savings associated with increased commonality. As multi-objective problem formulations allow for a nearly “infinite” number of objectives to be considered simultaneously, it is possible to reformulate the problem originally posed in Equation 1 to include all three objectives. This new problem formulation is shown in Equation 5.

Figure 6. Original Set of Non-Dominated Solutions Evaluated Using CI Metric



Maximize: *Market share of preference (given in Equation 2)*
Profit (given in Equation 3)
CI (given in Equation 4)
by changing: *Feature content*
1 < Price markup variable for each attribute level < 2
(5)
with respect to: *$49 < Price for each product < $699*
No identical products in the product line
Lower and upper level bounds on each attribute

Efforts to develop effective tradespace exploration tools (Stump et al., 2004; Stump et al., 2009; Daskilewicz and German, 2011) facilitate multidimensional visualization, filtering of unwanted solutions, and detailed exploration of interesting regions of the solution space. Tradespace exploration tools, like Penn State's ARL Trade Space Visualizer (ATSV) (Stump et al., 2004; Stump et al., 2009), can also be linked to optimization algorithms to enable real-time user interaction and design steering. For the purpose of this paper, the three-dimensional solution space is projected into two dimensions, as shown in Figure 7. This is done to show how various stratifications of common CI values allow for trades in market share of preference and profit.

This two-dimensional projection shows that increased commonality within a design solution leads to a reduced trade in the other business goals. Further, it is seen that the overall ranges of CI found in this optimization exceed that found when only share of preference and profit are considered. Point A in Figure 4 required a degree of repair to make the solution more intuitive to customers and managers. Now that the problem has been reformulated, Point B can be identified as a solution that trades market share for increased profit and increased solution commonality.

Figure 7. Two-Dimensional Projection of the Three-Dimensional Solution Space

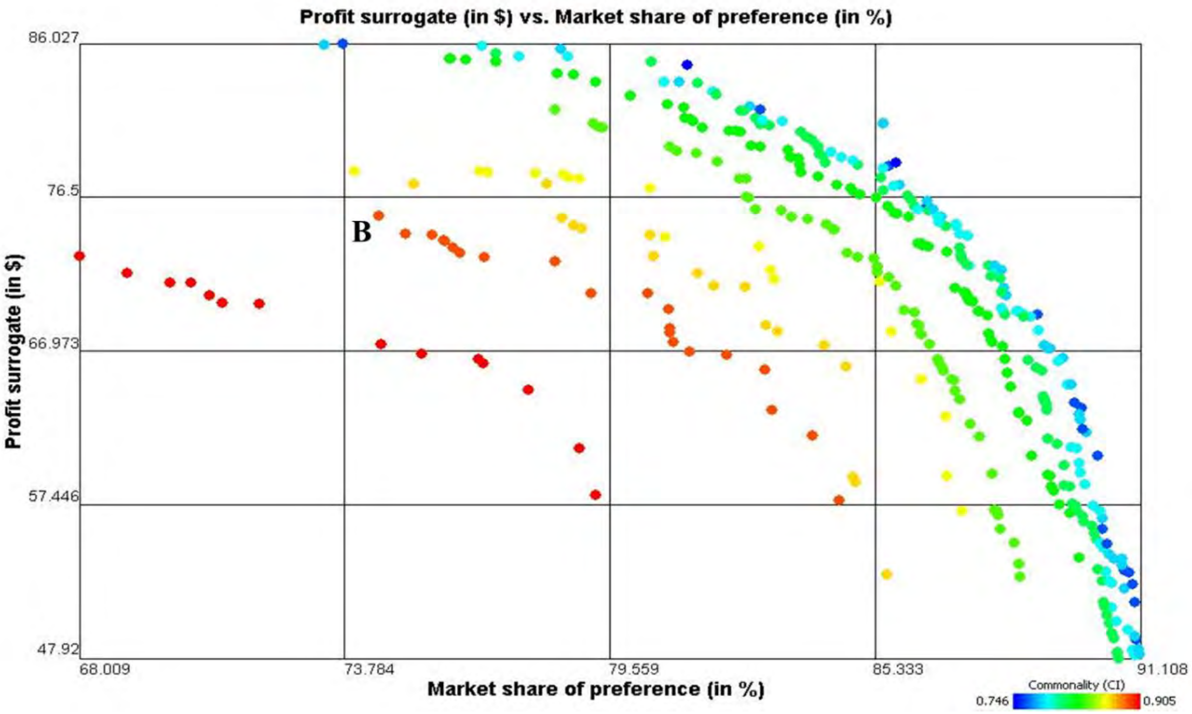


Table 5 reports the detailed design configurations associated with Point B. This solution uses only three different display sizes. The Photo/video/camera attribute is represented by the same attribute level across all products, and the only difference in the Web/app/ped attribute is the addition (or removal) of a pedometer from the various product offerings. The two cheapest products use a dial input, while the remaining products all use a touchscreen. Additionally, there is some laddering of the Storage attribute—it increases with larger display sizes and is used to help differentiate one of the mid-size display variants.

Looking at the prices of this product line it can be seen that there is focus placed on the "middle" of the market. That is, there is one lower-end product ($128, capturing 11.81% share) and few upper end products ($350+). Rather, each design is differentiated from the addition/removal of a pedometer and changing the background color and overlay associated with the case. Background color and overlay are particularly attractive features on which to offer variety, as they require very little engineering rework to change.

Table 5. Design Configuration of the 10 Products Represented by Point B in Figure 7

	Photo/video/camera	Web/app/ped	Input type	Display size	Storage	Price (in $)	Share (in %)
P1	Photo, video and hi-res camera	Web and app only	Dial	1.5 in diag	16 GB	$128	11.81%
P2	Photo, video and hi-res camera	Web and app only	Dial	4.5 in diag	16 GB	$178	8.12%
P3	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$202.60	9.97%
P4	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$205.10	7.85%
P5	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	16 GB	$235.10	5.20%
P6	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	16 GB	$240.10	13.39%
P7	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	32 GB	$335.10	5.81%
P8	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	32 GB	$318.40	4.65%
P9	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	32 GB	$353.40	3.69%
P10	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	32 GB	$355.90	4.02%

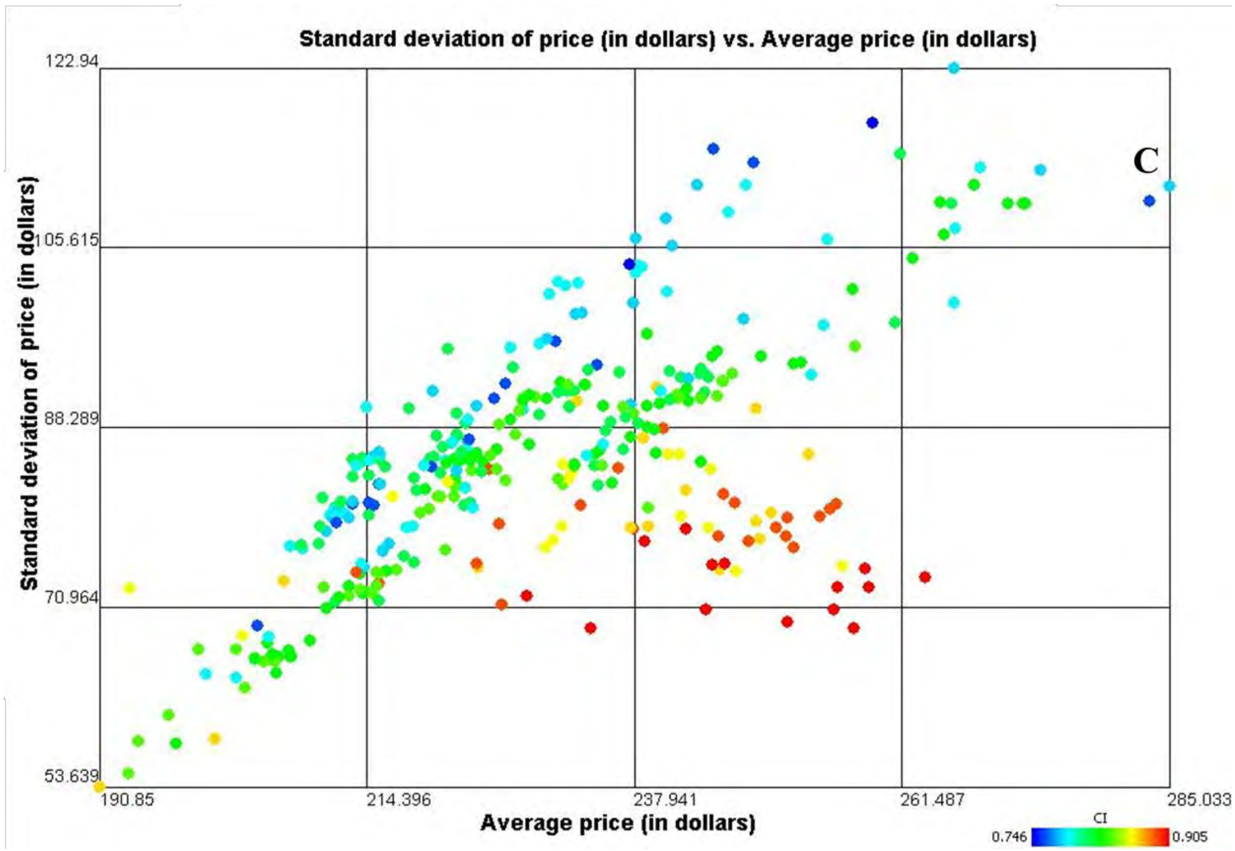
Noticing that the configurations represented by Point B offered very few high-end options the solutions found in Figure 7 were explored further. The first step of this analysis was to calculate the average price and standard deviation of price of the product line. The results from the three-dimensional optimization problem given by Equation 5 were then plotted on these axes, as shown in Figure 8. Design strings with CI values closer to 1 had a smaller variation in average price. This is to be expected as higher degrees of commonality in a product line limit the number of ways that the products can be differentiated.

Point C was then selected for detailed analysis as it had the highest average product line price, variation in price throughout the line, and a lower level of commonality. This solution was found to focus more on the extremes of the market. Two products were targeted at lower-end users. Four of the ten products targeted higher-end users, as they had product prices above $350. This product line also used four different display sizes and even a quick look at the results in Table 6 shows potential challenges in the final solution.

- In the photo/video/camera attribute the camera is either removed entirely (P6) or a lo-res camera is used (P10), while the rest use the hi-res camera capable of photo and video;
- The pedometer is either added or removed from the ten products;
- Only one product (P3) uses a display size of 3.5 in diag;

- Storage options are generally well-laddered, except for one product (P3) that is smaller in display than the most of the other products but also more expensive;
- Product P8 is estimated to capture only 1.56% of the market, raising questions about whether it should actually be offered.

Figure 8. Exploring Trends in Product Line Price as a Function of CI



When discussing the results of this study with Chris Chapman of the conference steering committee, he raised a significant point that share of preference estimates have been shown to be unreliable when dealing with product alternatives that show a great deal of similarity. This raised concerns that the outcomes shown in the previous figures could be an artifact of, or artificially influenced by, the logit calculation. One approach to handle product similarity is to estimate similarity from respondents' answers and use this estimate to adjust the covariance structure in hierarchical Bayesian estimation (Dotson et al., 2009). However, product similarity determined in this way is not necessarily closely tied to the engineering attributes of a product, so we investigate CI as a simpler structure that directly reflects engineering elements.

To provide a greater robustness in market simulation, Sawtooth Software advocates the use of HB and Randomized First Choice (RFC) to simulate respondent choice in the hypothetical market (Huber et al., 1999). However, RFC can be computationally expensive so the optimizations were re-run using a First Choice decision rule. The results of these simulations are discussed in the next section.

Table 6. Design Configuration of the 10 Products Represented by Point C in Figure 8

	Photo/video/camera	Web/app/ped	Input type	Display size	Storage	Price (in $)	Share (in %)
P1	Photo, video and hi-res camera	Web and app only	Dial	1.5 in diag	16 GB	$125.50	10.34%
P2	Photo, video and hi-res camera	Web and app only	Dial	1.5 in diag	16 GB	$128.00	9.38%
P3	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	3.5 in diag	64 GB	$383.63	8.33%
P4	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$200.50	11.76%
P5	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$203.00	4.34%
P6	Photo and video only	Web, app, and ped	Touchscreen	4.5 in diag	32 GB	$324.50	8.91%
P7	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	64 GB	$375.50	5.57%
P8	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	32 GB	$322.40	1.56%
P9	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	32 GB	$352.40	8.99%
P10	Photo, video and lo-res camera	Web, app, and ped	Touchscreen	6.5 in diag	64 GB	$434.90	4.13%

4. Commonality as an Objective—CI with First Choice

To explore solution behavior under a First Choice analysis when commonality is treated as an objective, the optimization problem posed in Equation 5 was re-solved. The two-dimensional projection of the solution space is shown in Figure 9. Like the results presented in Figure 7, there are distinct performance bands associated with varying levels of CI. However, unlike Figure 7 the solutions found in Figure 9 do not show the same "smooth" curve that represented the tradeoff between market share of preference and profit when using a share of preference analysis. For the solutions found during the multi-objective optimization, the value of CI ranged from 0.81 to 0.937. The first solution explored was the solution that captured the largest market share. As shown in Table 7, this product line used only three display sizes. Two of the products, *P5* and *P6* captured 1.46% and 0.49% of the market. To reduce the amount of information shown in Table 7, these products have been removed as they likely would not be offered.

Figure 9. Two-Dimensional Projection of the Solution Space When Using a First Choice Analysis

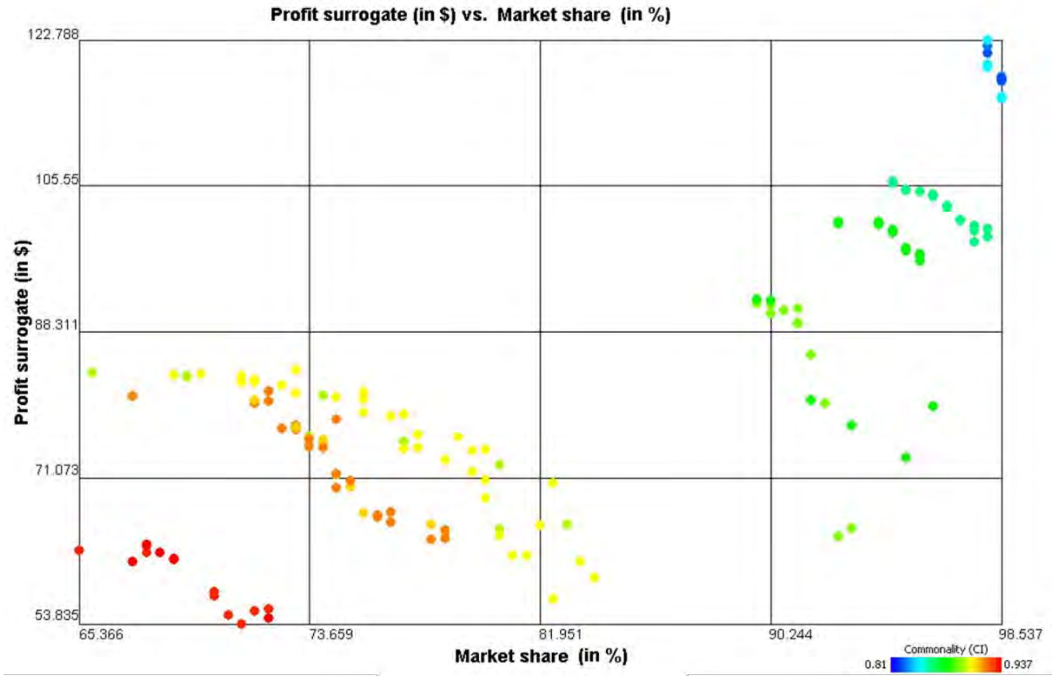


Table 7. Design Configurations for the Maximum Market Share Solution Using FC

	Photo/video/camera	Web/app/ped	Input type	Display size	Storage	Price (in $)	Share (in %)
P1	Photo, video and hi-res camera	Web and app only	Dial	1.5 in diag	16 GB	$132.28	21.46%
P2	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	1.5 in diag	16 GB	$201.28	6.83%
P3	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$202.00	26.83%
P4	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$208.78	12.68%
P7	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	64 GB	$423.58	8.29%
P8	Photo and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	160 GB	$460.08	7.32%
P9	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	64 GB	$417.58	4.39%
P10	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	160 GB	$479.55	8.78%

Examining the product configurations shown in Table 8, the following conclusions can be drawn:

- The configuration of P8 would likely be modified so that video capabilities are included. This would make the selected attribute level common across all products for the Photo/video/camera attribute.
- The inclusion/exclusion of the pedometer is how the optimization algorithm achieves a degree of horizontal product segmentation.
- Two products with the smallest display size are created—one with a dial and one with a touchscreen. However, the price and share estimates for these products suggest that they both could be offered.
- Only three display options are used, and they are the ones most often used when using a share of preference analysis.
- There is a more intuitive product laddering structure for storage size that establishes a vertical segmentation strategy.
- The product prices found using the FC analysis are significantly higher than those found using a share of preference analysis. Storage size increases appear to be driving these larger prices.

Exploring a maximum commonality solution illustrates a slightly different behavior. As shown in Table 8, there is complete commonality in the Photo/video/camera and Input type attributes. For each of the two Display types, two distinct segments can also be created. While these sub-segments are primarily driven by the choice of Storage size, the inclusion of a pedometer can also be used to offer product distinctiveness. Finally, within each sub-segment the configuration of the products is the same for the first five (engineering-driven) attributes. For these sub-segments, variability is achieved using changes to the background color and overlay, with small changes in price accompanying these modifications.

When used as a performance measure, the inclusion of commonality allows for a richer understanding of the tradeoff between business goals. If the CI value for a product line is close to 1, there is often little room for variability in product price. Such product lines will target a specific portion of the market—in this example products between $200 and $300—and use the limited amount of variability to differentiate the designs. As the value of CI gets closer to 0, product configurations will be found that also target the lower and upper ends of the market. These product lines will have more "extreme" products and the content structure of the solution may not be intuitive.

A limitation of using the CI is that it only provided a benchmark value that can be used to compare solutions. Even when the CI approaches 1, commonality is still not strictly enforced in the solution structure. Research in product family optimization offers product formulation strategies to address this by encoding commonality as a design parameter. This encoding is discussed in the next section.

