

**PROCEEDINGS OF THE
SAWTOOTH SOFTWARE
CONFERENCE**

October 2007

Copyright 2008

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

FOREWORD

We are pleased to publish these Proceedings of the 2007 Sawtooth Software Conference, held in Santa Rosa, California, October 17-19, 2007. We are grateful to the 180 attendees that enrolled this year. Even more papers were included (26) on the program than in previous years, so there was (and is) much to digest.

The focus of the 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 members of the audience. Topics included conjoint/choice analysis, market segmentation, MaxDiff, general web interviewing, scaling techniques, brand image research, data fusion, and hierarchical Bayesian estimation.

The papers are in the words of the authors, with generally very little copy editing done on our part. We express our gratitude to these authors for sacrificing time and effort toward making this conference one of the most useful and practical quantitative methods conferences in the industry. Beyond preparing PowerPoint slides and practicing a talk, it requires a special effort to write a paper (some can find it agonizing), and this written record will deliver value to the industry for years to come.

Sawtooth Software

January, 2008

CONTENTS

THE WEAKEST LINK: A COGNITIVE APPROACH TO IMPROVING SURVEY DATA QUALITY	11
<i>David G. Bakken, Harris Interactive</i>	
EVALUATING FINANCIAL DEALS USING A HOLISTIC DECISION MODELING APPROACH	23
<i>Paul Venditti, Donald Peterson, Matthew Siegel, General Electric</i>	
ISSUES AND CASES IN USER RESEARCH FOR TECHNOLOGY FIRMS.....	43
<i>Edwin Love, University of Washington School of Business Christopher N. Chapman, Microsoft Corporation</i>	
MINIMIZING PROMISES AND FEARS: DEFINING THE DECISION SPACE FOR CONJOINT RESEARCH FOR EMPLOYEES VERSUS CUSTOMERS	51
<i>L. Allen Slade, Covenant College</i>	
A CART-BEFORE-THE-HORSE APPROACH TO CONJOINT ANALYSIS.....	59
<i>Ely Dahan, UCLA Anderson School</i>	
TWO-STAGE MODELS: IDENTIFYING NON-COMPENSATORY HEURISTICS FOR THE CONSIDERATION SET THEN ADAPTIVE POLYHEDRAL METHODS WITHIN THE CONSIDERATION SET	67
<i>Steven Gaskin, Applied Marketing Science, Inc. (AMS), Theodoros Evgeniou, INSEAD Daniel Bailiff, AMS, John Hauser, MIT</i>	
A NEW APPROACH TO ADAPTIVE CBC.....	85
<i>Richard M. Johnson, Bryan K. Orme, Sawtooth Software, Inc.</i>	
HB-ANALYSIS FOR MULTI-FORMAT ADAPTIVE CBC	111
<i>Thomas Otter, Goethe University</i>	
EM CBC: A NEW FRAMEWORK FOR DERIVING INDIVIDUAL CONJOINT UTILITIES BY ESTIMATING RESPONSES TO UNOBSERVED TASKS VIA EXPECTATION-MAXIMIZATION (EM)	127
<i>Kevin Lattery, Maritz Research</i>	
REMOVING THE SCALE FACTOR CONFOUND IN MULTINOMIAL LOGIT CHOICE MODELS TO OBTAIN BETTER ESTIMATES OF PREFERENCE	139
<i>Jay Magidson, Statistical Innovations Inc., Jeroen K. Vermunt, Tilburg University</i>	
AN EMPIRICAL TEST OF ALTERNATIVE BRAND MEASUREMENT SYSTEMS.....	155
<i>Keith Chrzan, Doug Malcom, Maritz Research</i>	