Table 8. Design Configurations for the Maximum Commonality Solution Using FC

	Photo/video/camera	Web/app/ped	Input type	Display size	Storage	Price (in $)	Share (in %)
P1	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$177.70	20.98%
P2	Photo, video and hi-res camera	Web and app only	Touchscreen	4.5 in diag	16 GB	$182.75	15.61%
P3	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	64 GB	$318.20	2.44%
P4	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	64 GB	$330.70	5.85%
P5	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	4.5 in diag	64 GB	$335.75	13.66%
P6	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	16 GB	$213.70	5.85%
P7	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	16 GB	$248.75	3.90%
P8	Photo, video and hi-res camera	Web and app only	Touchscreen	6.5 in diag	64 GB	$336.70	2.44%
P9	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	64 GB	$354.20	0.98%
P10	Photo, video and hi-res camera	Web, app, and ped	Touchscreen	6.5 in diag	64 GB	$366.70	4.88%

5. Commonality as a Design Parameter—Product Family Optimization

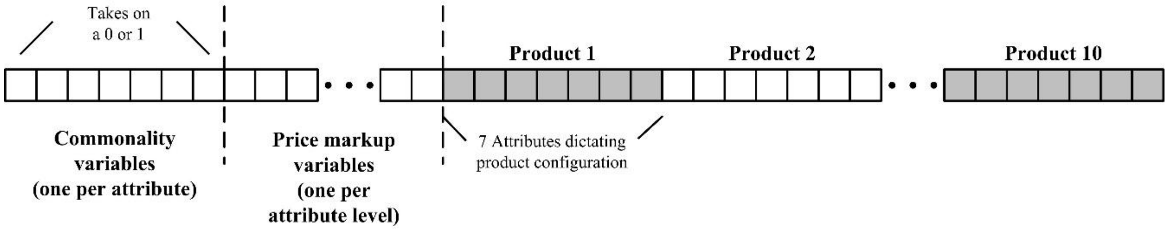
Product families are groups of related products that are created from a set of common components, modules, or subsystems. The common components amongst the related products are referred to as the product platform. Referring back to Figure 5, Screen A and Input A would be considered the product platform for Product line 1 as they are common across all three variants.

The non-trivial nature of defining both the platform and the individual product configuration has led to significant research within the engineering design community. Simpson (2005), for example, reviews over forty such approaches. The approach presented in this paper introduces a way to introduce commonality as a design parameter in the form of restricted commonality (Khajavirad et al., 2009). In a restricted commonality formulation, component sharing is limited to all-or-nothing. This means that a component is either common throughout the entire family or it is allowed to be unique amongst all variants. Restricted definitions of commonality are often a simplifying assumption to reduce computational complexity.

Early efforts solving restricted commonality problems divided the nature of the problem into two stages—platform definition and product configuration. In the first stage, the optimization algorithm would define which variables should be part of the platform (platform definition). The second stage would then solve for the optimum product configurations by determining the extent of variety needed in the remaining attributes (product configuration). However, research has shown that dividing the problem into multiple stages can lead to sub-optimal solutions (Messac et al., 2002).

Single-stage approaches typically require the modification of heuristic optimization approaches. As shown in Figure 10, commonality variables can be introduced into the design string of a genetic search by adding one gene for each product attribute considered. These genes have a binary property. If the commonality gene is set to 0 for Attribute 1, then all products in the line can set their own attribute level for Attribute 1. Conversely, if the gene is set to 1, all products in the line have a common Attribute 1. For coding purposes, the common attribute level is often established by that of Product A.

Figure 10. Encoding Scheme for an All-or-Nothing Commonality Variable


Takes on
a 0 or 1
Product 1
Product 2
Product 10
Commonality variables (one per attribute)
Price markup variables (one per attribute level)
7 Attributes dictating product configuration

Since commonality is no longer being treated as a performance measure, the optimization problem can be reduced to two objectives; market share and profit. When formulating the MP3 problem there are 123 design variables to control—7 for the commonality variables, 46 for the price markup variables, and 70 for feature content. A multi-objective genetic algorithm was used to find the non-dominated solution set for the problem formulation given in Equation 6, and the performances of these solutions are shown in Table 9.

Maximize:	*Market share using first choice analysis* *Profit (given by Equation 3)*	
by changing:	*Feature content* *Commonality variables* *1 < Price markup variable for each attribute level < 2*	(6)
with respect to:	*\$49 < Price for each product < \$699* *No identical products in the product line* *Lower and upper level bounds on each attribute*	

Table 9. Product Family Optimization Results Using Formulation in Equation 6

Solution	Market share (in %)	Profit (in $)	Number of platformed variables in solution	CI
1	99.02%	$68.44	0	0.683
2	98.54%	$115.27	0	0.651
3	98.05%	$118.82	0	0.635

The business performance of the solutions in Table 9 mimic those found in the upper-right hand corner of Figure 9. In both simulations a First Choice rule was applied to respondent selection, and solutions were found that exploited the price markup variables to find configurations for which people would be willing to pay. This allowed for both profit and market share to be maximized in an almost cooperative manner. The algorithm was able to exploit the design space in this manner because there was no penalty for not (or advantage to) embracing commonality.

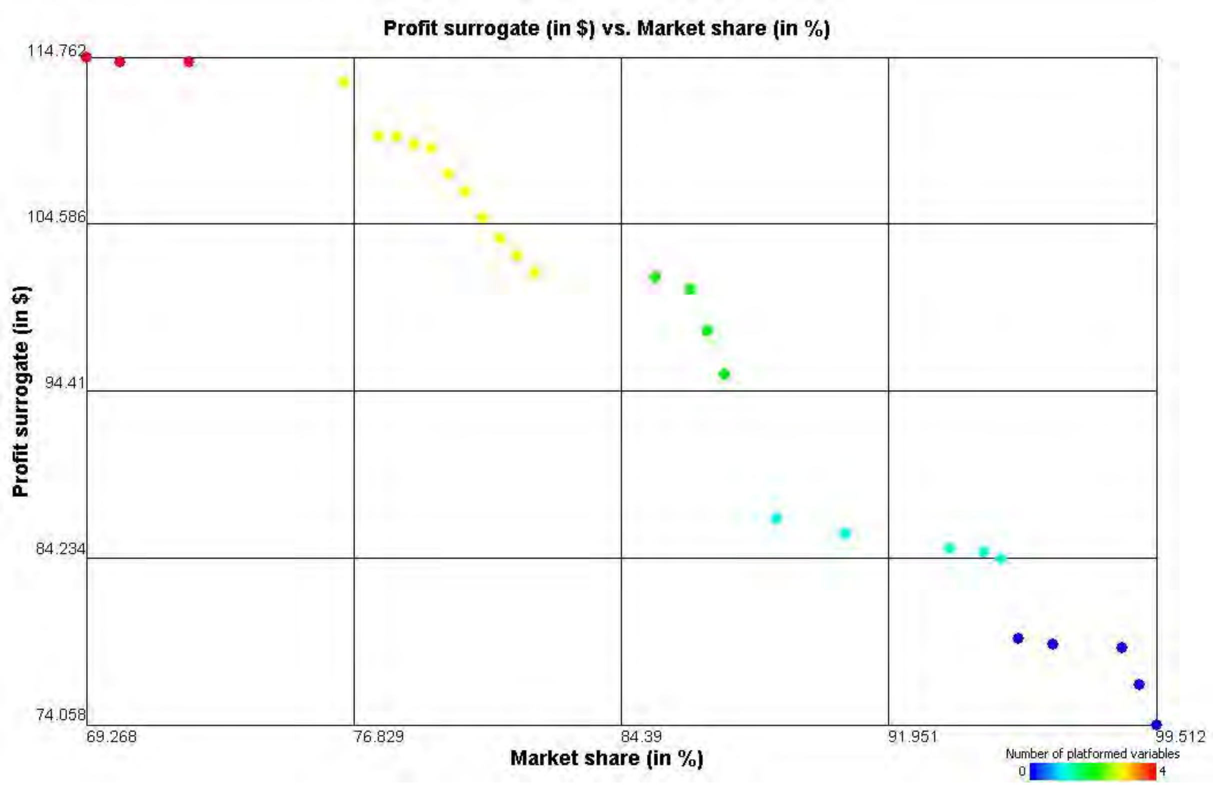
In the previous sections of this paper, commonality was explored using a benchmark measure. While it was possible to quantify the commonality within a product line, the CI number does not provide insight into how that commonality is achieved. Rather, all a larger CI value tells you is that the number of unique parts has been reduced. Conversely, by including the commonality variables as a design parameter, commonality within an attribute can be strictly enforced. This knowledge allows for the optimization problem to be reformulated to provide a benefit for embracing commonality. As shown in Equation 7, the cost of each product is now reduced by 10% for each platformed (common) attribute.

Maximize: *Market share using first choice analysis*
Profit (given by Equation 3)
by changing: *Feature content*
Commonality variables
1 < Price markup variable for each attribute level < 2
(7)
with respect to: *$49 < Price for each product < $699*
No identical products in the product line
Lower and upper level bounds on each attribute

cost reduction: *10% cost reduction per product for each platformed attribute*

The results of this multi-objective optimization are shown in Figure 11. This figure shows that distinct performance clusters are possible depending on the amount of commonality that is embraced. In cases where there are no strictly platformed variables (all commonality variables in the design string have a 0), the solutions are able to exploit the product price and variety to capture the largest amounts of market share. As the number of attributes in the product family increase, the lack of possible variety leads to a reduction in share as respondents choose either the competitor product or the outside good. However, the cost reductions associated with increased platforming allow greater overall profits to be achieved. The product family strategy adopted by a company requires understanding and navigating conflicting business goals.

Figure 11. Business Impact of Increased Product Platforming When Cost Reductions Are Modeled



Product line solutions with the least number of platformed variables made either the Background color or the Background overlay the common component. Increased product platforming saw the Background overlay and Input type made common across the line, followed by Screen size, and then Web/app/ped. None of the solutions found in this optimization platformed the Camera, Display Size or Storage attributes. As seen in the previous sections, these attributes were often used to expand product variety to hit different areas of the market. The information presented in Table 10 helps confirm these results as those attributes are used to reach both horizontal and vertical segmentations of the market.

Additionally, when using a product family, each attribute maintains the same active attribute level across all solutions:

- Background color = Silver (3 solutions)
- Background overlay = Custom pattern and graphic when used alone (2 solutions), Custom graphic when used with other attributes (18 solutions)
- Input type = Touchscreen (18 solutions)
- Screen size = 4.5 in diag (14 solutions)
- Web/app/ped = Web, app and ped (3 solutions)

Table 10. Exploring the Expansion of the Product Family

Product family extent	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1 Platformed variable (5 solutions)	Background color (3)			
	Background overlay (2)			
2 Platformed variables (4 solutions)	Background overlay	Input type		
3 Platformed variables (11 solutions)	Background overlay	Input type	Screen size	
4 Platformed variables (3 solutions)	Background overlay	Input type	Screen size	Web/app/ped

6. Conclusions and Future Work

The results presented in this paper build on previous outcomes that have been seen in the author's previous work and other market research literature that explores the formulation of product configuration problems and the application of heuristic optimization techniques (Green and Krieger, 1985; Balakrishnan et al., 1996; Besharati et al., 2006; Belloni et al., 2008; Wang et al., 2009). Significant increases in computing power and the availability of code—both in commercial software and as open-source packages—have made large product design problems tractable. However, these algorithms are not necessarily as fire-and-forget as they may seem. The distribution of product content and content laddering within a product line must make intuitive sense to managers (who are devoting resources to realize the product line) and customers (who must make a purchasing decision).

Results presented in this paper show that special attention must be spent when formulating product line optimization problems. Poorly posed problems can provide a space that can be exploited by the optimization algorithm, yielding solutions that may hold in simulation but offer little practical viability. Sections 3–5 of this paper demonstrate two different strategies for integrating commonality decisions into the formulation of the design problem. The Commonality Index was demonstrated as a useful benchmarking tool that could easily be incorporated into a problem's performance space by treating it as an additional objective. By doing so, a richer understanding of the trades that needed to be made between conflicting business goals could be realized.

An advantage of the Commonality Index is that it is easy to calculate and does not require detailed information associated with component costs, manufacturing details, etc. However, the CI can only provide a benchmark value; changes in CI are directly tied into more/less unique components being used. Detailed commonality information is not easily extracted. Because of the nature of this measure it can be difficult to model cost savings and other design/manufacturing efficiencies as a function of CI.

The second approach demonstrated in this work requires a reformulation of the design problem so that a vector of commonality variables can be incorporated into the genetic string. An advantage of this approach is that restrictive commonality can be easily modeled and estimates of cost savings and other efficiency advantages can be more directly modeled. The core architecture of the product line is simultaneously discovered along with the configuration of individual products, resulting in increased computational expense. However, incorporating commonality as a design parameter provides a more tunable approach that can be extended to handle multiple platforms and various architectures.

Throughout this work, solutions with greater commonality were shown to map to less extreme products and often catered to middle segments of a respondent market. The importance of simulation technique—particularly a logit decision rule versus a First Choice decision rule—was also explored. This consideration is particularly important in these simulations because of the high degree of similarity between product configurations. Such similarities can unfairly influence the estimated performance of a product line.

Trades between market share and profit were seen for both share of preference and First Choice simulations. When designing a product family using the problem formulation given by Equation 6, this trade was not as clear. Richer problem formulations should consider the challenges associated with build complexity and issues that might arise in the supply chain.

Future work in this area should aim to formally prove the outcomes presented in this paper. Additionally, there is a need to better understand how optimization outcomes are influenced by the choice rule followed. Figures 7 and 9 illustrate the different tradespace representations offered by each simulation type and illustrate the impact that the tools available have on possible outcome. Finally future work should explore the most effective way to structure the optimization problem for multiple platform possibilities.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from the National Science Foundation through NSF CAREER Grant No. CMMI-1054208. Any opinions, findings, and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Scott Ferguson

References

Balakrishnan, P. V., Gupta, R., and Jacob, V. S., 1996, "Genetic Algorithms for Product Design," *Management Science*, **42**(8): 1105–1117.

Belloni, A., Freund, R. M., Selove, M., and Simester, D., 2008, "Optimal Product Line Design: Efficient Methods and Comparisons," *Marketing Science*, **54**(9): 1544–1552.

Besharati, B., Luo, L., Azarm, S., and Kannan, P. K., 2006, "Multi-Objective Single Product Robust Optimization: An Integrated Design and Marketing Approach," *Journal of Mechanical Design,* **128**(4): 884–892.

Chapman, C., and Alford, J., 2011, "Product Portfolio Evaluation Using Choice Modeling and Genetic Algorithms," *Proceeding of the 2010 Sawtooth Software Conference*, Newport Beach, CA.

Collier, D. A., 1981, "The Measurement and Operating Benefits of Component Part Commonality," *Decision Sciences*, **12**(1): 85–96.

Daskilewicz, M. J., and German, B., J., 2011, "Rave: A Computational Framework to Facilitate Research in Design Decision Support," *Journal of Computing and Information Science in Engineering,* **12**(2): 021005:1–9.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., 2002, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, **6**(2): 182–197.

DeWalt, 2015, http://www.dewalt.com/tool-categories/cordless-drills.aspx

Dotson, J., Brazell, J. D., Howell, J. R., Lenk, P., Otter, T., MacEachern, S. N., and Allenby, G. M., 2009, "A Probit Model with Structured Covariance for Similarity Effects and Source of Volume Calculations" Available at SSRN: http://ssrn.com/abstract=1396232 or http://dx.doi.org/10.2139/ssrn.1396232.

Ferguson, S., and Foster, G., 2013, "Demonstrating the Need and Value of a Multiobjective Product Search," *2013 Sawtooth Software Users Conference*, October 14–18, Dana Point, CA.

Green, P. E., and Krieger, A. M., 1985, "Models and Heuristics for product Line Selection," *Marketing Science*, **4**(1): 1–19.

Huber, J., Orme, B. K., and Miller, R., 1999, "Dealing with Product Similarity in Conjoint Simulations," *1999 Sawtooth Software Conference Proceedings*, San Diego, CA, pp 253–66.

Khajavirad, A., Michalek, J. J., and Simpson, T. W., 2009, "An Efficient Decomposed Multiobjective Genetic Algorithm for Solving the Joint Product Platform Selection and Product Family Design Problem With Generalized Commonality," *Structural and Multidisciplinary Optimization*, **39**: 187–201. DOI 10.1007/s00158-008-0321-9.

Martin, M. V., and Ishii, K., 1996, "Design for Variety: A Methodology for Understanding the Costs of Product Proliferation," *Proceedings of the 1996 ASME Design Engineering Technical Conferences*, Irvine, CA, DTM-1610.

Martin, M. V., and Ishii, K., 1997, "Design for Variety: Development of Complexity Indices and Design Charts," *Proceedings of the 1997 ASME Design Engineering Technical Conferences*, Sacramento, CA, DFM-4359.

Matlab, the Mathworks, 2014, Matlab 2014a.

Messac, A., Martinez, M.P., Simpson, T. W., 2002, "Effective Product Family Design Using Physical Programming," *Engineering Optimization,* **34**(3): 245–261.

Mulhern, M. G., 2007, "Determining Product Line Pricing by Combining Choice Based Conjoint and Automated Optimization Algorithms: A Case Example," *Proceedings of the 2007 Sawtooth Software Conference*, Santa Rosa, CA, 271–278.

RCR Yachts, 2015, http://www.rcryachts.com/new

Sawtooth Software, 2003, "Advanced Simulation Module for Product Optimization v1.5 Technical Paper," Sequim, WA.

Sawtooth Software, 2009, "The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper," Sawtooth Software, Inc., Sequim, WA, http://www.sawtoothsoftware.com/download/techpap/hbtech.pdf.

Thevenot, H. J., and Simpson, T. W., 2006, "Commonality Indices for Product Family Design: A Detailed Comparison," *Journal of Engineering Design, **17**(2):* 99–119.

Turner, C., Foster, G., Ferguson, S., Donndelinger, J., and Beltramo, M., 2012, "Creating Targeted Initial Populations for Genetic Product Searches," *2012 Sawtooth Software Users Conference*, Orlando, FL.

Turner, C., Foster, G., Ferguson, S., Donndelinger, J., 2014, "Creating Targeted Initial Populations for Genetic Product Searches in Heterogeneous Markets," *Engineering Optimization*, **46**(12): 1729–1747, doi. 10.1080/0305215X.2013.861458.

Stump, G., Yukish, M., Martin, J., and Simpson, T., 2004, "The ARL Trade Space Visualizer: An Engineering Decision-Making Tool," *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY, AIAA-2004-4568.

Stump, G. M., Lego, S., Yukish, M., Simpson, T. W., and Donndelinger, J. A., 2009, "Visual Steering Commands for Trade Space Exploration: User-Guided Sampling with Example," *Journal of Computing and Information Science in Engineering*, **9**(4): 044501:1–10.

Simpson, T. W., 2005, "Methods for Optimizing Product Platforms and Product Families: Overview and Classification," In: Simpson, T., W, Siddique, Z., and Jiao, J. (eds), Product Platform and Product Family Design: Methods and Applications. Springer, New York, pp 133–156

Wang, X., Camm, J., and Curry, D., 2009, "A Branch-and-Price Approach to the Share-of-Choice Product Line Design Problem," *Management Science*, **55**(10): 1718–1728.