ALTERNATIVE APPROACHES TO MAXDIFF WITH LARGE SETS OF DISPARATE ITEMS – AUGMENTED AND TAILORED MAXDIFF	169
<i>Phil Hendrix, immr, Stuart Drucker, Drucker Analytics</i>	
PRODUCT OPTIMIZATION AS A BASIS FOR SEGMENTATION	189
<i>Chris Diener, Lieberman Research Worldwide</i>	
JOINT SEGMENTING CONSUMERS USING BOTH BEHAVIORAL AND ATTITUDINAL DATA	199
<i>Luiz Sá Lucas, IDS-Interactive Data Systems</i>	
DEFINING THE LINKAGES BETWEEN CULTURAL ICONS.....	221
<i>Patrick Moriarty, OTX, Robert Maxwell, 12 Americans</i>	
CLUSTER ENSEMBLE ANALYSIS AND GRAPHICAL DEPICTION OF CLUSTER PARTITIONS	239
<i>Joseph Retzer, Ming Shan, Maritz Research</i>	
MODELING HEALTH SERVICE PREFERENCES USING DISCRETE CHOICE CONJOINT EXPERIMENTS: THE INFLUENCE OF INFORMANT MOOD	251
<i>Charles E. Cunningham, Heather Rimas, Ken Deal, McMaster University</i>	
DETERMINING PRODUCT LINE PRICING BY COMBINING CHOICE BASED CONJOINT AND AUTOMATED OPTIMIZATION ALGORITHMS: A CASE EXAMPLE.....	271
<i>Michael G. Mulhern, Mulhern Consulting</i>	
USING CONSTANT SUM QUESTIONS TO FORECAST SALES OF NEW FREQUENTLY PURCHASED PRODUCTS	279
<i>Greg Rogers, Procter & Gamble</i>	
REPLACEMENT MODELING: A SIMPLE SOLUTION TO THE CHALLENGE OF MEASURING ADDING AND SWITCHING IN A POLYTHERAPY CHOICE ALLOCATION MODEL	287
<i>Larry Goldberger, Adelphi Research by Design</i>	
DATA FUSION TO SUPPORT PRODUCT DEVELOPMENT IN THE SUBSCRIBER SERVICE BUSINESS.....	303
<i>Frank Berkers, Gerard Loosschilder, SKIM Group, Mary Anne Cronk, Philips Lifeline Systems</i>	
MULTIPLE IMPUTATION AS A BENCHMARK FOR COMPARISON WITHIN MODELS OF CUSTOMER SATISFACTION	313
<i>Jorge Alejandro, Kurt A. Pflughoeft, Market Probe</i>	

MAKING MAXDIFF MORE INFORMATIVE: STATISTICAL DATA FUSION BY WAY OF LATENT VARIABLE MODELING	327
<i>Lynd Bacon, YouGov/Polimetrix, Inc.</i>	
<i>Peter Lenk, Stephen Ross Business School at the University of Michigan</i>	
<i>Katya Seryakova, Knowledge Networks, Inc.</i>	
<i>Ellen Veccia, Knowledge Networks, Inc.</i>	
ENDOGENEITY BIAS – FACT OR FICTION?	345
<i>Qing Liu, University of Wisconsin, Thomas Otter, Goethe University</i>	
<i>Greg Allenby, Ohio State University</i>	
CBC/HB, BAYESM AND OTHER ALTERNATIVES FOR BAYESIAN ANALYSIS OF TRADE-OFF DATA	355
<i>Well Howell, Harris Interactive</i>	
RESPONDENT WEIGHTING IN HB	365
<i>John Howell, Sawtooth Software</i>	

SUMMARY OF FINDINGS

The thirteenth Sawtooth Software Conference was held in Santa Rosa, California, October 17-19, 2007. The summaries below capture some of the main points of the presentations. We hope that these introductions will help you get the most of the 2007 Sawtooth Software Conference Proceedings.

The Weakest Link: A Cognitive Approach to Improving Survey Data Quality (David G. Bakken, Harris Interactive): David reminded us that our inferences and theories of consumer behavior are only as good as the data on which they are based. As researchers, we often apply conventional wisdom, “judgment” and some empirical evidence in designing questionnaires. But, often in our haste to take studies to field, we fail to pretest and refine our instruments. David reviewed previous work by psychologists regarding how humans interact with surveys. The four step model of survey response involves comprehension, retrieval, judgment, and response. He advocated the use of “Think Aloud Pre-Testing” in which respondents (10-20 per wave) verbalize their thoughts while answering survey questions. These tests should be conducted over multiple days to allow survey changes to be implemented and re-tested. Based on many such tests, David offered some observations regarding how respondents interact with web-based surveys and how they can be improved. Current problem areas include: grid questions, survey navigation, error messages, multi-lingual surveys, and CBC questionnaires.

Evaluating Financial Deals Using a Holistic Decision Modeling Approach (Paul Venditti, Don Peterson, and Matthew Siegel, General Electric): Paul described a very interesting approach that he and his co-authors are implementing within GE to evaluate complex financial deals. In the past, analysts have spent many hours evaluating financial deals and presenting the details of those deals to a committee of three individuals. Paul described how the characteristics of those deals could be defined using about 20 “conjoint” attributes. A modified ACA survey was developed to study three key individuals at GE who approve deals. The standard stated importance question in ACA was substituted with a constant-sum question implemented via an Excel worksheet. The final part-worth utilities were further modified by implementing a few non-compensatory rules (red flags). A market simulator based on the three respondents was found to be highly predictive of whether deals were approved or rejected in the months following the surveys (accuracy of about 80%). Paul’s work demonstrated that effective conjoint models (to profile tiny populations) can be built using tiny sample sizes. Conjoint analysis can provide good data for implementing sophisticated decision support tools in non-traditional contexts.

Issues and Cases in User Research for Technology Firms (Edwin Love, University of Washington School of Business, and Christopher N. Chapman, Microsoft Corporation): Edwin and Christopher described how conducting market research for technology products presents unique challenges. For example, innovative features are often not well-understood by respondents, and different user groups will have different levels of understanding. Also, features might not actually yet exist while the research is being conducted. The presenters commented that vague descriptions of attributes such as “easy setup” can skew user responses (toward expressing strong preference for nondescript features), and the results create the illusion of specific value where none may exist. They further recommended segmenting respondents based on product experience: owners vs. intenders. Edwin and Christopher illustrated the challenges of

conducting market research for technology products via three case studies: a digital pen project, a webcam, and a digital camera.

Minimizing Promises and Fears: Defining the Decision Space for Conjoint Research for Employees versus Customers (L. Allen Slade, Covenant College): Conjoint analysis can be a valuable tool in both consumer and employee research. However, the researcher must recognize the key differences in how the firm interacts with the respondents. Allen affirmed that customers are less interdependent with the firm than are employees. And, different employees (depending on role and experience/training) are more highly interdependent with the firm than others. With employee research, the worry is of creating false promises of rewards or unwarranted fears of takeaways. Allen suggested that researchers ask themselves three key questions prior to including something in a conjoint survey for employees: 1) Would we be willing to actually do this?, 2) How does this intervention compare to the others we are considering?, and 3) How would an employee or customer react to taking this survey? Using an actual case study at Microsoft (total rewards optimization), Allen illustrated how applying these three questions led to effective research without undue promises or fears.

A Cart-Before-the-Horse Approach to Conjoint Analysis* (Ely Dahan, UCLA Anderson School): With traditional conjoint studies, respondents are often asked to complete long surveys, they are required to rate products they don't like, and the resulting part-worth utilities often contain reversals in the utilities. Ely described a novel, computer-administrated and adaptive method of employing a traditional full-profile conjoint design. Rather than estimate part-worth utilities after respondents take the surveys, CARDS (conjoint adaptive ranking database system) begins with a researcher-constructed database of typically thousands of potential sets of consistent part-worth utilities. Respondents are shown a set of product concepts and asked to choose which products they prefer. After the respondent provides a few answers, the database of utilities is queried to determine if certain product concepts that haven't yet been evaluated are clearly inferior (and should not be chosen next in order). Those products are deleted from the screen, allowing respondents to focus on those product concepts that are relevant to identifying which set of utilities best fits them, while forcing respondents to maintain consistent ordering. The benefit is much shorter questionnaires. The downsides are that early answers matter a lot, and there is no real error theory. Plus, the quality of the results depends on how well researchers can develop the database of potential sets of utilities.

(*Winner of Best Presentation award, based on attendee ballots.)

Two-Stage Models: Identifying Non-Compensatory Heuristics for the Consideration Set then Adaptive Polyhedral Methods within the Consideration Set (Steven Gaskin, AMS, Theodoros Evgeniou, INSEAD, Daniel Bailiff, AMS, and John Hauser, MIT): Steven reviewed the scientific evidence that suggests that people buy products by first forming a consideration set and then choosing a product from within the consideration set. This two-stage approach helps people deal with a large number of alternatives in the choices they face. By reflecting this process in our choice models, Steven argued that we can more accurately model choices, create more realistic and enjoyable surveys, and handle more features than conventional CBC. He presented a survey design in which respondents may use non-compensatory (cut-off rules) to form consideration sets. Respondents are then asked to tradeoff considered products within a more standard-looking CBC task. He and his co-authors employed FastPace CBC to estimate

the utilities for the n most important compensatory features for each respondent. Steven reported results showing that respondents preferred the adaptive survey over standard CBC.

A New Approach to Adaptive CBC (Rich Johnson and Bryan Orme, Sawtooth Software): Existing CBC questionnaires have weaknesses: they are viewed as tedious and not very focused on the particular needs of each respondent. The experimental plans have assumed compensatory behavior, and previous research has shown that many respondents apply non-compensatory heuristics to answer conjoint questionnaires. Rich and Bryan presented a new technique for adaptive CBC that helps overcome these issues. Their approach mimics the purchase process of formulating a consideration set using non-compensatory heuristics (such as “must have” or “must avoid” features), followed by a more careful tradeoff of alternatives within the consideration set using compensatory rules. This new approach involves three core stages: 1) Build-Your-Own (BYO) Stage, 2) Screening Stage, and 3) Choice Tasks Stage. They conducted a split-sample experiment comparing the new approach to traditional CBC. They found that respondents liked the adaptive survey more and felt it was more realistic—even though it took about double the time as traditional CBC. Furthermore, part-worths developed from ACBC were more predictive of holdout tasks than traditional CBC, despite the methods bias in favor of CBC for predicting the CBC-looking holdouts.