A MACHINE LEARNING APPROACH TO CONJOINT ANALYSIS: BOOSTING AND BLENDING ENSEMBLES

KEVIN LATTERY
SKIM GROUP

1.0 INTRODUCTION

Interpretable models have been a centerpiece of classical predictive modeling. For instance, regression coefficients have meaning and we can tell from them how much change will result from changes to specific variables. But with the rise of computers and the field of machine learning a new kind of predictive modeling is also being done. This new approach to building predictive models no longer cares whether the model is interpretable or corresponds to human psychology. All that matters is that one can build a *computer algorithm* to make accurate predictions. For instance, the way a computer recognizes text or visual patterns is very different from the way humans do it. And in machine learning that is just fine. One only cares that the computer has some algorithm that makes accurate predictions.

One of the success stories from machine learning is the use of ensembles. An ensemble approach first generates multiple diverse models. Each specific model makes predictions on its own. The models are diverse in the sense that each model makes different predictions, each with its own unique strengths and weaknesses. One takes the diverse models and then blends the predictions to reduce bias from any one model and generate more robust and accurate out-of-sample predictions.

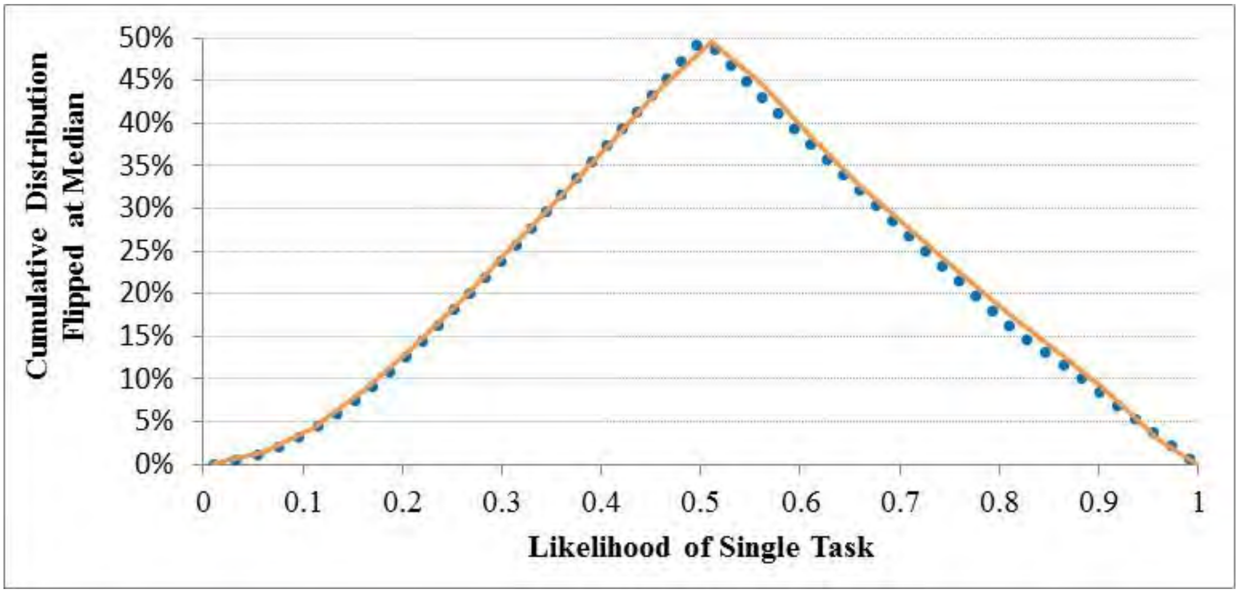
Predictive ensembles originated with classification trees. Rather than a single classification tree, ensembles use a collection of trees. Averaging over the predictions of all the trees is more robust than a single tree and tends to improve predictive accuracy significantly. This ensemble of trees was called Random Forests. A later development was Boosted Trees, which created different classification trees using an adaptive algorithm, rather than random generation.

The Netflix competition is one of the most famous success stories of using ensembles for prediction. For each Netflix subscriber the Netflix algorithm could predict the member's rating of other movies not seen by them. Netflix sponsored a competition to better predict respondent ratings and enable Netflix to make better movie recommendations. The Netflix prize was substantial: $1 million dollars to the best predictive algorithm, with a minimum requirement of 10% improvement over their current algorithm. The winner, as well as all of the leading methods, was an ensemble approach. In fact, the competition became dominated by teams that grew in size as they brought in diverse models developed by other competitors. The winning methods employed hundreds of models created by diverse people which were blended.

This paper applies a similar ensemble approach to conjoint analysis. We demonstrate the power of blending diverse conjoint models, and show how this ensemble approach can significantly improve our predictions. Of course, there are many different ways to build an ensemble of predictive models. So this paper only touches upon a broad area of research. We have no doubt that others can develop even better ensembles and further improve their predictive powers.

1.1 A Simple Example

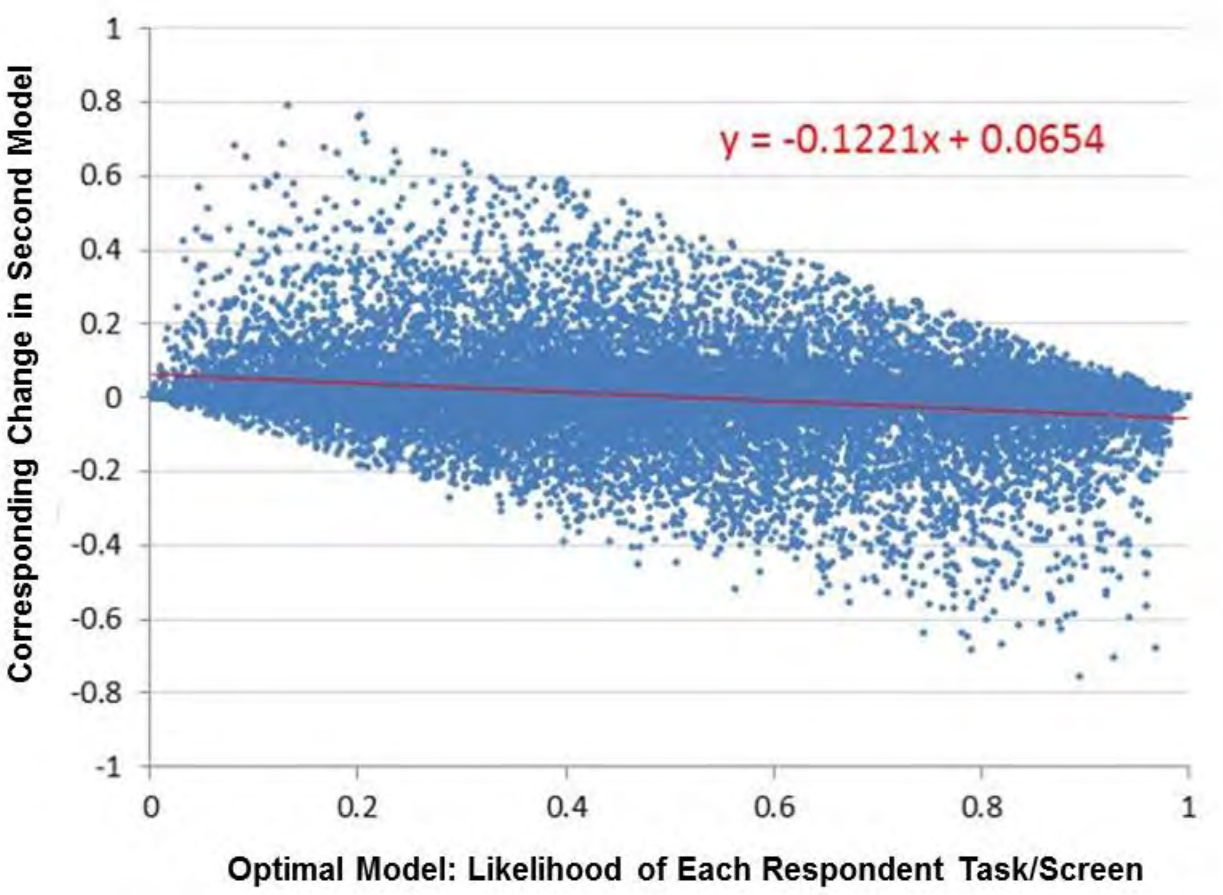
To show how and why ensemble methods work, let's look at a simple example. In this case we have a sample of 1500 respondents, with each respondent doing 12 conjoint tasks (each task has 5 alternatives). The best 30-segment latent class solution from LC Gold Choice has a total LL of -14,915. But one can find many other great latent class models that fit almost as well. Below we show another latent class model also with 30 segments. It has a slightly poorer fit with a total LL of -14,999. The chart below shows the distribution of the likelihood for each of these two models: the solid line is slightly better than the dashed line.



Note: this and later graphs show a cumulative distribution function, but "flip" it vertically when it reaches 50% (the median). The sharp peak in the middle is the result of the flip; the overall curve is roughly sigmoidal as for many cumulative distribution functions.

This chart clearly shows the overall *distribution* of errors is quite similar across these two models. Sure, the solid line is slightly better, but both models have about the same number of conjoint tasks with a given likelihood.

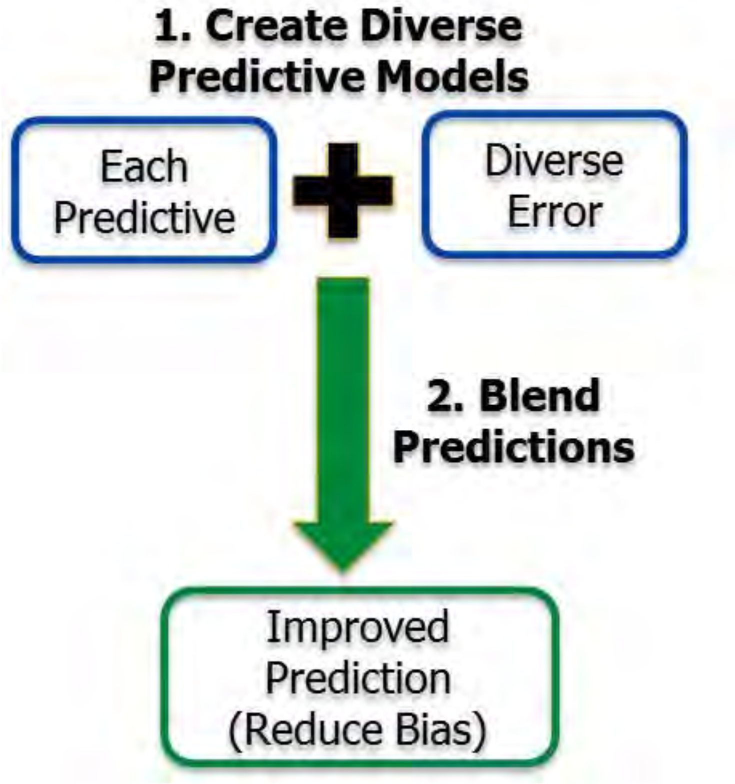
But now let's turn from the overall distribution of errors to a direct comparison between the two solutions on each specific task. The scatterplot on the next page shows the fit of the optimal solution on the x-axis for each of the 18,000 tasks (1500 respondents x 12 tasks). The y-axis shows the difference in fit for the second model on that same exact task (one respondent doing one task).



What this chart shows is the *two models have different strengths and weaknesses*. This is the key power of ensembles. The left side of the chart shows tasks which are fit poorly by the optimal model are fit better by the second model. By combining the two models one improves predictions on the poorly fit tasks in the optimal model. Of course, the combination also reduces the fit of those which were fit very well. The net result is a reduced variance in the error across tasks, which should translate to more robust predictions and better out-of-sample prediction. This is true even though the overall distribution of error is nearly identical.

1.2 Basic Overview of Ensembles

The overall flow of predictive ensemble methods is rather simple:


1. Create Diverse Predictive Models
Each Predictive
Diverse Error
2. Blend Predictions
Improved Prediction (Reduce Bias)

The first step is to create diverse predictive models. It is important that the models differ from each other in their errors. Models with the same error will make the same predictions and do not add any new information as part of the ensemble. However, we still want each model in the ensemble to be predictive, meaning they are accurate. The ideal scenario is highly predictive models that are diverse.

The second step is to blend these diverse models. Blending is done by *combining the predictions*. It is not about creating some single master model. The simplest form of blending is averaging. This amounts to simply taking the predictions of each model and averaging them. More complex blending schemes weight each model differently (model 5 has more weight than model 3). The most complex blending schemes give different models different weights by respondents (for me, model 3 has more weight than model 5).

The first part of this paper will focus on how to create diverse predictive models. There are two general approaches to creating diverse models in an ensemble. The first way is to generate diverse models independently. The second approach is to create models that are complementary to each other, using knowledge of previous ensemble members. For example, based on the first two models, one can create a third that complements their combined weaknesses. This is an example of what is called “boosting.” Section 2 will discuss independent models. Section 3 discusses the generation of complementary models via boosting. In sections 2 and 3 the only blending we will do is simple averaging of predictions. Section 4 introduces more complex blending methods. Section 5 looks briefly at how our work can be extended.

2.0 Independent Models

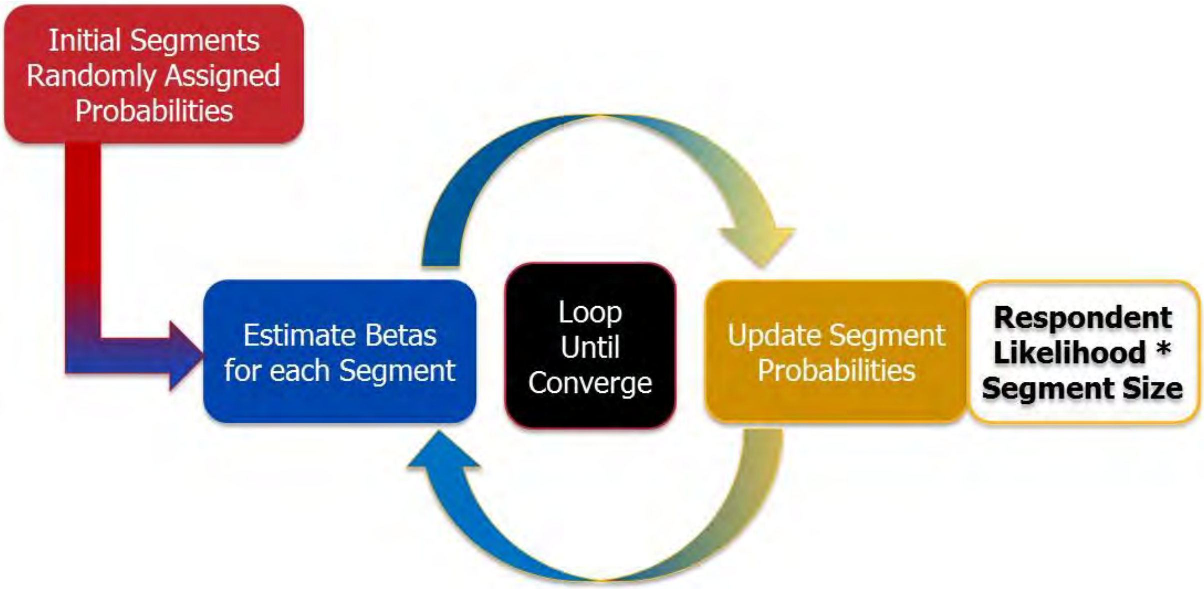
There are many ways to generate diverse conjoint models for a specific data set. For instance, one can use different algorithms (Latent Class, HB, Probit). One can also code the design matrix differently (different ways to code price for example) or use different utility structures (utility maximization or random regret). We have had great success using different nested logit structures in some cases. We have also specified different interactions and different constraints. In this paper we wanted to develop an approach that could be used in any conjoint study. In particular we will focus on developing different latent class models.

Latent class models can be used to fit any conjoint data, except perhaps very small data sets where splitting the data into subgroups would be inappropriate. In addition to being widely applicable, latent class models enable us to generate diverse models easily. The reality is that for any segmentation one derives, there is another segmentation that looks different, makes different predictions, and fits the data nearly as well. For some, the presence of so many great segmentation solutions is a dilemma—which do I choose? But for generating members of an ensemble, the plethora of great segmentations is a blessing.

2.1 Generating Diverse Latent Class Solutions

Latent Class solutions are developed by an iterative procedure. One begins with initial segment solutions. These are just starting seeds, and can be randomly assigned. In the analyses here, we start by assigning each respondent nearly equally to all segments. Specifically we use a random uniform distribution between 1 and 2, and then normalize so the sum is 1. This guarantees that the maximum segment probability for a respondent is no more than twice as high as the minimum. Given the segment assignments, one computes betas for each segment. Then for a specific respondent we compute the likelihood fit of the respondent data to each segment solution. The probability of a respondent being in a segment is then proportional to the likelihood times the size of the segment, and this replaces the initial random probability.

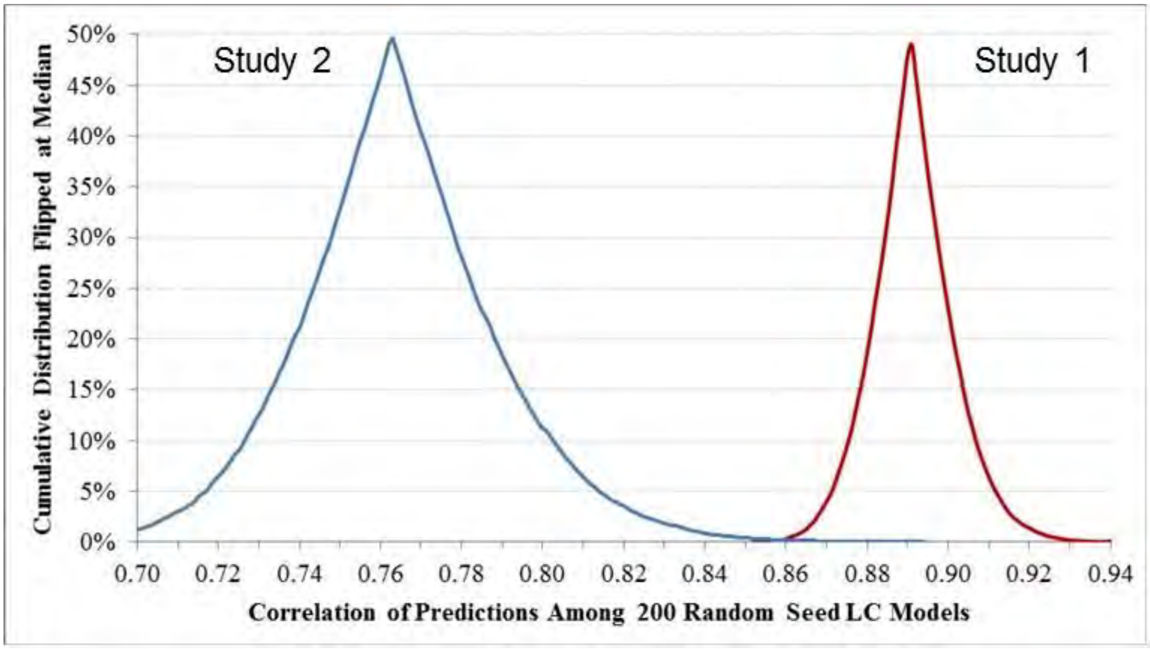
This process iterates as shown in the chart below, until we meet a defined convergence limit:


Initial Segments Randomly Assigned Probabilities
Estimate Betas for each Segment
Loop Until Converge
Update Segment Probabilities
Respondent Likelihood * Segment Size

Each successive iteration typically improves the total log likelihood. Convergence here is considered achieved when the last 10 iterations improve total log likelihood less than .1% (MaxLL/MinLL > .999, where max and min are over last 10 iterations), or some similar rule. After convergence, the iteration with the best fit (typically the last one) is chosen.