HB-Analysis for Multi-Format Adaptive CBC (Thomas Otter, Goethe University): The three-stage interview proposed by Johnson and Orme is innovative, but the formulation of a model extracting the common preference information is a challenge. Thomas first showed that such a model is required, as simply discarding any of the data collected before the CBC part results in inconsistent inferences in an HB setting. Thomas then investigated different models: a multinomial likelihood for all parts of the interview allowing for task-specific scale factors, task-specific “wiggles” in the preference vector using the same likelihood, a binary logit likelihood for the screener part and a multichoice likelihood for this same part. Thomas found that the scale factor did vary considerably between the sections. However, accounting for task specific scales had only a small effect on the predictive ability of the models. Moreover, his results suggest that a binary logit or a multichoice likelihood for the screener part of the interview are preferable to the explosion into multinomial choices both in terms of the implied story about how the data are generated and the empirical fits.

EM CBC: A New Framework for Deriving Individual Conjoint Utilities by Estimating Responses to Unobserved Tasks via Expectation-Maximization (Kevin Lattery, Maritz Research): Kevin demonstrated how EM algorithms can be used to estimate individual-level utilities from CBC data. EM is often applied in missing values analysis. In the context of CBC, each respondent could be viewed as having been shown all the tasks in a very large design plan, but having completed only a subset of them. The missing answers are imputed via EM. Once missing answers have been imputed, there is enough information available to estimate part worths for each individual. Utility constraints may be implemented as well. Kevin faced a few challenges in implementing EM for CBC. He found that if he allowed EM to iterate fully to convergence, overfitting would occur. Therefore, he relaxed the convergence criterion. Kevin also found that the estimated probabilities for the tasks respondents did versus those that were missing varied in their means and standard deviations. So he adjusted the results from each task so that means and variances of the missing data were comparable to the observed data. He then repeated the EM process again until the missing data converged. Kevin compared utilities

estimated under EM to those estimated via HB, and found that the EM utilities performed as well or better than HB utilities for three data sets.

Removing the Scale Factor Confound in Multinomial Logit Choice Models to Obtain Better Estimates of Preference (Jay Magidson, Statistical Innovations, and Jeroen K. Vermunt, Tilburg University): Jay reintroduced the audience to the issue of scale factor. The size of the parameters in MNL estimation is inversely related to the amount of certainty in the respondents' choices. Because different groups of respondents may have different scale factors, it is not theoretically appropriate to directly compare the raw MNL estimates between groups. Jay showed how such comparisons can lead to incorrect conclusions. He then turned attention toward an extended Latent Class choice model to isolate the scale parameter. Using that model, he showed how latent class segmentations can differ for real data sets as compared to the generic latent class model that doesn't separately model scale. In one particular comparison, Jay found that the amount of time respondents spent answering a CBC questionnaire was directly related to segment membership from standard latent class estimation (without estimating the scale factor).

Jay also demonstrated how scale estimation can be incorporated into DFactor Latent Class models. Jay concluded that removing the scale confound in latent class modeling will result in improved estimates of part-worths and improved targeting to relevant segments based on an improved understanding of segment preferences and levels of uncertainty.

An Empirical Test of Alternative Brand Measurement Systems (Keith Chrzan and Doug Malcom, Maritz Research): Keith and Doug presented results from three commercial studies that compared different ways of collecting brand image data. Those methods included: Likert ratings, comparative ratings, MaxDiff, pick any, semantic differential, and yes/no scaling. They argued that the brand image measurement system should produce 1) credible brand positions (face validity), 2) strong differences among brands (discriminant validity), and 3) powerful predictions of brand choice (predictive validity). The first two research studies they reported on demonstrated that Likert ratings and pick any data were generally inferior to the other methods. The third study they reported compared semantic differential, comparative ratings, yes/no, and pick any data. They concluded that, of those four methods, comparative ratings had the most discriminating power, followed by semantic differential. Pick any data measured little beyond the halo effect (a complicating issue wherein brands/objects liked overall tend to get higher ratings across the board on the attributes). To help control for the Halo Effect, the authors double-centered the scores prior to making comparisons.