This algorithm produces a locally optimal solution (or nearly optimal) based on the starting seeds. Different starting seeds will lead to a different local optimum. In fact, we used 200 different starting seeds to produce 200 different latent class solutions. None of the 200 solutions fit the data as well as LC Gold Choice, which we assume to be the globally optimal solution. And that's alright, because we are interested in blending diverse models rather than finding the single best one.

So how much diversity did we generate? The chart below shows two studies that generated different degrees of diversity. In both cases, the exact same algorithm was used. But Study 1 on the right had far less diversity than Study 2. We chose Study 1 as our first benchmark just because it had much less diversity. We wanted to see how well ensembles work even when there is less diversity.



It is important to recognize that given a specific study, the predictions from different models will have some degree of correlation. After all, the models are predicting the same set of data. So one should not expect correlations near 0. But of course if the correlations are very high, like .95 or higher, meaning there is a lack of diversity, then the ensemble is not adding much value over a single model.

How much diversity is generated depends upon many factors. Based on our experience with about 10 studies, we see the following factors driving diversity within Latent Class solutions.

1. The number of segments in the latent class. Given a specific data set, more segments tends to drive more diversity. With a larger sample, one can define more segments. With more segments, there are likely to be more great latent class solutions. With a

smaller sample, one may only have 2–3 segments per model, and more care must be taken to ensure diversity across different solutions.

2. The number of parameters and alternatives in the conjoint tasks. More parameters and more alternatives generate more diversity. Study 1 had only 18 degrees of freedom, while study 2 had 23. This is not many parameters. We chose these two studies because we wanted to show that one can create diversity even with smaller numbers of parameters to estimate.

2.2 Reducing and Blending Models in Ensemble

As we noted before, 200 Latent Class solutions were generated from different starting seeds. While we could blend all 200 solutions, one of the questions we asked was whether we could reduce the number of models in an ensemble without sacrificing predictive accuracy. We experimented with many methods, including TURF and various optimization procedures. But a relatively simple method worked very well: backward elimination.

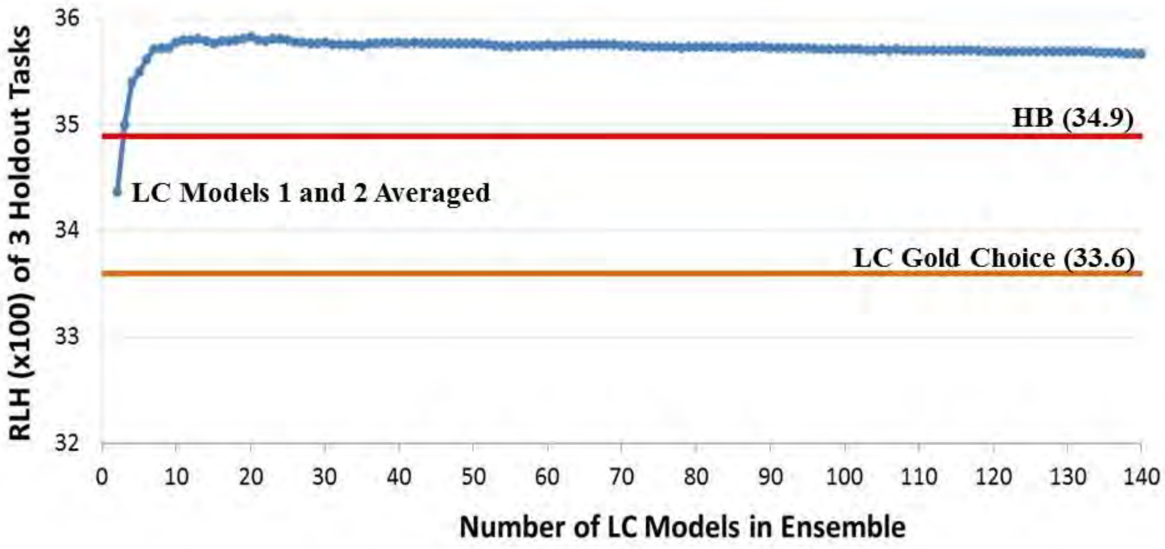
We begin with a 200 x 200 pairwise correlation matrix of the predictions from each of the models. We then select the highest correlation. There will typically be two items with this correlation (but there could be more if there is a tie). We then remove one of the items using a tiebreaker rule. The tiebreaker is the sum of the correlations for that item with all the other items. In particular, the item with the highest correlation sum loses the tiebreaker and is removed. That leaves us with a 199 x 199 correlation matrix. We repeat the process removing the next item.

The result of this backward elimination is a ranking of the models in the ensemble from 2 until 200. Note that stepwise forward model addition would have sought the same objective, but we found it did not work as well. We cumulate the models as if they were developed using stepwise forward selection. The table below shows the correlation among the first five models. This was for Study 1 shown above.

	RLH x 100	Correlation Among Models 1–5				
1	33.0		0.855	0.855	0.864	0.871
2	32.8	0.855		0.857	0.859	0.867
3	33.0	0.855	0.857		0.865	0.862
4	33.4	0.864	0.859	0.865		0.867
5	33.2	0.871	0.867	0.862	0.867	

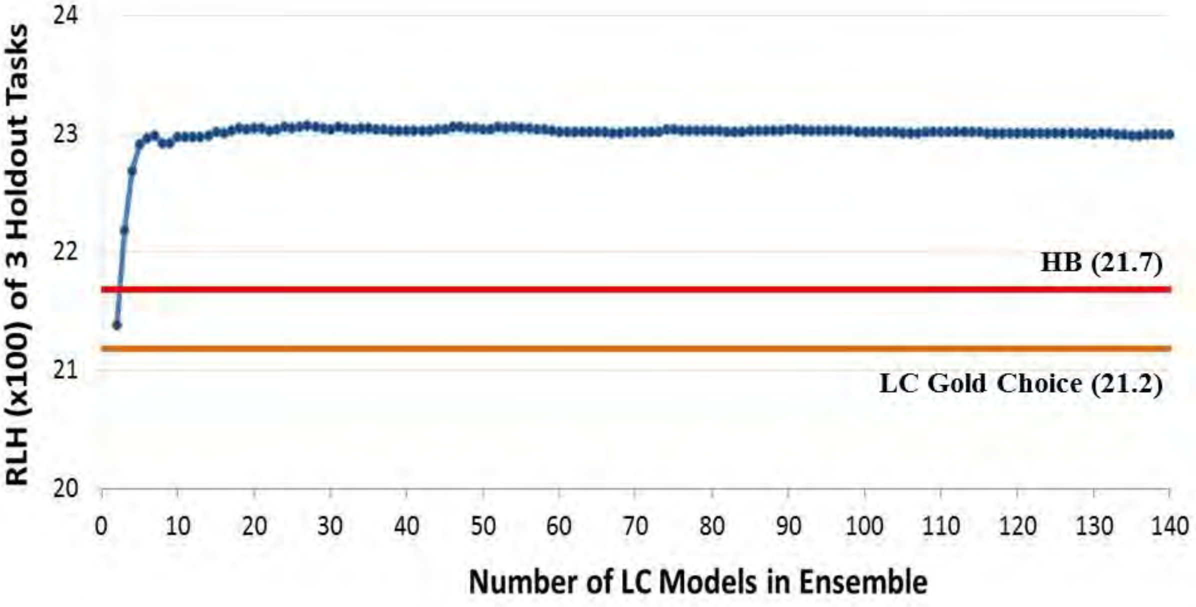
The correlations among these 5 models are relatively low compared with the initial range of .85 to .94. We have also included the RLH in the table to show how well each of these models fit the data. Note that the highest fit here has an RLH of 33.4. This is slightly lower than our analyses of the same data using LC Gold Choice (33.6) and quite a bit lower than HB (34.9).

So we have a collection of slightly poorer models. The magic is in the power of the group. And to show we have nothing up our sleeves, we blended these models using simple averaging of predictions. The chart below shows the impact of cumulative blending.



Blending the first two models from our backward elimination process beats LC Gold Choice. Averaging the first three beats HB. By the time we blend 10 models we have achieved as much predictive accuracy as blending all 200 models. We achieve the highest prediction from blending the first 13 models from our stepwise procedure. Note that as we go across the chart's x-axis we are just successively adding models to the cumulative mix, holding all the previous models constant. We have shown only the first 140 models in the chart above, but the RLH is very flat after that.

The chart below shows similar results for Study 2.



Again we see that HB predicts better at the respondent level than a single LC Gold Choice model, a pattern we typically see. Also, as previously seen, averaging 3 Latent Class models is enough to beat HB. By the time we average 7 models we predict as well as averaging across all

200 models. In general we have found that we can get better respondent level fit than HB with about 3–5 Latent Class models averaged together.

3.0 Creating Diverse Models via Boosting

Our success generating diverse Latent Class models via random seeds coupled with backward elimination was very satisfying. Eager for more, we thought we might achieve even greater success by creating an ensemble of models via boosting. After all, boosting almost always outperforms its random counterparts. In the world of classification trees, Boosted Trees typically predict much better than Random Forests. We expected a boosted version of Latent Class Ensembles to significantly outperform our random seed version.

Boosting is a method for developing models in the ensemble sequentially. So the second model is developed in a way that is complementary with respect to the first model. And this continues for each sequential model, where model n is generated specifically to complement the previous n-1 models. An excellent, very readable text on boosting is *Boosting*, by Robert E. Schapire and Yoav Freund, its inventors.

3.1 Adapting Boosting to Conjoint

AdaBoost and Stochastic Gradient Boosting are two of the more prominent boosting algorithms. Both were developed in the context of decision trees. But the question remained about how to adapt them to work for multinomial logistic conjoint models. We think adapting boosting to conjoint ensembles remains a topic for further research, with more potential than we have currently developed. The next section discusses some of our work on adapting boosting to conjoint.

One approach is to change our non-classification type model into a classification problem. AdaBoost.RT for example is a boosting procedure for general regression that converts regression into classification via a fit threshold. Basically, if the fit of a model to a respondent exceeds a specific threshold value, then we consider that respondent classified correctly (coded as 1), otherwise they are classified incorrectly (coded -1 or 0). So it is simply a matter of picking a good threshold value, and applying boosting to the newly defined classification problem. With an MNL conjoint, we might say, for instance, that if the likelihood for the actual choice is 0.5 or more, the choice is classified as correct.

In our research we estimated the first latent class model. From there we derived a threshold likelihood based on the distribution of the likelihoods across each respondent task. We found the best threshold to be a value where the resulting correct classification rate would be in the range of 60%–75%. Note that for boosting algorithms to work, the correct classification must be at least 50%. There was no specific percentage that worked well across all studies, so we think it best with this approach to test four different threshold values: those that result in 60%, 65%, 70%, and 75% correct classification of the initial Latent Class model.

In addition to setting a fixed threshold value (a likelihood value for the task) based on the first LC model, we tried an adaptive approach where the threshold was varied to keep the correct classification rate (like 65%) constant. Sometimes this offered a slight improvement. For now, we prefer to keep the likelihood threshold value constant across ensemble generation. That said, we are also moving away from the threshold approach, as we have found an alternative approach that works better.

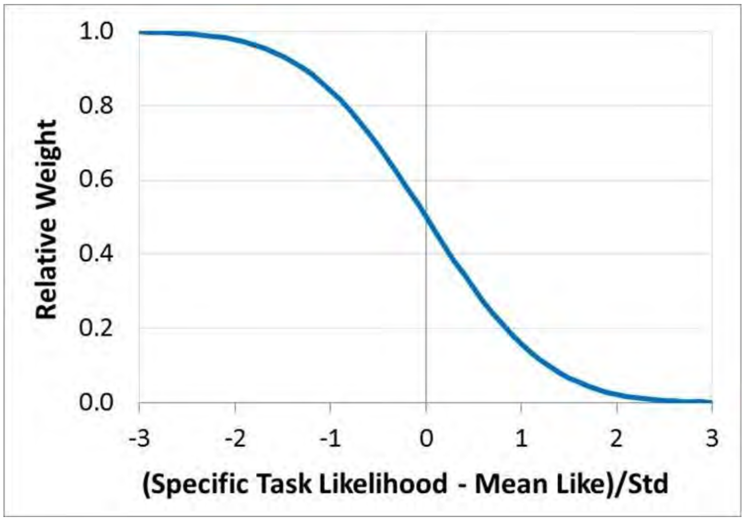
AdaBoost works by reweighting the data. So after a first model is fit, the errors are computed for each case. Cases with more error are given higher weights for the second model. This drives the second model to make different tradeoffs than the first model: fitting cases which were poorly fit at the expense of cases that were fit well. This continues at each sequential stage. The cumulative predictions from accumulating n models are computed, and the data is reweighted based on those errors. Then the (n+1)th model is fit.

We tested several algorithms based on errors rather than thresholds. For instance, the AdaBoost.R2 algorithm works by computing an adjusted error for each specific error e_i on the interval [0,1]. The revised weights for each case are:

$[\bar{e} /(1 - \bar{e})] \wedge (1- e_i)$, where $\bar{e}$ is the mean of the e_i over all respondent tasks, and $\bar{e} < .5$.

Typically, these revised weights are computed based on the errors from the last model and then multiplied by the previous weights. An alternative is to use the error from the cumulative model.

The plethora of adjustments and different functions led us to the following more theoretically sound version: weighting based on the number of standard deviations from the mean. Since we have the likelihood value for each case, we use the fit rather than error. Specifically we compute the likelihood of each specific case given the cumulative model. We then calculate the number of standard deviations from the mean fit (the z-value for each case). The weights are then simply 1 - Cumulative Normal, also known as the Q-function. This is shown in the chart below:



This relatively simple calculation performed at least as well as the many other algorithms we tried, and has a stronger theoretical basis. (Of course if one used error rather than fit, one would use the cumulative normal function which is just the mirror image of this.)

One important detail is that the “individual case” we are reweighting for each successive model is a single respondent task. So if respondent Spock did 12 tasks, he would have 12 weights, one for each task. The alternative is to look at weights for each respondent. So Spock would have only one weight based on the fit across all his conjoint tasks. We found this did not create as much diversity in our ensemble, but we may explore this option more in the future. The advantage of taking a single respondent task as the unit is that Spock's segment assignment in the Latent Class model is based more on those tasks that did not fit well.

One final practical issue with boosting is that conjoint analysis typically has different goals from other areas where boosting is applied. Specifically in conjoint analysis we are interested in respondent level fit *and* aggregate predictions. In contrast, most applications of boosting are concerned only with improving respondent level predictions. A boosted tree prediction is not concerned that across all respondents, 42% of them should choose this. But in a conjoint analysis, we are very concerned with aggregate predictions—they are one of the key outputs. We want to know how many people will choose a specific product configuration vs. other alternatives.

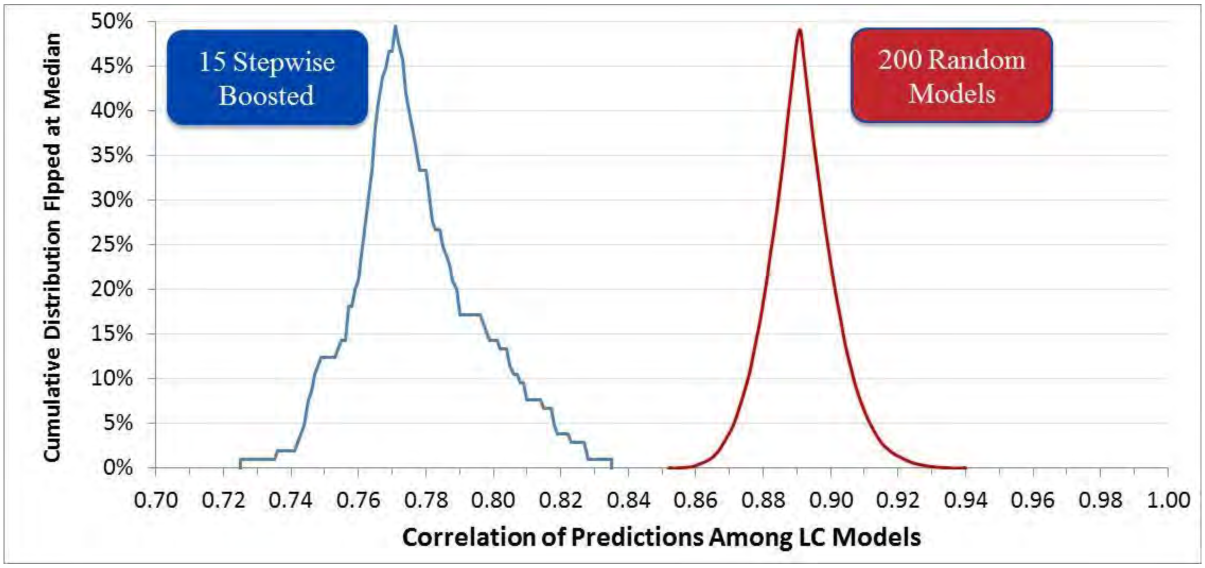
Our initial applications of boosting enabled us to improve respondent level fit, but sacrificed aggregate predictions. The problem is that when we reweight the data in a given stage, we are effectively creating new shares for those conjoint tasks. In our case, we used a fixed blocked design. We observed for instance that 52% of respondents (n = 300) choose alternative one in a specific task, but when we reweighted the data only 38% of respondents choose that alternative. Reweighting the data shifted the aggregate shares, and that carried over into our predictions of aggregate shares.

So the solution we use now is a two-stage approach. When we develop each model in the boosted series, we first fit a latent class model using the boosted weights. This gives us segment membership probabilities for each respondent. We then fit another latent class model. In this second stage however we remove the boosted weights. Of course, we keep the segment membership probabilities—these are the new starting point for the second model. In effect we use boosted weights to create different starting seeds, rather than random generation.

One alternative to this two-stage approach is to do post-hoc adjustments to aggregate shares. We got very accurate aggregate shares when we exponentiated the predicted shares, (Predicted Share)^k, and then repercentaged the results to a sum of 1. We solved for k to minimize MAE of predicted share. We always found $k > 1$, as the boosted ensemble shares tended to be flatter (more nearly equal for all products than they should be). We have not presented the results for this post-hoc share exponentiation in this paper. But we will continue to explore this approach as the improvement in fit was sometimes very significant.

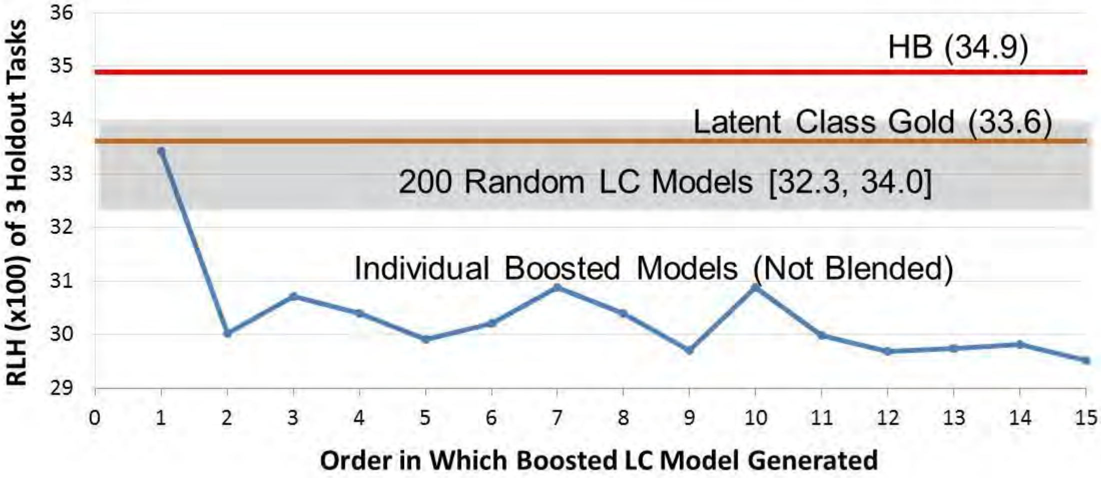
3.2 Results of Boosting

The good news is that boosting clearly creates more diversity across the models. The chart below shows the results for generating 15 boosted models (15 x 15 correlation matrix) vs the 200 random seed generated models. The distribution on the right is Study 1 as shown previously. The corresponding boosted model shifts the correlations to the left, and is more jagged since we only have 15 models. Note that all 15 models have lower correlations with each other than any pair of the 200 random seed generated models.



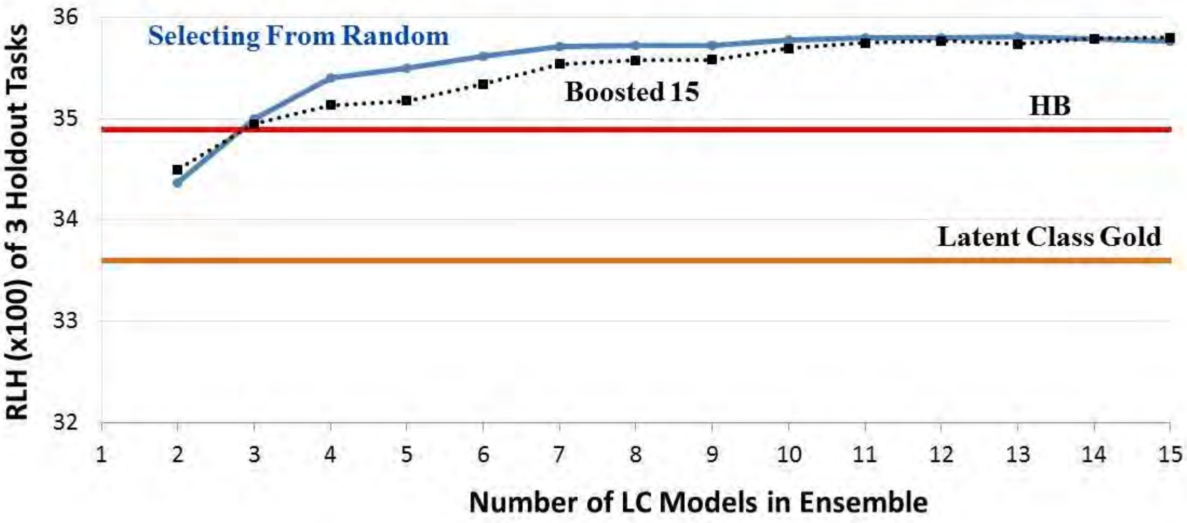
The downside to the boosted models is that each boosted model has a poorer fit. Even though we are using a two-stage process, and we run latent class after removing the boosted weights, we still get significantly poorer fit. It seems the segment probability seeds are so far from optimal that the locally optimal solution is quite a bit worse.

The chart below shows the fit of the 15 boosted LC models for the Study 1 example:



The shaded area shows the range of fit for the 200 random seed LC models, with RLH x 100 in the range of 32.3 to 34.0. The first boosted model has no boosted weights and is in this interval. But the second model has an RLH of only 30. All 14 weight-boosted models have significantly lower RLH across 3 holdout tasks. Ideally we would like our boosted models to be in the shaded area, or at least close to it. Unfortunately across all of our studies we consistently found the weight-boosted models to have significantly lower fit.

It turns out that when we blend the boosted models we still get significant improvement over a single latent class model. In fact, the results are comparable to our random seed approach.



The random seed with backward elimination is the line on top, but by the time we get to blending 11 models the boosted and random models have the same fit. Again, we find this across other case studies as well.

One advantage of the boosted approach is time. Even with the two stages of our boosting approach, it is much quicker to generate 15 boosted models than 200 random seed models. Of course, 200 random seed models is more than one really needs. Even 50 random seed models tends to generate enough diversity and nearly as good results. In addition, the random seed approach can take advantage of parallel processors, since each model is independent. With 2 or 3 parallel processes random generation can be faster than boosting.

4.0 Blending

All blending of models shown prior to this used the simplest form of blending: a simple average of the predictions. For example, when we had 15 boosted models, we took the predictions of the 15 models and averaged them.

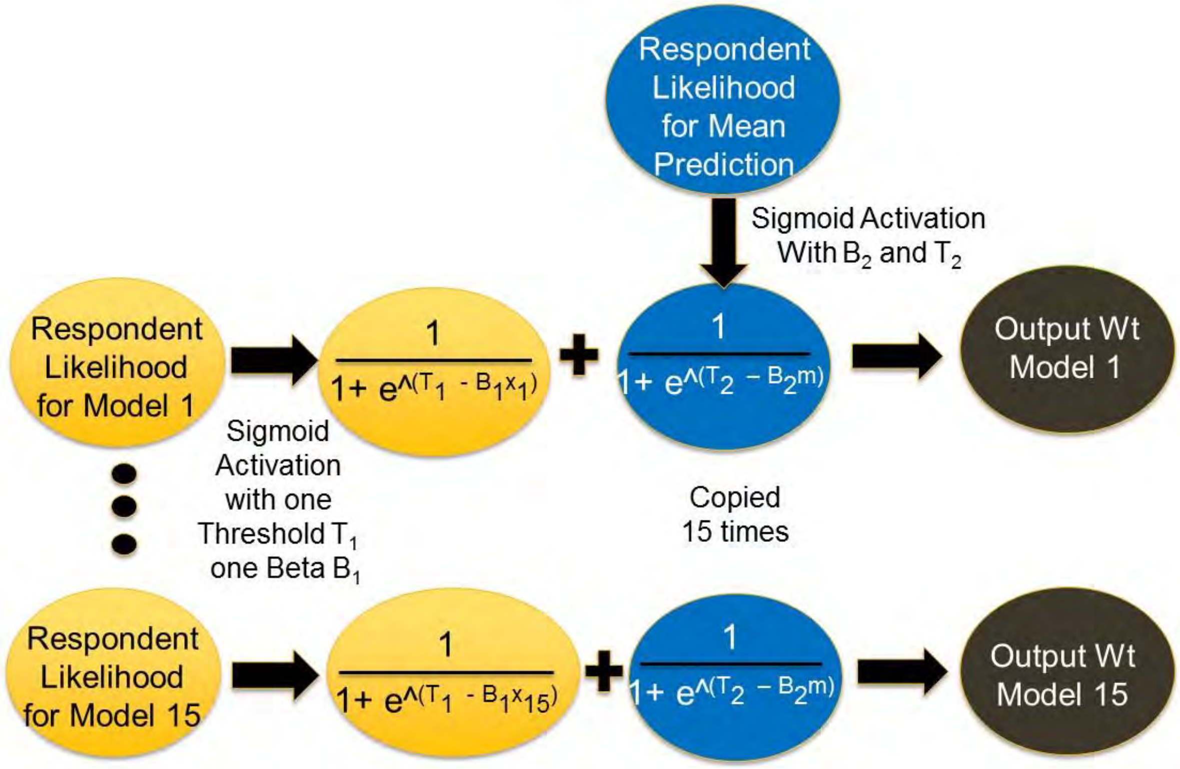
As a reminder, blending is done on the predictions. Our expectation was that we could create significant improvements by developing more complex blending algorithms vs. simple averaging. But what we found was that averaging works very well.

On a similar note, the winners of the Netflix competition created a simple baseline blend using linear regression. It had a RMSE of .87526. Their final blend which was a neural network of neural networks had a RMSE of .87297. That's only an improvement of almost .3%. But it's a slightly bigger improvement than the 2nd place team, and without it they would not have won. In a business context the complex blend makes no practical difference, and would likely not be worth the extra trouble to estimate and implement.

Our improvements from more complex blending were also small, but better than the Netflix winners. We tested a few approaches.

We could only find a very small improvement by using a weighted average of the models. We tried many methods for developing these weights. Due to the multicollinearity of predictions from our models, simple regression did not work well. The best method we found for developing a simple weighted average came from a Shapley Value type regression. We applied Shapley Value to a logistic regression model. The independent variables were the predictions from each model in the ensemble. The dependent variable was the observed choices. We ran all possible logistic regressions, and computed the difference in log likelihood for a variable when it is present vs. absent. The average of those differences became the weight for the model.

Better results were obtained by generating weights for the ensemble models that varied by respondent, so each respondent had their own weights for averaging across the models. We estimated these respondent level weights using a simple neural network. For the sake of explanation, assume we have 15 models in the ensemble. We also have a mean prediction which is the simple average across all 15 models. For each respondent we compute the likelihood of each of the 15 models, x_i. We also compute the likelihood for the mean model m. Then the neural network is defined by 4 parameters, 2 multiplicative values B_1 and B_2, as well as two corresponding thresholds T_1 and T_2.


Respondent Likelihood for Mean Prediction
Sigmoid Activation With B_2 and T_2
Respondent Likelihood for Model 1
$\frac{1}{1+ e^{\wedge}(T_1 - B_1x_1)}$
+
$\frac{1}{1+ e^{\wedge}(T_2 - B_2m)}$
Output Wt Model 1
Sigmoid Activation with one Threshold T_1 one Beta B_1
Copied 15 times
Respondent Likelihood for Model 15
$\frac{1}{1+ e^{\wedge}(T_1 - B_1x_{15})}$
+
$\frac{1}{1+ e^{\wedge}(T_2 - B_2m)}$
Output Wt Model 15

The result of this is respondent specific weights for each of the 15 models in the ensemble. Note that as B_1 goes to 0 this becomes a simple unweighted average.

The table below shows the resulting fit (for our Study 2 example) when we applied blending to the random seed model with backward selection to get 15 models (which are then averaged). We have also included LC Gold and HB tuned results for comparison.

	3 Holdout Tasks		Holdout Sample
	RLH	MAE	MAE
LC Gold	33.62	1.21%	2.23%
HB Tuned Beta	34.93	1.72%	2.30%
LC Boost 15 (Avg)	35.80	1.25%	2.23%
LC Select 15 Random (Avg)	35.80	1.31%	2.19%
LC Select 15 SV Blend	35.82	1.31%	2.21%
LC Select 15 Neural Net Blend	35.95	1.27%	2.21%

Rows 3 and 4 blend ensembles via averaging. The RLH values are those shown before. The last two rows show the slight improvement over the random seed model (with backward selection of 15 models). We have not found much improvement over simple averaging. The neural network with 4 parameters works better than the Shapley Value blend and is estimated very quickly (a minute or so depending upon data). It does however result in respondent level weights for each of the 15 models in the ensemble which is additional complexity for the simulator.

5.0 More Model Structures in the Ensemble

All prior ensemble work in this paper focused on Latent Class solutions. In addition, each solution involved the same coding of the design and the same modeling structure. We did this to show that even in this simple scenario ensembles could work. But one can do much better.

For instance, in some studies with many alternatives we have developed ensembles by specifying different nested logit structures. For each nested logit structure we generated multiple latent class models. Selecting and blending across the different nested logit structures worked very well. In studies with price we have had success modeling the price variable in different ways. Of course, nested logits and pricing are not relevant to all conjoint studies.

One general approach that can be used in all studies is blending HB and Latent Class models. In the table below, we first averaged the 15 LC models. We then blended that result with HB, giving the LC models a weight of 51.4% and HB 48.6% (optimal in-sample weights to maximize RLH). Blending HB with LC ensembles typically improves respondent level fit. However, in our experience HB tends to predict aggregate shares more poorly than Latent Class. So while blending HB and Latent Class improves respondent fit, it tends to lower prediction of aggregate shares.

This point is part of a more general finding. Blending models in an ensemble tends to produce the same error in *aggregate share* as the average of its members' aggregate errors. The improvement is in the *respondent level* fit. For instance, a single Latent Class model tends to have the same degree of aggregate share accuracy as the blended ensemble. The table below shows the results for Study 2, including the first three rows of the previous table for reference.

	3 Holdout Tasks (N = 1500)		Holdout Sample (N =300)
	RLH	MAE	MAE
LC Gold Choice	33.6	1.2%	2.2%
HB Tuned (Raw Beta *.64)	34.9	1.7%	2.3%
LC Ensemble Select 15 Random	35.8	1.3%	2.2%
LC Ensemble Boost 15	35.8	1.3%	2.2%
LC Ensemble Random Multiple Codings	36.1	1.3%	2.2%
LC Select 15 (51.4%) + HB Blend	**37.2**	1.4%	2.2%

The table above shows that the single Latent Class Gold Choice model with 30 segments has the best MAE across 3 holdout tasks (at 1.2%). HB is worst (at 1.7%), and that is after tuning (before tuning MAE is 2.1%). Blending Latent Class models slightly increases the MAE from 1.2% to 1.3% because most of the ensemble models had slightly higher MAE. The improvement from ensembles is at the respondent level. Blending models in an ensemble increases respondent heterogeneity, and reduces bias from a single latent class solution.

The best respondent level RLH comes from blending the Latent Class ensemble and HB, with a significant lift in RLH to 37.2. The corresponding MAE is only slightly poorer than Latent Class ensembles. In our judgment, the lift in respondent fit is worth the small sacrifice in MAE for this case study.

Our other case study in this paper shows a similar pattern.

	3 Holdout Tasks (N = 407)		Holdout Sample (N =406)
	RLH	MAE	MAE
LC Gold Choice	21.2	3.7%	3.5%
HB Tuned	21.7	4.1%	3.5%
LC Select 15 Random	23.0	3.8%	3.5%
LC Boost 15	22.9	3.7%	3.6%
LC Multiple Codings	23.1	3.8%	3.5%
LC Select 15 (70.3%) + HB Blend	23.7	3.8%	3.5%

We see Latent Class Gold Choice again has the best MAE, though only slightly. Ensembles do not help lower MAE. But they do improve respondent level fit, and blending Latent Class ensembles with HB improves respondent fit the most.

Conclusion

Latent Class is an excellent source for developing diverse models for an ensemble. There are many ways to segment data and this is opportunity for ensembles. We found that just using different random seeds generates sufficient diversity. We also found that we did not need hundreds of models in the ensemble. We could obtain equally good (or slightly better) predictions with a reduced set of 10–20 models. We described a fast and easy backward elimination procedure to find this reduced set.

Boosted models in an ensemble typically outperform their random counterparts. Unfortunately none of the boosting algorithms we evaluated performed better than the random seed models with backward elimination. We described a new boosting algorithm that works for any regression problem with a continuous fit statistic, like log likelihood. This worked better than other boosted methods, and about equally as well as the random seed method.

For blending we found that simple averaging of predictions works very well. We described a simple neural network approach that gives different weights for different respondents for each of the ensemble models. In some cases, this may perform better than simple averaging.

We found that blending latent class models maintains the excellent aggregate fit of latent class, but does not improve it. However, it does improve respondent heterogeneity and respondent level fit. Averaging a few latent class models can beat the respondent level fit of HB. But we found the best fit by combining HB and Latent Class Ensembles.

Ensemble methodology is a broad topic and we have only scratched the surface in this paper. Each specific study has many potential ways to apply ensemble methodology. We mentioned a few of them here: nested logits, different codings, constraints, interactions, model specifications. Ensembles recognize and exploit the diversity of different modeling methods, and the creativity of capturing different insights from different models. We look forward to seeing others develop ensemble models in their own way.

Kevin Lattery

COMMENT ON LATTERY'S CONJOINT ANALYSIS ENSEMBLES

BRYAN ORME
SAWTOOTH SOFTWARE, INC.

BACKGROUND

At the 2015 Sawtooth Software Conference, Kevin Lattery presented an intriguing paper entitled, "A Machine Learning Approach to Conjoint Analysis: Boosting and Blending Ensembles." Ensembles involve blending multiple conjoint utility models to improve overall predictive validity. The predictive validity of the blended ensemble often exceeds any specific set of conjoint part-worths within the ensemble. In other words, the whole is typically better than *any* of the parts. Lattery pointed out that ensembles may blend many conjoint part-worth utility estimation approaches, but he illustrated the potential gains specifically using ensembles of just latent class solutions.

Lattery demonstrated that an ensemble of sub-optimal latent class solutions (suboptimal in the sense that he purposefully broke out early, prior to convergence) provided individual-level hit rates that slightly exceeded HB's hit rates for two sample CBC datasets. Though latent class is typically thought of as an aggregate prediction method (predicting shares of preference for groups rather than individuals), it is well-known that pseudo individual-level utilities may be developed by taking the weighted average of the class utility vectors, where the weights are each individual's likelihood of belonging to each class. Such pseudo individual-level utilities are usually found to be less predictive of holdout choice tasks than HB utilities. Rich Johnson (Johnson 1997) discussed reasons for this and illustrated it with multiple data sets in his 1997 paper introducing ICE (Individual Choice Estimation). Therefore, it intrigued me when Lattery demonstrated that creating ensembles of pseudo individual-level utilities from a relatively small number sub-optimal latent class runs could outperform HB for his two data sets.

SAMPLE CBC DATASET FOR VALIDATION

Because my interest was piqued, shortly after the conference concluded, I conducted a limited investigation on a single CBC dataset provided to me by our partners, SKIM Group, The Netherlands (where Lattery currently works). This investigation was meant to be my own internal proof of concept, but others may be interested in the results—especially the extension of Lattery's approach to investigate ensembles of HB solutions.