Alternative Approaches to MaxDiff with Large Sets of Disparate Items—Augmented and Tailored MaxDiff (Phil Hendrix, immr and Stuart Drucker, Drucker Analytics): Phil and Stuart investigated some enhancements to standard MaxDiff questionnaires to help deal with large numbers of items while still achieving strong individual-level scores. The authors argued that with more than about 40 items, MaxDiff becomes very tedious for respondents if individual-level estimates are required. To deal with this issue, the authors proposed that respondents first perform a Q-Sort task, wherein they drag-and-drop items into one of K buckets (they used 4 buckets in their research). The information from the Q-Sort task can be added to the MaxDiff information to improve the estimates. The Q-Sort task can also be used to create customized MaxDiff questions that principally draw on items of greatest preference/importance. Phil and Stuart conducted a split-sample study comparing standard and two forms of augmented MaxDiff exercises. They found that overall the aggregate parameters were very similar across the

methods. But, both forms of augmented MaxDiff exercises outperformed ordinary MaxDiff in terms of holdout predictions. They also found that respondents found the Q-Sort + MaxDiff methodology more enjoyable than standard MaxDiff alone.

Product Optimization as a Basis for Segmentation (Chris Diener, Lieberman Research Worldwide): Chris motivated his presentation by reviewing the strategic goals and outcomes of traditional segmentation approaches. With attitudinal segmentations, one finds strong segments in terms of attitudinal differences, but those differences often do not translate into segments that differ strongly in terms of product preferences. With segmentation based on product features, the hope is that the segments have targetable differences and that the preferences translate to profitable product line decisions. If product optimization is used as the focus, then there is a stronger linkage with profitable product line decisions. Of all the methods of optimization, Chris stated that he prefers Genetic Algorithms. But, Chris pointed out that segmentation based on product optimization provides no guarantee that the segments will demonstrate targetable differences in terms of attitudes, media usage, or demographics. To improve the odds that the segments are useful, Chris advocated data fusion processes which combine information from attitude segmentation and product optimization segmentation, especially when the strategic priority is on product development and you are confident in being able to find an attitudinal story.

Joint Segmenting Consumers Using both Behavioral and Attitudinal Data (Luiz Sa Lucas, IDS Market Analysis): Luiz discussed segmentation methods that incorporate both behavioral and attitudinal data. Behavior data alone are often not satisfactory to use in segmentation schemes, because the segments do not necessarily map to anything useful in terms of descriptive demographics or attitudinal data. By the same token, attitudinal data alone are not sufficient because attitudes don't necessarily correlate strongly with behaviors. Luiz reviewed multiple procedures for incorporating both behavior and attitudinal data in segmentation, including Reverse Segmentation, Weighted Distance Matrices, Concomitant Variables Mixture Models, Joint Segmentation, and LTA models. Luiz finished by discussing different fit metrics for determining the appropriate number of clusters.

Defining the Linkages between Cultural Icons (Patrick Moriarty, OTX and Scott Porter, 12 Americans): Patrick and Scott described a mapping methodology in which cultural icons (celebrities, brands, politicians) are placed within a perceptual map. The data are in part driven by a MaxDiff questionnaire. The goal is to provide a unique understanding of the strength of linkage between brands, personalities, and media properties based on consumer attraction. Their research identified that religion and marital status are the two social identities that on average most define individuals. But, identity may also be measured by the degree to which people express connection with cultural icons. The authors explained that cultural icons can also be measured and characterized, in terms of four key components: Recognition, Attraction, Presence, and Polarization. As an example of how their mapping methods can drive strategy, they showed relationships between either Hillary Clinton or Rudy Giuliani, segments of the population, and popular consumer brands.

Cluster Ensemble Analysis and Graphical Depiction of Cluster Partitions (Joseph Retzer and Ming Shan, Maritz Research): Joe described a relatively new technique in unsupervised learning analysis called Cluster Ensemble Analysis that has been suggested as a generic approach for improving the accuracy and stability of cluster algorithm results. Cluster ensembles begin by

generating multiple cluster solutions using a “base learner” algorithm, such as K-means. Multiple solutions may be generated in a variety of ways. The basic idea is to combine the results of a variety of cluster solutions to find a consensus solution that is representative of the different solutions. Joe further demonstrated how the quality of cluster solutions can be graphically depicted in terms of Silhouette plots. The silhouette shows which objects lie well within the cluster and which are somewhere in between clusters. He finished by showing how cluster ensemble analysis can improve cluster results for a particularly difficult sample data set that has non-spherical clusters.

Modeling Health Service Preferences Using Discrete Choice Conjoint Experiments: The Influence of Informant Mood (Charles Cunningham, Heather Rimas, and Ken Deal, McMaster University): Chuck presented the results of a research study that investigated how depression influences performance on discrete choice experiments designed to understand patient preferences. Previous evidence in the literature suggests that people with depressive orders can have impaired information processing and a related host of decision making deficits. Because Chuck and his co-authors often use discrete choice experiments in health care planning issues, and because the incidence of depression is relatively high within populations they often survey, these issues were of interest to them. They found that although depression did not increase inconsistent responding to identical holdout tasks (test-retest reliability), it did influence health service preferences and segment membership. Chuck also reviewed basic principles for designing and analyzing holdout questions.