The CBC dataset I used was very robust, with 2005 respondents, 15 choice tasks per respondent, and 3 concepts per task. The design was a 3x4x5x4x15 attribute levels experiment. Although the original CBC dataset included a dual-response none, I only analyzed the forced choice among the three product alternatives and ignored the secondary none question.

The CBC dataset I analyzed did not include any holdout choice tasks, so I selected tasks 10 through 12 for each respondent to hold out for validation purposes. Because the design included dozens of versions (blocks), these 3 holdout tasks had a lot of variation in attribute level composition and contexts across respondents. With a fixed design of three holdout tasks, one runs the risk that the characteristics of those specific three tasks may have some negative bearing on validation comparisons. Given that it actually covers hundreds of unique holdout choice tasks, selecting each respondent's 10^{th} through 12^{th} tasks to hold out is quite robust.

In contrast to Lattery's approach of computing hit rates on a continuous root likelihood scale (which can be affected by scale factor differences between utility estimation methods), I used the raw hit rate approach (1=hit, 0=miss) to evaluate internal predictive validity, which is not affected by scale factor differences. Raw hit rates have less precision than continuous likelihood of hit rates, but this is ameliorated in my situation due to the rich pool of 2005 respondents x 3 choice tasks = 6015 unique holdout tasks available for validation.

HB as Baseline Comparison

As a baseline for comparison, I used Sawtooth Software's CBC/HB software with its default settings of Degrees of Freedom=5, Prior Variance=2. But to ensure convergence and obtain potentially slightly better precision I increased the iterations above the defaults to 10K burn-in iterations followed by 100K used iterations. I specified the default part-worth main effects model and estimated the model using 12 of the original 15 choice tasks (recall that tasks 10 through 12 are held out). For each respondent, I counted how many of the holdout choice tasks I could predict correctly using the part-worth utilities, achieving a hit rate of 64.84% for the sample.

It is well known that the priors can have an effect on the predictive validity of HB models. McCullough illustrated this nicely in his 2009 paper at the Sawtooth Software Conference (McCullough 2009). He examined multiple data sets and did a grid search to find which combination of Degrees of Freedom and Prior Variance led to highest predictive hit rate. For one data set, he showed that the defaults (Degrees of Freedom=5, Prior Variance=2) in Sawtooth Software's CBC/HB system led to a 61.8% hit rate. Tuning the priors to Degrees of Freedom=30, Prior Variance=0.25 led to hit rates for the same dataset of 65.1%. I'm currently working with Walter Williams on an automated method for any CBC or MaxDiff dataset that doesn't require fixed holdout tasks to find the optimal HB priors settings. We employ jackknifing (holding out a subset of existing "random" tasks in each replicate, typically just 1 or 2 tasks per respondent) and bootstrap resampling with HB estimation. We plan to share our results at a forthcoming Sawtooth Software event. I used this procedure to find that for the dataset in this current investigation, the optimal priors settings are Degrees of Freedom=105 and Prior Variance=0.25. Note that this was found via a jackknifing procedure that leverages *all* tasks in the dataset (including the original dual-response None), not just optimizing for tasks 10 through 12. So, I am not cherry-picking by selecting these priors when building HB models that will be used to predict holdout tasks 10 through 12. After fitting a single HB model with these optimized priors (again using 10K burn-in iterations followed by 100K used iterations), the hit rate for held-out tasks 10 through 12 was 65.78%, representing almost a 1% absolute increase in hit rate over the default HB run.

To summarize, here are the hit rates so far for this dataset:

Default HB (D.F.=5, PriorVar=2)	64.84%
Optimized HB (D.F.=105, PriorVar=0.25)	65.78%

Latent Class Ensembles

Next, let's turn to Lattery's Latent Class ensembles approach. I wasn't able to precisely follow Lattery's recommendations because I was using Sawtooth Software's Latent Class tool rather than Latent Gold. For example, Lattery described creating sub-optimal Latent Class runs by breaking out of the iterations once the last 10 iterations had provided a sum total of no more than 0.1% gain in log-likelihood. Sawtooth Software's latent class software breaks out after the last iteration fails to increase log-likelihood by a user-defined threshold. So, with some experimentation for this dataset, I found that if I broke out after a fixed 30 iterations, the sum total of the gains over the last 10 iterations was about 0.1% on average across replicates (this of course will differ for other datasets).

In his paper, Lattery described a process of backwards pruning to isolate a relatively small ensemble of a dozen or so latent class runs that represent a lot of diversity. Rather than undertake that effort, I took Lattery's recommendation that a simple averaging across a few dozen latent class runs would also do suitably well (though it would be more complex than Lattery's approach to manage for building market simulators).

Following Lattery's recommendations (though with the simplifications noted above), this is the latent class ensemble procedure I employed for predicting holdouts at the individual level:

1. Create dozens of sub-optimal latent class runs (using a different random starting seed each time) by breaking out well prior to convergence. I used 30-group latent class solutions[1].
2. For each latent class run, create[2] pseudo individual-level utilities (by taking a weighted combination of the latent class utility vectors according to each respondent's likelihood of belonging to each class).
3. For each respondent and each latent class run, use the logit rule to estimate "shares of preference" for the holdout tasks. For each respondent, average the shares of preference across latent class runs to create consensus shares of preference for the holdout tasks.
4. For each respondent and holdout choice task, if the concept the respondent chose had the highest predicted share of preference, score a hit. Otherwise, score a miss.
5. Summarize the hits and misses across all respondents and holdout choice tasks.

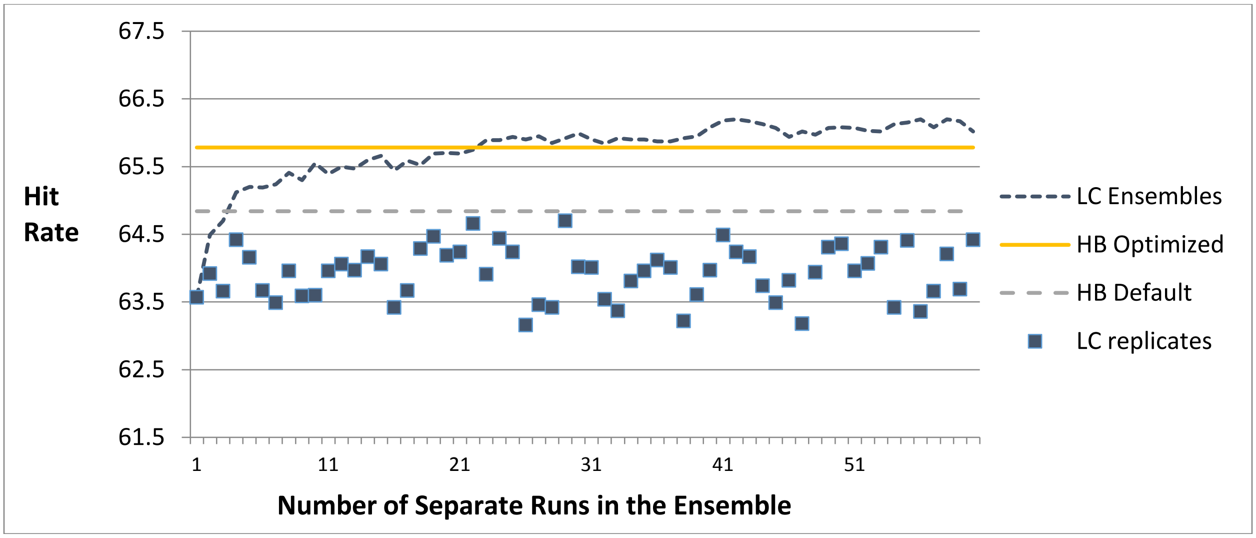
Performing steps 3 through 5 is not facilitated by Sawtooth Software tools (I used a third-party statistical software package), though Sawtooth Software's latent class software handles steps 1 and 2 quite nicely.

Figure 1 shows results for the two benchmark HB runs previously described as well as for latent class ensembles (for 1 to 60 latent class replicates in the ensemble.

[1] Lattery advocated using higher dimension latent class solutions if sample size affords it, because greater dimension latent class solutions provide more variety across replicates. He reported good results for 30-segment solutions. With my robust sample size of n=2005, I was also able to use 30-group solutions.

[2] Upon reviewing this document, Lattery clarified that rather than take the weighted average of part-worth utilities to create his individual-level predictions for latent class, he took the weighted average of the probability predictions for the holdouts. We both believe the differences between our two approaches should be very small, though Lattery prefers to take averages of predictions rather than averages of part-worths.

Figure 1



The Y axis shows the hit rates for different utility estimation methods (I have charted just the range of 61.5 to 67.5 to enhance the differences for readability, though the reader should note that all methods produced hit rates that were not too different from one another). The X axis shows the results for each successive latent class replication in the ensemble. In the case of the blue markers, they are not cumulative but are treated independently. As expected, these are less successful than HB runs or the latent class ensembles approach. But, in the case of the dotted blue trend line for latent class ensembles, the results are cumulative through each of 60 replications. For example, the dotted blue line at the 10th replication shows the results that can be achieved by creating an ensemble solution across the first 10 replications.

For this dataset, after leveraging just four latent class replications, the latent class ensemble hit rate exceeds that of default HB. After about two dozen replications used, the latent class ensemble slightly edges out the optimal HB run for this data set (though the differences in predictive validity are extremely small).

What about using an optimal latent class run where we do not break out early? Are ensembles actually helping us? To test this, I ran latent class 30 separate times, breaking out only after successive iterations failed to increase the log-likelihood by more than 0.1. The best try out of the 30 was assumed to be near-optimal. Although I am not certain that I was able to find the globally optimal solution (given the complexity of a 30-group solution), I am pretty confident that it is quite close to optimal (and it reflects much greater fit than the sub-optimal runs used in the ensemble). Interestingly enough, the hit rate for the "optimal" latent class run was 63.11%, which is actually below every one of the 50 sub-optimal latent class runs used in the ensemble! Why a latent class run with higher log-likelihood fit should provide slightly lower individual-level hit rate fit than sub-optimal latent class runs is strange, but it may represent overfitting of the 30-group solution. Perhaps with a lower dimension solution we wouldn't see the illogical outcome.

Ensembles for HB

Lattery mentioned that ensembles are not just limited to latent class, but could be applied to any part-worth utility estimation method, including HB. He also emphasized that ensembles could potentially be made stronger by mixing model specifications: interactions vs. no interactions, constrained models vs. unconstrained models, etc. However, my interest was to see, as in the latent class ensembles approach reported above (which held the model specification constant), if I could similarly achieve a lift in HB hit rates by creating an ensemble of sub-optimal HB runs based on the same model specification.

Lattery provided no guidance on this, so my first attempt was to arbitrarily choose to break out early from HB runs after 1K burn-in iterations and 5K used iterations. Each of the sub-optimal HB replicates provided slightly lower hit rates than the 10K+100K version reported above. With this first try at HB ensembles, I did not find similar gains for HB ensembles as for latent class, I think due to the relative lack of diversity of the individual-level part-worth utilities across the HB replicates. Ensemble methods benefit from quality of solutions *and* diversity. As an example, for respondent #24, the correlation squared for the pseudo individual-level utilities across multiple sub-optimal latent class replicates averaged 0.812, whereas the average correlation squared for HB (point estimates, meaning the average of the used draws for each respondent) across multiple sub-optimal HB runs was 0.968.

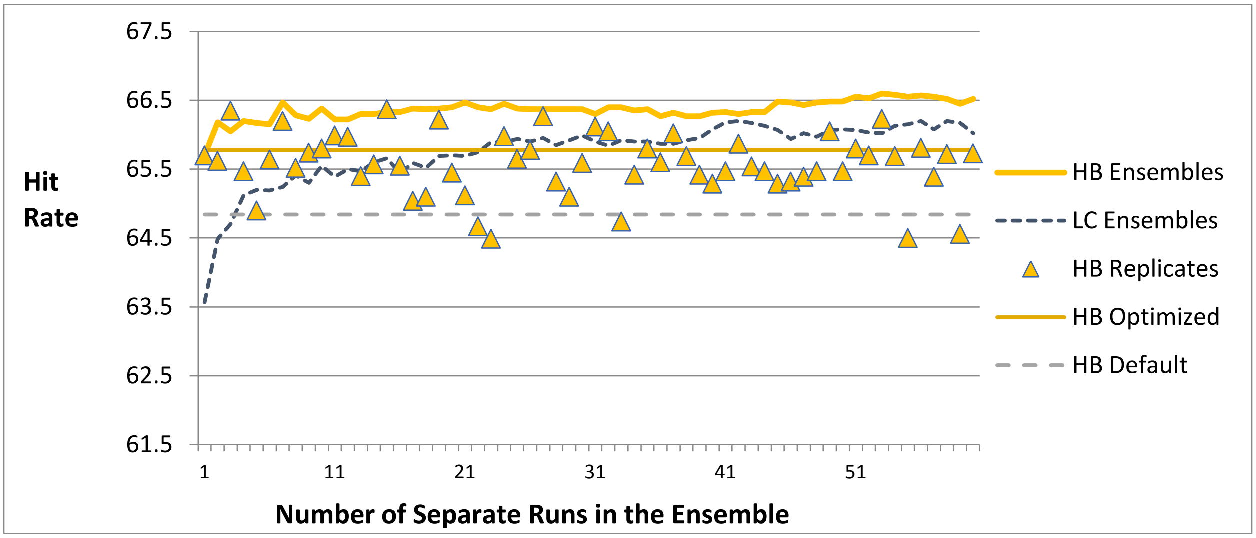
My second attempt at HB ensembles was more successful. Rather than holding the same HB settings constant across replicates (except for random starting points), I decided to do something that could force greater diversity across replicates and wouldn't reuse the same HB model repeatedly. I decided to apply different covariates across the HB replicates (again applying the "optimized priors" settings as before). However, I didn't have any additional information for this dataset beyond the choice tasks, so my covariates couldn't come from external questionnaire data. Rather, I used discrete latent class segment assignment as covariates[3], varying from 3 to 5 group solutions. To encourage variability across the covariates, I used sub-optimal latent class runs using different random starting seeds where for each I purposefully broke out after just 12 latent class iterations. For the HB runs with covariates, I used 10K burn-in iterations followed by 40K used iterations[4]. This seemed to do the trick and I now saw modest gains for ensembles of HB runs.

Figure 2 retains the latent class ensembles results from Figure 1 for comparison, but adds two series (gold markers and a gold line) to the previous chart, representing the hit rate achieved by each HB with covariates run and the hit rate achieved by blending an ensemble of those runs.

[3] Not a formally proper procedure according to Bayesian statisticians, who argue that this is "double-dipping" the heterogeneity.

[4] Given such a large dataset (n=2005) and using covariates, each separate HB run took about 30 minutes on my laptop, which I configured in the CBC/HB software package to run in batch mode. To make HB ensembles quicker in practice, one can set multiple instances of CBC/HB software running in batch mode, such that multiple HB runs can be run in parallel. For example, with a 4-core processor, one could estimate HB replicates much faster than when running one at a time. An ensemble of 20 HB replicates (of the hefty size of this CBC dataset) could be completed in about 2 hours of runtime (though manual setup would likely be an additional hour or two).

Figure 2



The HB with covariates runs (indicated by gold triangle markers) typically did no better nor worse individually than the HB Optimized run. But, creating an ensemble of the HB runs that applied latent class membership as covariates led to a modest amount of improvement. The results were slightly better than the latent class ensembles, though the differences in hit rates are again all very small.

Combining Latent Class and HB Runs in the Ensemble

Naturally, I wondered whether combining the latent class and HB runs into one larger ensemble could provide a better solution than the best ensemble result achieved to this point. The combined latent class and HB runs within one ensemble lifted the hit rate to 66.82, another three-tenths point higher than the previous best results. This is reassuring, since ensemble analysis is supposed to benefit from both quality and variety across the collection of individual solutions. Latent Class ensembles seem to add good quality and variety to the HB ensemble.

To summarize, here are the hit rates for this dataset, sorted from less to more successful:

Best Fit 30-group Latent Class Solution	63.11%
Default HB (D.F.=5, PriorVar=2)	64.84%
Optimized HB (D.F.=105, PriorVar=0.25)	65.78%
Latent Class Ensemble (60 replicates)	66.02%
HB Ensemble (60 replicates)	66.52%
HB + Latent Class Ensemble (120 replicates)	66.82%

Comments and Conclusions

If one is interested in prediction accuracy for CBC models at the individual level, latent class ensembles work very well! In terms of individual-level predictions, an ensemble of sub-optimal latent class replicates performs better than a single near-optimal latent class run.

We shouldn't think of HB as the only method for achieving high-quality individual-level utility estimates. Many past presentations at the Sawtooth Software conference have shown different methods that do nearly as well as HB or even better in certain situations. Examples include:

- Rich Johnson's ICE method (1997)
- Kenneth Train's mixed logit (2000)
- Lattery's EM approach (Lattery 2007)
- Latent Class with C factors (McCullough 2009)
- Empirical Bayes (See the Discover-CBC white paper at http://www.sawtoothsoftware.com/1267)

Lattery (2015) has now shown us (and I've replicated his findings here) that latent class ensembles also offer another valuable alternative for CBC modeling. There is no reason to believe that the same benefits wouldn't extend to other related choice data such as MaxDiff and MBC as well. Also, as Lattery mentioned, more sophisticated ensembles may be explored by blending different model specifications (e.g., main effects vs. interactions, linear vs. part-worth models, constrained vs. unconstrained). Such extensions could perform even better than what I've demonstrated here.

HB ensembles where the models are made to vary across replicates via a variety of latent class segmentations as covariates performed slightly better than latent class ensembles (following Lattery's suggested procedure, except for certain details noted above) for this dataset. This limited investigation was more of a proof of concept rather than testing whether ensembles of HB solutions generally work better than ensembles of latent class solutions. For this data set, both approaches worked extremely well and combining the approaches within a single ensemble worked even better. This follows ensemble theory nicely, as variety and quality within the ensemble should boost overall predictive validity.

For this dataset and the limited ways I chose to build variation into utility run replicates, an ensemble of 20 HB utility solutions was adequate to achieve near-optimal results; and an ensemble of about 40 Latent Class utility solutions achieved near-optimal results.

Is all this extra effort involved in building diverse part-worth utility models and stitching together ensemble simulators worth it for practitioners? The last few decades have seen many innovations in choice modeling where the gains have been measured in terms of just a very few points in terms of hit rates for holdout choice tasks (hit rates are notoriously stubborn to lift much). Consultants are deeply concerned about the robustness, face validity, and predictive performance of their market simulators, especially when conducting sensitivity analysis and product optimization searches to direct marketing strategy. If practitioners are looking for an extra edge that can improve their choice simulators without spending hard cash for larger sample sizes or burdening respondents with longer questionnaires . . . if practitioners are looking for yet another way to distinguish themselves from other consulting shops, then conjoint/choice analysis ensembles indeed appear to be a promising new approach.