Determining Product Line Pricing by Combining Choice Based Conjoint and Automated Optimization Algorithms: A Case Example (Michael Mulhern, Mulhern Consulting): Mike presented the results of a recent study where the purpose was to develop an optimal pricing strategy for a product line decision. Six price levels were included in the study, and based on the plot of average utilities, there appeared to be two “elbows” in the price function. The elbows seemed to represent optimal pricing points for mid-price and a higher-price products. Mike used the Advanced Simulation Module to conduct optimization searches to maximize revenue. He found that the optimization routines also identified those same two price points as optimal positions. The different optimization algorithms (exhaustive, grid, gradient, stochastic, and genetic) produced identical results irrespective of the starting points (with the exception of the gradient search method, which had some inconsistencies). Mike’s client also asked whether the optimal price points would change depending on different assumptions for the base case. Altering the base case and re-running the optimizations revealed similar recommendations in most cases. Mike was able to report what the client eventually did and how actual sales volume compared to the simulation’s predictions. The client followed some of the recommendations, but ignored others. The sales results suggest that ignoring the recommendations provided by the optimization simulations was costly. A poorly positioned mid-price product foundered, as would have been predicted by the model.

Using Constant Sum Questions to Forecast Sales of New Frequently Purchased Products (Greg Rogers, Procter & Gamble): Greg compared two relatively common methods for measuring buyer intent for an FMCG category: CBC and constant sum allocations (both computer-administered). Not surprisingly, the constant sum allocation (out of 10 purchases) data were more “spiked” on the 0%, 10%, 50%, and 100% allocation probabilities relative to the probabilities projected from the pick-one CBC data. Greg expanded the analysis to include a Dirichlet model (to estimate base trial for a new item) that incorporated the issues of trial and

frequency. Greg concluded that analyzing the brand choices from simple constant sum scales using a Dirichlet model results in comparable base trial estimates to those derived by CBC. This finding has implications for researchers that cannot use other methods like purchase intent (requires database to interpret) or CBC (can be relatively complex and costly) to estimate trial for new products.

Replacement Modeling: A Simple Solution to the Challenge of Measuring Adding and Switching in a Polytherapy Choice Allocation Model (Larry Goldberger, Adelphi Research by Design): In pharmaceutical research, doctors sometimes will prescribe multiple drugs to treat particular condition. When this occurs, the standard allocation models that assume that each patient is assigned a single drug therapy is violated. When this happens, the allocations may sum to more than 100%, so the allocation total is no longer fixed. Larry demonstrated a Polytherapy Allocation Model that does not assume that the total sum allocated per task is 100%. The proposed solution models the likelihood that a new product will substitute for an existing product, and does not constrain the sum to 100%. Larry also reviewed other common approaches to the problem, and discussed the limitations. He discussed the common binary logit approach to the problem, and how the cross-effects can often lead to reversals.

Data Fusion to Support Product Development in the Subscriber Service Business (Frank Berkers, Gerard Loosschilder, SKIM Group, and Mary Anne Cronk, Philips Lifeline Systems): Data fusion can involve combining different datasets to learn more than the original datasets had to offer individually. The authors explained how they used data fusion to help develop new strategies with respect to a subscriber service for Lifeline monitoring (the leader in North America for Personal Emergency Response Systems). Specifically, the authors were able to develop a plan of action to approach customers with increased communication regarding specific offers depending on the pattern of signals received from the subscriber. This provided an “early warning system” that would flag subscribers as in danger of deactivating their service. By implementing this system, subscriptions could be prolonged, resulting in greater profitability to the firm. The combination of behavioral patterns and background characteristics gave a better and clearer warning of imminent deactivation, and the type of deactivation, than the separate data sources could provide. Furthermore, the combined information provided greater clarity in deciding what services to offer, and when to offer them to subscribers.

Multiple Imputation as a Benchmark for Comparison within Models of Customer Satisfaction (Jorge Alejandro and Kurt Pflughoeft, Market Probe): Kurt emphasized that many studies must deal with missing data, and the degree of missingness can be significant. Different missing value routines will lead to different degrees of bias and imprecision for statistical estimates. The authors examined a variety of techniques to deal with or impute missing data: Casewise and Pairwise deletion, the Missing Indicator Method, Mean Substitution, Regression-based Imputation, Expectation Maximization (EM), and Multiple Imputation. They used a real dataset involving customer satisfaction for a bank, and induced missingness. After deleting values to induce missingness, they estimated regression models and compared the results to the same models prior to having missing data. They determined that Multiple Imputation appeared to be the best performer in terms of reducing bias and generally was more realistic in terms of standard errors. The Missing Indicator Method and Overall Mean substitution were generally biased, as the authors expected. Point estimates of EM worked well with regression, however SPSS’s imputed dataset was biased. Pairwise deletion performed well in this experiment in estimating stable beta coefficients.