Bryan Orme

References

Johnson, R. M. (1997), "ICE: Individual Choice Estimation," Sawtooth Software.

Lattery, Kevin (2007), "EM CBC: A New Framework for Deriving Individual Conjoint Utilities by Estimating Responses to Unobserved Tasks via Expectation-Maximization (EM)," Sawtooth Software Conference.

Lattery, Kevin (2015), "A Machine Learning Approach to Conjoint Analysis: Boosting and Blending Ensembles," Sawtooth Software Conference.

McCullough, Richard (2009), "Comparing Hierarchical Bayes and Latent Class Choice: Practical Issues for Sparse Data Sets," Sawtooth Software Conference Proceedings, pp 273–284.

Train, Kenneth (2000), "Estimating the Part-Worths of Individual Customers: A Flexible New Approach," Sawtooth Software Conference.

THE UNRELIABILITY OF STATED PREFERENCES WHEN NEEDS AND WANTS DON'T MATCH

MARC R. DOTSON
GREG M. ALLENBY
FISHER COLLEGE OF BUSINESS, THE OHIO STATE UNIVERSITY

1. INTRODUCTION

We buy products that respond to our needs and improve our lives. We find value in marketplace offerings because of our pursuits and because of the things that concern us. Knowing what we need allows us to form preferences and make decisions, both in the context of our lives and in response to researchers who survey us about our beliefs and buying intentions. In this paper we examine consumers' ability to provide reliable statements of preference for offerings that they may not need, where needs are measured as the concerns and interests of individuals engaged in behaviors related to a product category.

Screening respondents for inclusion in survey samples is a common practice in marketing, serving to identify candidate respondents with preferences relevant to the offerings of a firm. Researchers qualify respondents by asking if they engage in behaviors related to a focal product category, if they take part in the decision-making process, and if they are willing and able to make purchases. Our analysis indicates that this may not be enough, particularly when describing products in terms of a subset of their attributes and benefits. Products can be described in terms of dozens of product attributes, with a relatively small subset examined in any particular study. We find that respondents do not reliably state their preferences when choice options are described in terms that aren't relevant to them. Asking respondents about their preferences for things they don't need leads to inconsistent and noisy responses, not responses that indicate consistently low preference.

Academics and practitioners have not diligently investigated the importance of motivating conditions, or needs, in models of choice using stated preference data. Exceptions include Fennell and Allenby (2014), who demonstrate that data on needs and wants are conceptually and empirically distinct, and Chandukala et al. (2011), who use information on consumer frustration with products to identify unmet demand. Additionally, Yang et al. (2002) show that motivating conditions result from the intersection of individuals with their environmental context. In contrast, researchers in marketing have tended to focus on goal pursuit (Bagozzi and Dholakia, 1999), which is a more aggregate measure that does not easily relate to the specific attributes and benefits of an offering.

We develop a model for use in conjoint analysis that incorporates motivating conditions and measures the effect of relevance on preference reliability. We accomplish this by modifying the random utility model's error term, which represents consistency of utility expression in the choices that are made. Instead of assuming that the extent of this choice certainty is constant across respondents and products, we structure the scale parameter of the error term for each product to be a function of the consumer's needs being addressed. This specification allows for an individual to be more or less consistent in the preferences they form for a given product based on whether or not the product is relevant to them.

We find differences in the reliability of preferences when comparing responses where needs are addressed by the product versus responses where needs are not addressed, providing support for the claim that screening on category participation alone may not be enough for choice experiments. We use data from a conjoint study in the pre-packaged dinner category that provides a unique opportunity to relate needs and wants. Including these less-reliable responses in standard models of stated preference results in possible parameter bias.

The remainder of the paper will be organized as follows. Section 2 provides an overview of the literature related to the concept of relevance and how it might impact consumer choice. Section 3 develops our model and includes simulation results showing that it is statistically identified. Section 4 details our empirical application. In Section 5 we compare results from our model and alternative models. Concluding remarks are offered in Section 6.

2. Relevance and Choice

Marketing as a discipline is inseparable from the concept of relevance. At its core, marketing is concerned with understanding the needs of consumers and developing products that people will want to buy. Past research has framed relevance primarily with respect to the pursuit of goals. A goal is defined as a cognitive representation of a desired end state (Fishbach and Ferguson, 2007; Bagozzi and Dholakia, 1999). Motivation refers to the psychological force enabling consumers to try and remove the disparity between their current state and their desired end state (Lewin, 1935; Ajzen and Sheikh, 2013). Individuals with an outcome-oriented motivation will seek out products with certain benefits that move them closer to obtaining their imagined end state (Touré-Tillery and Fishbach, 2011; Petty and Cacioppo, 1981; Haley, 1968). In other words, relevance is seen as a product's benefits being aligned with the goals of the consumer.

This focus on the alignment of benefits and goals disregards the motivating conditions that drive consumers to the marketplace to begin with. Just as goals represent a desired end state, motivating conditions describe the consumer's current state (Fennell and Allenby, 2014). Failing to understand where consumers are coming from provides an incomplete picture regarding the needs that products, offerings, and promotions must be designed to address. Malär et al. (2011) demonstrate the importance of considering both current and desired states and find that, in general, brands should be positioned to align with the needs of the current state rather than the aspirations of the desired state. We continue this focus on the current state by defining relevance as an alignment between known motivating conditions and a product's perceived benefits.

Our interest is in exploring the mechanism through which relevance impacts consumer choice. This requires that we consider how utility is expressed. The literature on random utility models comprises two parts, deterministic and random, with the random component reflecting the consistency of consumer choices. Consider the standard random utility model:

$$U_{jh} = x_j'\beta_h + \epsilon_{jh}, \tag{1}$$

where the deterministic component $V_{jh} = x'_j\beta_h$ for respondent h is related to the attributes (x_j) and benefits of product j, and ε_{jh} is the random component thought to arise from unobservables. It is typically assumed that ε_{jh} is independent and identically distributed. The scale parameter (σ) of the random component represents preference consistency.

Traditionally, relevance is thought to affect choice through the deterministic component of the model by allowing the coefficients β_h to be cross-sectionally related to an appropriate set of covariates (z_h) through a random-effects model:

$$\beta_h = \Gamma' z_h + \xi_h. \tag{2}$$

Thus, if large positive values of the variables z_h indicate consumers for whom the product is relevant, it is expected that some elements of the Γ coefficient matrix would be positive and would predict large positive coefficients in β_h. Allenby and Ginter (1995b) specify z_h as the demographic variables of age, income, and gender and examine their relationship to part-worths (β_h) in a conjoint analysis, and Rossi et al. (1996) examine the information contents of demographics in general. Lenk et al. (1996) examine the role of expertise and other variables on personal computer purchases, and more recently Chandukala et al. (2011) examine the role of motivating conditions in explaining variation in β_h. The matrix of coefficients Γ in Equation (2) maps cross-sectional variation in the variable z_h to variation in the coefficients β_h, and provides a flexible model describing correlates of utility formation and preference.

In practice, the influence of z_h on explaining heterogeneity in β_h in Equation (2) has not met with much success. Rossi et al. (1996) show that the inclusion of demographic variables as covariates only explains between 7 and 33% of the variability in β_h. Horsky et al. (2006) only see a 5% improvement in the log-marginal density when moving from an intercept model (-3, 304.6) to a model that includes covariates (-3, 149.1), as in Equation (2). Similarly, Chandukala et al. (2011) only see a 1% improvement of the log-marginal density when moving from an intercept model (-19,090.11) to a model that includes covariates (-18,976.53). Heterogeneity in model coefficients has largely been explained by unobservable factors (ξ_h) rather than observable factors ($\Gamma' z_h$).

Louviere et al. (2002) detail both the importance of studying the scale term (σ), which represents preference consistency, and what future paths this line of research might take. They describe the scale term as an expression of "unobserved variability" and discuss its importance for developing more complete models of choice as well as testing theories of choice processes, especially when we consider the influence of context on choice. Fiebig et al. (2010) further explore the concept by testing what they term "the generalized multinomial logit model" and its variants. In this model, they allow for both part-worth and scale heterogeneity. They conclude that the scale distribution does more to explain heterogeneity than the distribution over the coefficients. While the claim that people differ solely in preference certainty and not in preferences is untenable, the illustration of the influence of the scale term is notable.

Our research is intended to add to the literature on modeling the scale term in the random utility model. Specifying that the random component, or error term, in the random utility model is independent and identically distributed is clearly a simplifying assumption. However, little has been done to explain variability in the scale term from a behavioral standpoint. One exception is Dellaert et al. (1999), who show that choice difficulty has an influence on consistency of choice. Our model provides another behavioral interpretation by framing "unobserved variability" in terms of the impact of relevance on preference consistency.

3. Model Development

We develop our model by starting with the multinomial logit, detailing where our work deviates from standard choice models, and concluding with a validation of our model, specific to our empirical application, via a simulation study.

3.1 The Multinomial Logit

The multinomial logit model has been a workhorse in choice modeling, including conjoint analysis. In a conjoint study, respondents are presented with a fixed number of product alternatives, with the attributes composing each alternative set by an experimental design. The respondent is typically asked to choose the single alternative they most prefer. This process is then repeated, with each choice task consisting of differently configured alternatives. The standard multinomial logit model assumes extreme value error terms in the random utility model that are independent and identical. This results in the following closed-form expression of the likelihood for a single choice task with K alternatives:

$$\begin{aligned} Pr(j)_h &= Pr\left(V_{jh} + \epsilon_{jh} > V_{kh} + \epsilon_{kh} \text{ for all } k \neq j\right) \\ &= \int_{-\infty}^{\infty} F([V_{jh} - V_{1h} + \epsilon_{jh}]/\sigma) \cdots \\ &\qquad F([V_{jh} - V_{Kh} + \epsilon_{jh}]/\sigma) f(\epsilon_{jh}/\sigma) d\epsilon_{jh} \\ &= \frac{\exp[V_{jh}/\sigma]}{\sum_{k=1}^{K} \exp[V_{kh}/\sigma]} \end{aligned} \tag{3}$$

where $V_{jh} = x'_j\beta_h$ is the deterministic component of random utility for respondent h. It is important to note in this expression that it is $x'_j\beta_h/\sigma$ that is identified. Thus the part-worths β_h for respondent h will be a function of both the marginal utility $x'_j\beta_h$ and the scale of the error term σ (Swait and Louviere, 1993; Sonnier, Ainslie, and Otter, 2007). It is typical to set $\sigma = 1$.

3.2 The Heteroscedastic Multinomial Logit

We relax the standard assumption of homoscedastic errors to model the impact of relevance on choices. Following Allenby and Ginter (1995a), we assume the error terms are distributed extreme value, but allow for the scale parameters to differ across individuals and alternatives. We relate the scale of the error for each choice alternative to covariates through a functional form that indicates whether a choice alternative is relevant to the individual.

We investigate covariates z_h that represent needs, or motivating conditions, associated with the attributes and benefits of a product (see Fennell and Allenby, 2014). The need variables are measured on a binary scale (i.e., absent or present) and describe the current condition of the respondent. The product attributes x_j are also binary. In our application the needs are intentionally matched to corresponding product attributes (i.e., dim(x_j) = dim(z_h) = M). The variable z_h describes from whence respondents come, and x_j describes product features that might be of interest to them. We investigate a model where the scale parameter for the jth alternative for respondent h is related to the joint presence of a motivating condition and a corresponding attribute:

$$\sigma_{jh} = \exp\left[\gamma \cdot I\left(\sum_{m=1}^{M} x_{jm} z_{hm} = \sum_{m=1}^{M} z_{hm}\right)\right]. \tag{4}$$

In this specification, which assumes a priori that we have a one-to-one mapping between attributes and needs, γ will measure the effect of relevance on consistency in utility expression. When the product is irrelevant (i.e., all of a respondents needs aren't addressed by wants), $\sigma_{jh} = 1$. A negative value of γ would produce a small σ_{jh} and thus indicate that relevance (i.e., all of a respondents needs being met by wants) leads to more certainty in utility expression and choice. Assuming independent but no longer identically distributed errors means we no longer have a closed-form expression for the choice probabilities.

The problem with this simplified model is that needs aren't typically mapped one-to-one with product attributes. Rather, consumers seek out products with benefits that address their needs where the benefits each product provides depends on the consumers' beliefs, particularly about the associated brand. To model choice as a function of the consumer's needs, brand beliefs, and price sensitivity, we need an extended model of behavior.

3.3 An Extended Model of Relevance and Preference Reliability

With an extended model, we can model the scale parameter in terms of the joint presence of a motivating conditions and the perceived benefits that address them:

$$\sigma_{jh} = \exp\left[\gamma \cdot I\left(\sum_{m=1}^{M} z_{hm} \geq 1\right) \cdot I\left(\sum_{m=1}^{M} B_{jhm} z_{hm} = \sum_{m=1}^{M} z_{hm}\right)\right]. \tag{5}$$

The covariates z_h are again a binary vector indicating active needs. B_h is a binary matrix of respondent h's brand beliefs regarding benefits, where the jth row is a vector of beliefs about brand j. We assume an a priori one-to-one mapping between the M needs and M benefits, thus for every need included in the analysis there is a corresponding benefit that satisfies it. The indicator functions specify that when respondent h has active needs and that the product in question has a brand that the respondent believes is able to address all of their needs, γ will measure the effect of relevance on preference certainty. We exponentiate the expression in Equation (5) to ensure that the scale term is positive. If either of the indicator functions don't hold then $\sigma_{jh} = 1$ as is typically assumed. A negative value of γ would produce a small σ_{jh} and thus indicate that relevance leads to more preference certainty. We expect choices among relevant options to be associated with greater preference certainty, therefore we expect the estimate of γ to be negative.

Without assuming identically distributed error terms, the standard multinomial logit model used for discrete choice no longer has a closed form expression. Additionally, we expect that in the conjoint survey the relevant choices will be the ones that are picked first. To ensure that we have enough data to identify γ, we anticipate using ranked data rather than first-choice only. To make use of ranked data, we employ the exploded multinomial logit model (Chapman and Staelin, 1982). The exploded multinomial logit decomposes each choice task with K alternatives into K − 1 independent choice tasks, each with successively fewer alternatives. Ranking the K-th alternative is deterministic given the previous K − 1.

With ranked data, we are now interested in the probability that the first ranked alternative, denoted by $U_{(1)}$, has a utility expression that is greater than or equal to the second ranked alternative, $U_{(2)}$, and so on. Thus the random utility components for the *i*th ranked alternative are denoted $V_{(i)h}$ and $\varepsilon_{(i)h}$. The exploded multinomial logit assumes that individuals rank their most preferred alternative first, their second preferred alternative second, and so on, and that the choice probabilities for each ranking are independent. The probability of an observed rank ordering is:

$$Pr(U_{(1)} > U_{(2)} > \cdots > U_{(K)})_h = \prod_{i=1}^{K-1} \frac{\exp[V_{(i)h}/\sigma]}{\sum_{k=i}^{K} \exp[V_{(k)h}/\sigma]}. \tag{6}$$

Relaxing the assumption of homoscedastic errors leads to the heteroscedastic exploded multinomial logit model. Without a closed-form expression, the probability of an observed rank ordering is:

$$\begin{aligned} &Pr(U_{(1)} > U_{(2)} > \cdots > U_{(K)})_h \\ &= \prod_{i=1}^{K-1} Pr\left(V_{(i)h} + \epsilon_{(i)h} > V_{(k)h} + \epsilon_{(k)h} \text{ for } k = i+1, \ldots, K\right) \\ &= \prod_{i=1}^{K-1} \int_{-\infty}^{\infty} \left[\prod_{k=i}^{K} F([V_{(i)h} - V_{(k)h} + \epsilon_{(i)h}]/\sigma_{(k)h})\right] f(\epsilon_{(i)h}/\sigma_{(i)h}) d\epsilon_{(i)h}. \end{aligned} \tag{7}$$

Both needs z_h and brand beliefs B_h are included as additional data in our model. The benefit evaluation utilizes the exploded multinomial logit as detailed in Equation (6). The brand-price evaluation utilizes the heteroscedastic exploded multinomial logit as detailed in Equation (7) with the structured scale term as detailed in Equation (5). To bridge the two likelihoods, we model the brand intercepts as $\beta_{0h} = B_h\beta_{0h}$. In other words, brand intercepts are a sum of the benefit part-worths from the benefits respondent *h* believes each brand provides as indicated by their brand beliefs.

3.4 Simulation Experiment

We validate our extended model of behavior by generating data according to the model and recovering parameters using a random walk Metropolis-Hastings estimation algorithm. We employ Simpson's rule to numerically integrate for the heteroscedastic exploded multinomial logit. The simulation experiment matches the dimensions used in our empirical application: a single γ and 31 β_h's for each of 567 respondents.

After 60,000 iterations the Markov chain converges to the true stationary (i.e., posterior) distribution. In Table 1 we demonstrate that we have recovered the true parameter values for γ and the mean of the model of heterogeneity over β_h, where each parameter estimate is within or near the bounds of a 95% credible interval.

Table 1 Simulation Results and 95% Credible Intervals

Parameter	True Value	Posterior Mean	Lower	Upper
γ	-0.30	-0.11	-0.36	0.16
δ_1	2.50	2.64	2.46	2.84
δ_2	2.25	2.36	2.11	2.56
δ_3	2.10	2.24	1.99	2.45
δ_4	2.00	2.08	1.85	2.30
δ_5	1.90	2.05	1.88	2.23
δ_6	1.80	1.84	1.63	2.07
δ_7	1.75	1.79	1.61	1.97
δ_8	1.63	1.57	1.40	1.75
δ_9	1.50	1.47	1.23	1.69
δ_{10}	1.15	1.26	1.08	1.43
δ_{11}	1.00	1.10	0.91	1.28
δ_{12}	0.75	0.77	0.61	0.91
δ_{13}	0.50	0.71	0.52	0.89
δ_{14}	0.35	0.34	0.15	0.52
δ_{15}	0.20	0.16	-0.04	0.40
δ_{16}	0.10	0.12	-0.16	0.34
δ_{17}	0.05	0.05	-0.16	0.26
δ_{18}	0.01	-0.07	-0.23	0.09
δ_{19}	-0.08	0.05	-0.14	0.27
δ_{20}	-0.40	-0.63	-0.86	-0.41
δ_{21}	-0.52	-0.66	-0.88	-0.46
δ_{22}	-0.95	-0.91	-1.11	-0.68
δ_{23}	-1.05	-1.27	-1.45	-1.08
δ_{24}	-1.15	-1.08	-1.28	-0.90
δ_{25}	-1.28	-1.20	-1.42	-0.97
δ_{26}	-1.40	-1.48	-1.7	-1.25
δ_{27}	-1.45	-1.42	-1.58	-1.22
δ_{28}	-1.50	-1.46	-1.67	-1.24
δ_{29}	-1.52	-1.66	-1.83	-1.47
δ_{30}	-1.60	-1.51	-1.72	-1.28
δ_{31}	-2.30	-2.39	-2.57	-2.22

4. Empirical Application

We employ data from a national survey of preferences for pre-packaged dinners conducted by a major packaged goods manufacturer. Because of the proprietary nature of the data, we are restricted from revealing information about the specific brands studied in the survey. A total of 567 respondents provided information on needs, benefits sought, brand beliefs, and preferences expressed in two conjoint experiments. One of the authors was involved with the sponsoring company in designing the survey and conjoint studies to be able to explore the kind of issues we address in this paper. In particular, the exploratory work that was employed to generate needs and map them to benefits makes it an ideal setting to study the effect of product relevance on choice within an extended model framework.

Prior to the conjoint experiments, respondents rated 30 potential motivating conditions or needs associated with pre-packaged dinners on a 5-point rating scale, from Not at All (1) to Completely (5) describing the respondent. We operationalize active needs for a respondent by them providing a top-box indication of the need (e.g., a "5" on a 5-point scale). We conducted a sensitivity analysis and found no difference between a top box and a top-two box indicator.

Table 2 lists each of the 30 needs and the 30 corresponding benefits. The needs are concrete and specific to the given purchase context without being category or even brand-specific. The needs are generated within the motivational classification framework discussed in Fennell and Allenby (2014). The class structure helps to identify qualitatively distinct types of motivating conditions within the given context. There are 7 different classes within the framework, with overarching groups of classes representing moving away from an undesirable state (classes 1 through 3), moving toward the source of motivation (classes 4 and 5), and avoiding expected excessive cost or harm (classes 6 and 7). The framework is used only to generate candidate items for inclusion in the survey. Once the needs data are collected, the general framework is not used and analysis proceeds with the responses alone.

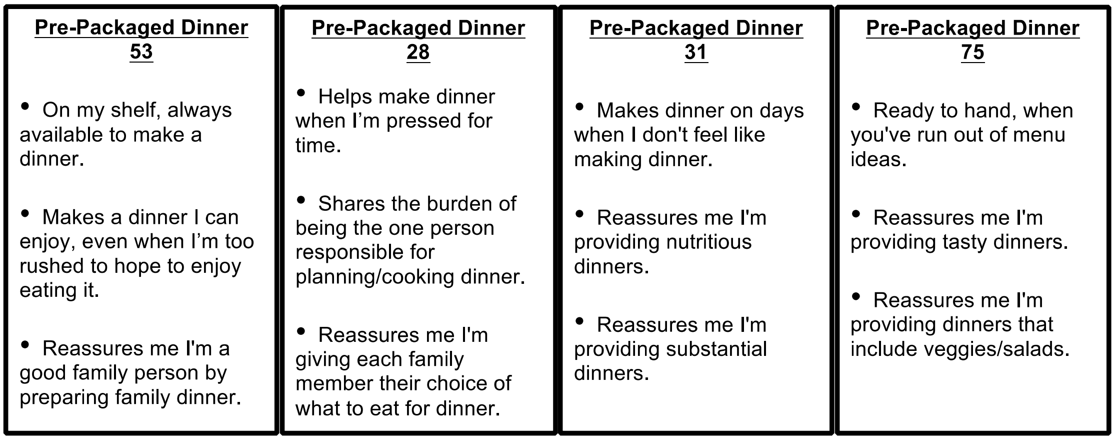
Classes 1 through 3 represent moving away from an undesirable state currently being experienced (e.g., need 5, "I was too rushed/pressured preparing dinner to enjoy eating it"), an undesirable state in the future (e.g., need 11, "I felt I'd be letting myself/my family down if I didn't provide a substantial dinner"), and a "default" undesirable state (e.g., need 16, "I felt that preparing weekday dinner is just a matter of routine"). Classes 4 and 5 represent an interest in mental exploration (e.g., need 19, "It interested me to tweak favorite family dinner recipes") and sensory enjoyment (e.g., need 20, "I was enjoying making dinner with foods of different textures"). Classes 6 and 7 represent avoiding expected excessive cost (e.g., need 26, "High cost kept me from serving a better dinner") and expected dissatisfaction (e.g., need 30, "I was upset to think there wouldn't be enough food for dinner"). The items included as needs in the study were generated from focus groups and packaging claims in the pre-packaged dinner category, using the above classification system as a guide. It is important to note, as show in Table 3, that the needs relate to the person while the benefits relate to the product. Once the items are generated, the structure used to guide their elicitation is ignored. Details of the motivational classes and their elicitation are provided in Fennell and Allenby (2014).

Table 2 Needs and Corresponding Benefits

No.	Needs	Benefits
1	I was worried that I hadn't anything available to make a dinner.	On your shelf, always available to make a dinner.
2	I was pressed for time to make dinner.	Helps make dinner when you're pressed for time.
3	It was a day when I just didn't feel like making dinner.	Makes dinner on days when you don't feel like making dinner.
4	I was worried that I was running out of menu ideas.	Ready to hand, when you've run out of menu ideas.
5	I was too rushed/pressured preparing dinner to enjoy eating it.	Makes a dinner you can enjoy, even when you're too rushed to hope to enjoy eating it.
6	I felt it a strain to have no relief from being the person to plan/cook dinner.	Shares the burden of being the one person responsible to plan/cook dinner.
7	I felt I'd be letting myself/my family down if I didn't provide a nutritious dinner.	Reassures me I'm providing nutritious dinners.
8	I felt I'd be letting myself/my family down if I didn't provide a tasty dinner.	Reassures me I'm providing tasty dinners.
9	I felt that preparing dinner is one way I show I'm a good family person.	Reassures me I'm a good family person by preparing family dinner.
10	I felt I'd be letting myself/my family down if I didn't give each family member their choice of what to eat for dinner.	Reassures me I'm giving each family member their choice of what to eat for dinner.
11	I felt I'd be letting myself/my family down if I didn't provide a substantial dinner.	Reassures me I'm providing substantial dinners.
12	I felt I'd be letting myself/my family down if I didn't provide a dinner that includes salad/veggies.	Reassures me I'm providing dinners that include veggies/salads.
13	I felt I'd be letting myself/my family down if I didn't provide a home cooked dinner.	Reassures me I'm providing home cooked dinners.
14	I felt I'd be letting myself/my family down if I didn't provide a dinner that includes meat.	Reassures me I'm providing dinners that include meat.
15	I felt I'd be letting myself/my family down unless everyone including my kids and spouse loved what I'd make.	Reassures me I'm providing dinners that everyone–kids and spouse–love.
16	I felt that preparing weekday dinner is just a matter of routine.	Suits my view that preparing weekday dinner is just a routine matter.
17	I felt that the conversation around the table at dinner would interest me.	Allows me appreciate dinner table conversation that interests me.
18	It interested me to make a dinner from different kinds of food, day to day.	Supports my interest in making many different kinds of food for dinner.
19	It interested me to tweak favorite family dinner recipes.	Supports my interest in tweaking favorite family dinner recipes.
20	I was enjoying making dinner with foods of different textures.	Allows me provide different textures of food for dinner.
21	I was relishing the added enjoyment of appetizing smells from a home prepared dinner.	Allows appetizing smells add to the enjoyment of dinner prepared at home.
22	I was concerned that my family would leave the dinner uneaten.	Helps ensure my family won't leave the dinner uneaten.
23	I was concerned that the kids would complain and refuse to eat dinner.	Helps ensure that kids don't complain and refuse to eat dinner.
24	I was concerned about burdensome clean-up afterwards.	Ensures there's no burdensome clean-up for me.
25	I was concerned about the burden of complicated or lengthy preparation.	Assures me I'm not burdened by complicated or lengthy preparation.
26	High cost kept me from serving a better dinner.	Allows me serve better dinners without high cost.
27	I was concerned about the problem too much salt in food would cause me/my family.	Guards against my family getting too much salt .
28	I was concerned not to prepare a depressing same old dinner.	Fights depressing same old thing dinner every time.
29	I was upset to think of having dinner food left over.	Reassures me there won't be food left over to upset me.
30	I was upset to think there wouldn't be enough food for dinner.	Reassures me there will be enough food.

After rating the 30 needs, each respondent completed 10 choice tasks each with 4 alternatives, where alternatives were benefit bundles. The alternatives were ranked, with 1 being the most preferred alternative. We thus explode the rank ordering to a depth of 3, which is at the recommended limit in Chapman and Staelin (1982). The "brand" of the pre-packaged dinners was fixed across choice tasks, so only the benefits changed. Only 3 of the 30 attributes were active for each of the 4 alternatives in each choice task. Figure 1 is an example of a single benefit-bundle choice task.

Figure 1: Example Choice Task



Pre-Packaged Dinner 53

- On my shelf, always available to make a dinner.
- Makes a dinner I can enjoy, even when I'm too rushed to hope to enjoy eating it.
- Reassures me I'm a good family person by preparing family dinner.

Pre-Packaged Dinner 28

- Helps make dinner when I'm pressed for time.
- Shares the burden of being the one person responsible for planning/cooking dinner.
- Reassures me I'm giving each family member their choice of what to eat for dinner.

Pre-Packaged Dinner 31

- Makes dinner on days when I don't feel like making dinner.
- Reassures me I'm providing nutritious dinners.
- Reassures me I'm providing substantial dinners.

Pre-Packaged Dinner 75

- Ready to hand, when you've run out of menu ideas.
- Reassures me I'm providing tasty dinners.
- Reassures me I'm providing dinners that include veggies/salads.

Respondents then indicated their beliefs regarding 6 brands in the pre-packaged dinner category. They indicated in a pick any/J format whether each brand provided each of the same 30 benefits used in the benefit-bundle conjoint, which also map one-to-one to the needs as shown in Table 2. After indicating their brand beliefs, each respondent completed 8 choice tasks each with 3 to 6 alternatives, depending on which brands they had either purchased or indicated were part of their consideration set. Each alternative in this second conjoint was a pair of one of the 6 brands and price. The alternatives were again ranked, with 1 being the most preferred alternative. We again explode the rank ordering to a depth of 3, the recommended limit in Chapman and Staelin (1982).

Following the structure specified in Equation (5) and the notation of ranked alternatives, $\sigma_{(j)h} = \exp(\gamma)$ if the respondent perceives the brand for the j-th ranked alternative as being able to provide the benefits that address their needs, with $\sigma_{(j)h} = 1$ otherwise. For example, and using Table 2 as reference, if a respondent only stated that the third need was applicable to him the last time he used a pre-packaged meal (i.e., he ranked "It was a day when I just didn't feel like making dinner" as the only need Completely (5) describing him) and he believes that the alternative's brand is able to provide the corresponding benefit—"Makes dinner on days when you don't feel like making dinner"—than that alternative would be relevant.

We offer the following model-free evidence for using our model in the context of this empirical application. If needs matter for preference certainty, as described above, one would expect the proportion of relevant choices for respondents with one or more needs should be largest for those items ranked first. We include the proportion of relevant choices for each rank in Table 3. Not only is the proportion of relevant choices highest for the first rank, but the next-largest proportions match for each subsequent rank as well.

Table 3 Proportion of Relevant Choices

Rank 1	0.053
Rank 2	0.039
Rank 3	0.034
Rank 4	0.016

5. RESULTS

Our results indicate differences in preference reliability for responses where needs are addressed versus responses where needs are not addressed. We compare three models. First, we estimate a standard exploded multinomial logit model, without including consumer needs, to serve as a baseline comparison. Second, we estimate an exploded multinomial logit model with the needs included as covariates in the random-effects specification of heterogeneity, as in Equation (2). Finally, we estimate the proposed heteroscedastic exploded multinomial logit mode. For each model, we used the first 9 choice tasks for estimation and the final choice task for out-of-sample fit. We ran each model for 80,000 iterations, using the final 4,000 iterations for inference.

Model fit is detailed in Table 4. Log-marginal density is a Bayesian measure of in-sample model fit. Hit probability is the posterior mean of the predicted probability for the observed alternative ranking. We see that the proposed model outperforms the alternative and baseline models in that it has the smallest value of the log-marginal density and the largest predictive hit probability. Interestingly, the baseline model performs slightly better than the alternative model. This serves as evidence that the mechanism by which relevance impacts consumer choice is indeed through choice certainty rather than part-worth heterogeneity.

Table 4 Model Fit

Model	LMD	Hit Probabilities
Baseline: $\sigma = 1$ and $\beta_{0h} = B_h\beta_{bh}$	-15,926.86	0.430
Alternative: $\sigma = 1$, $\beta_{0h} = B_h\beta_{bh}$, and $\beta_h = \Gamma' z_h$	-16,132.79	0.428
Proposed: $\sigma_{(j)h} = exp(\gamma)$ and $\beta_{0h} = B_h\beta_{bh}$	**-15,895.56**	**0.435**

We have two groups of responses, those where the part-worths are scaled by $\sigma_{(j)h} = 1$ and those where the part-worths are scaled by $\sigma_{(j)h}$ = exp(-0.378) = 0.685. Responses with $\sigma_{(j)h}$ = exp(-0.378) are more reliable as they give weight to the deterministic component of random utility. Responses where $\sigma_{(j)h} = 1$ are more unreliable as they give weight to the random component of random utility. The two groups of responses allow us to separate error based on genuine uncertainty regarding offerings from error that results from not actually caring about the choice alternatives and providing essentially random responses. Not distinguishing between these two very different sources of error results in possible parameter bias.

6. Conclusion

In this paper we show that respondents don't reliably state what they want when considering things they don't need. We accomplish this by developing a model that allows us to estimate the effect of relevance—of consumers' needs being addressed by perceived product benefits—on consistency in choice and utility expression. Our results indicate that the effect of relevance on choice is manifest in the scale of random utility and not in its location.

Our results also suggest the need for stricter screening criteria when studying aspects of an offering that might not be relevant to all respondents engaged in activities related to a product category. A study of outboard marine engines, for example, might focus on people owning boats for pleasure and recreation, and would naturally include features such as horsepower, acceleration, and fuel efficiency. Concerns about durability, however, might be more prevalent among people engaged in fishing where running over submerged logs is more likely. Obtaining accurate measures of preference for durability requires respondents for whom the issue of durability is relevant in their pursuits. Within our empirical application, we screened out respondents who didn't have any of the needs and found a marginal improvement in out-of-sample hit probability.

The study of motivating conditions and relevance has growing managerial importance. The amount of digitized individual-level data increasingly allows for relevant products and promotions to be offered to individual consumers. We need to better understand the role relevance plays in consumer choice to take full advantage of our growing access to individual-level information.

Marc R. Dotson

Greg M. Allenby

References

Allenby, Greg M, James L Ginter. 1995a. The effects of in-store displays and feature advertising on consideration sets. *International Journal of Research in Marketing* **12** 67–80.

Allenby, Greg M, James L Ginter. 1995b. Using extremes to design products and segment markets. *Journal of Marketing Research* **32**(4) 392–403.

Bagozzi, Richard P, Utpal Dholakia. 1999. Goal Setting and Goal Striving in Consumer Behavior. *Journal of Marketing* **63** 19–32.

Bhat, Chandra R. 1995. A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological* **29**(6) 471–483.

Chandukala, Sandeep R, Jeffrey P Dotson, Jeff D Brazell, Greg M Allenby. 2011a. Bayesian Analysis of Hierarchical Effects. *Marketing Science* **30**(1) 123–133.

Chandukala, Sandeep R, Yancy D Edwards, Greg M Allenby. 2011b. Identifying Unmet Demand. *Marketing Science* **30**(1) 61–73.

Chapman, Randall G, Richard Staelin. 1982. Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model. *Journal of Marketing Research* **19**(3) 288–301.

Dellaert, Benedict G C, Bas Donkers, Arthur Van Soest. 2012. Complexity Effects in Choice Experiment-Based Models. *Journal of Marketing Research* **49**(3) 424–434.

Fennell, Geraldine, Greg M Allenby. 2014. Conceptualizing and Measuring Prospect Wants: Understanding the Source of Brand Preference. *Customer Needs and Solutions* **1**(1) 23–39.

Fiebig, Denzil G, Michael P Keane, Jordan J Louviere, Nada Wasi. 2010. The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science* **29**(3) 393–421.

Griffin, Abbie, John R Hauser. 1993. The Voice of the Customer. *Marketing Science* **12**(1) 1–27.

Gutman, Jonathan. 1982. A Means-End Chain Model Based on Consumer Categorization Processes. *Journal of Marketing* **46**(2) 60–72.

Horsky, Dan, Sanjog Misra, Paul Nelson. 2006. Observed and Unobserved Preference Heterogeneity in Brand-Choice Models. *Marketing Science* **25**(4) 322–335.

Kim, Dong Soo, Roger A Bailey, Nino Hardt, Greg M Allenby. 2014. Benefit-Based Conjoint Analysis. Working Paper 1–40.

Kim, Yeung Jo, Jongwon Park, Robert S Wyer Jr. 2009. Effects of Temporal Distance and Memory on Consumer Judgments. *Journal of Consumer Research* **36**(4) 634–645.

Lavidge, Robert J, Gary A Steiner. 1961. A Model for Predictive Measurements of Advertising Effectiveness. *Journal of Marketing* **25**(6) 59–62.

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, Martin R Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science* **15**(2) 173–191.

Louviere, Jordan J, Deborah Street, Richard Carson, Andrew Ainslie, J R Deshazo, Trudy Cameron, David Hensher, Robert Kohn, Tony Marley. 2002. Dissecting the Random Component of Utility. *Marketing Letters* **13**(3) 177–193.

Luo, Lan, P K Kannan, Brian T Ratchford. 2008. Incorporating Subjective Characteristics in Product Design and Evaluations. *Journal of Marketing Research* **45**(2) 182–194.

Netzer, Oded, Olivier Toubia, Eric T Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M Feinberg, Eleanor M Feit, Sam K Hui, Joseph Johnson, John C Liechty, James B Orlin, Vithala R Rao. 2008. Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters* **19**(3–4) 337–354.

Rossi, Peter E, Robert E McCulloch, Greg M Allenby. 1996. The Value of Purchase History Data in Target Marketing. *Marketing Science* **15**(4) 321–340.

Salisbury, Linda Court, Fred M Feinberg. 2010. Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement, and Experimental Test. *Marketing Science* **29**(1) 1–17.

Strong, E K Jr. 1925. Theories of selling. Journal of Applied Psychology 9(1) 75–86.

Swait, Joffre, Jordan J Louviere. 1993. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *Journal of Marketing Research* **30**(3) 305–314.

Vakratsas, Demetrios, Tim Ambler. 1999. How Advertising Works: What Do We Really Know? *Journal of Marketing* **63**(1) 26–43.

van Osselaer, Stijn M J, Chris Janiszewski. 2012. A Goal-Based Model of Product Evaluation and Choice. *Journal of Consumer Research* **39**(2) 260–292.

Yang, S, Greg M Allenby, Geraldine Fennell. 2002. Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions. *Marketing Science* **21**(1) 14–31.