Making MaxDiff More Informative: Statistical Data Fusion by way of Latent Variable Modeling (Lynd Bacon, YouGov/Polimetrix, Inc., Peter Lenk, University of Michigan, Katya Seryakova and Ellen Veccia, Knowledge Networks): Lynd demonstrated three different ways to think about coding and estimating MaxDiff data: differences coding, coding as two separate choice tasks, and rank imputed exploding logit. All three methods produced very similar results. The authors then turned their attention to a weakness in MaxDiff experiments: the scores are scaled with respect to an arbitrary intercept (rather than a common origin) for each respondent. This makes it hard to compare a single score from one respondent to a single score from another. They applied a different model (cutpoint model for ratings) which allows them to estimate the scores for items on a common scale with a common origin. They demonstrated how using the new model can improve the ability of researchers to identify respondents to target according to overall preference for a feature. Another point they emphasized is that the lack of scale origin issue also extends to attributes within standard discrete choice methods. The new model can be applied in those situations as well.

Endogeneity Bias—Fact or Fiction? (Qing Liu, University of Wisconsin, Thomas Otter, Goethe University, and Greg Allenby, Ohio State University): In theoretically proper applications of regression modeling, the independent variables are truly independent. However, in some market research applications, the independent variables are not truly independent. Examples include sequential analysis, time series models with lagged dependent variables, and Adaptive Conjoint Analysis (ACA). Greg suggested that endogeneity bias will matter whenever an adaptive procedure is used to learn about respondents (so that informative questions can be determined) and these data are excluded from analysis. However, with ACA, all of the information from each respondent is included in the estimation. Endogeneity bias only depends on whether you rely on the likelihood principle, and therefore, explained Greg, “being Bayes” or not matters. The presence of endogenously determined designs in ACA doesn’t affect the likelihood of the data. Although a small degree of bias is introduced in ACA due to endogeneity, the bias is typically quite small and ignorable.

CBC/HB, Bayesm and other Alternatives for Bayesian Analysis of Trade-off Data (Well Howell, Harris Interactive): HB has become a mainstream tool for analyzing results of DCM and related techniques (such as MaxDiff). There are a number of tools available for HB estimation, including Sawtooth Software’s CBC/HB product, bayesm (R package), WinBUGS, and Harris Interactive’s Hlhbmk1 model. Well used three data sets to compare the different tools in terms of in-sample and out-of-sample fit. The speed of the different systems varied quite a bit, with CBC/HB being significantly faster than the other methods. Both the in-sample and out-of-sample fit was strongly affected by the tuning of the priors (the amount of shrinkage permitted). Tools other than CBC/HB offer some more advanced diagnostics and model specifications, including Gelman diagnostics for convergence, and respondent covariates in the upper level model.

Respondent Weighting in HB (John Howell, Sawtooth Software): When samples include subgroups that have been oversampled, it has been reported that this can pose some problems for proper HB estimation within CBC/HB software (which assumes a single, normally-distributed population). John investigated the degree to which this is a problem, and potential solutions. Using simulated data, John demonstrated that when subsamples are dramatically oversampled, it causes the means of smaller groups to shrink disproportionately toward the larger groups. This biases the sample means for the under-represented groups, and harms the accuracy of market

simulations. John found that much of the problem is due to diverging scale factors between smaller and larger subgroups. The scale for the oversampled groups is expanded, leading to stronger pull on the overall sample mean. John found that normalizing the scale post hoc can largely control this issue. He also found that implementing a simple weighting algorithm within HB (computing a weighted alpha vector) can potentially improve matters further when there are extreme differences in sample sizes between subgroups. John suggested that other methods he didn't investigate may improve estimation when some groups are oversampled, including developing models that estimate individual-level scale factors, models that involve less shrinkage (Students-t prior) or models that utilize multiple upper-level models. He concluded that regardless how the shrinkage problem is solved, models should be tuned for scale at either the individual or group level.

THE WEAKEST LINK: A COGNITIVE APPROACH TO IMPROVING SURVEY DATA QUALITY

DAVID G. BAKKEN
HARRIS INTERACTIVE

The survey questionnaire has been an important tool for gathering data since at least the early part of the 20th century. The questionnaire is a form of *verbal interrogation* (whether oral or written), and throughout the history of survey research, investigators have been concerned with the veracity of the information provided in response to the survey questions.

Considerable effort has been devoted to studying the sources of error in survey data. Even if we ignore representativeness error attributable to sampling processes, selection and non-response biases, a large number of factors may impact the veracity of answers to survey questions. A partial list of areas that have been investigated include asking for sensitive information such as income or history of sexual activity, the effects of question ordering, and the impact of different types of measurement *scales*. Empirical investigation, common sense and practical or anecdotal experience have, over the history of survey research, resulted in a shared set of practices for designing survey questionnaires. One example is the admonition to “avoid double-barreled questions.” A double-barreled question is one that asks a respondent to report simultaneously on two (or more) separate states, such that the possible answers to the question are not *collectively exhaustive*. Consider this question: “When you travel for business, do you usually stay at mid-priced hotels and dine at value restaurants?” There are only two implied responses, “yes” and “no.” However, there are more than two possible *states of being* represented by this question because it asks about two *independent* events: choice of hotel and choice of restaurant. Individuals who satisfy both conditions in the question should answer “yes.” All others (those who satisfy one or the other or neither condition) logically should answer “no.” If the researcher wants to know only the incidence of the joint occurrence of the two conditions, I suppose this is acceptable, as long as respondents follow the logic in answering the question. In most cases, the researcher probably wants to know something about each condition, and it is quite possible that some or many respondents do not apply the strict logical rule to answering the question.

This second possibility, that the respondent applies some other decision rule or process in arriving at an answer to the question, is the focus of this paper. Despite research on the sources of error in survey questionnaires, questionnaire design remains more art than science, and only recently have researchers begun to explore systematically the link between survey response and the cognitive processes that respondents use to generate those responses.

THE WEAKEST LINK?

Market researchers rely on survey data to build models of consumer decision-making and behavior, to draw inferences, and to predict response to marketing actions like the introduction of a new product or implementation of a price change. Many of the mathematical techniques used to analyze survey data, such as classical linear regression, implicitly or explicitly assume that the data are *measured* without error. Thus, while there might be error in the model (that is, prediction error), that error is a function of the data-generating process rather than the

measurement process. A simple regression model predicting income from years of education, for example, assumes that the data values for years of education and income are *accurate*. While it is true that *measurement error* can contribute to prediction error, we almost always use the simplifying assumption that any measurement errors are small and randomly distributed about the *true* data values. In some cases, suspect data values may be eliminated from the modeling. If the data do not satisfy this assumption, any inferences based on analysis of the data will be suspect. For that reason, we consider the quality of survey response to be the “weakest link” in the chain of components that comprise the typical survey-based market research study.

One reason that survey responses are the weakest link is that, despite efforts to study the response process, most research into survey response has focused on *observable* aspects of the process, such as differential predictive performance of various attitude measurement scales rather than the *unobserved* cognitive processes that respondents employ to generate answers to survey questions. All survey researchers will benefit from a greater understanding of the way that respondents’ cognitive processes affect survey data. Moreover, market researchers will benefit from understanding the power of think-aloud pre-testing for detecting and fixing survey questions that are likely to generate poor quality data.

Throughout this paper, I am primarily interested in the impact of survey design on *internal* and *construct validity* rather than external validity (the representativeness or generalizability of the responses from the sample to a population). Construct validity is of particular interest. If we hope to test propositions about hypothetical constructs such as attitudes towards a brand or customer loyalty, we need to have confidence that the data we observe reflect the construct rather than an artifact of the cognitive processes respondents use to generate answers to survey questions.

MODELS OF THE SURVEY RESPONSE PROCESS

One barrier to developing better survey questionnaires is lack of a comprehensive theoretical framework for testing propositions about the cognitive mechanisms that underlie survey response. More general theories of memory, information processing, and judgment are useful but not quite specific enough to guide our efforts to improve the quality of survey data. One such model has been proposed by Tourangeau, Rips and Rasinski (2000). This model decomposes the survey response process into four distinct steps: comprehension, retrieval, judgment, and response. The *comprehension* step encompasses those cognitive processes that a respondent uses to determine the *meaning* of the survey question. The *retrieval* step comprises those mechanisms by which we search long term memory for information relevant to forming a response to the question. In the *judgment* step, the respondent assesses the accuracy of the retrieved information and draws inferences based on the retrieved information. Finally, in the *response* step, the respondent matches his or her *internally generated* response with the alternatives made available in the survey.

Other models of the survey response process have been proposed. These models are not necessarily inconsistent with the four-step model of Tourangeau, *et al.* For example, so-called “high road/low road” theories (e.g., Cannell, Miller and Oksenberg, 1981) posit that some respondents follow something like the four-step process (the high road) while others, in effect, short circuit the process by looking for *external cues* (such as question context) to inform their survey responses. According to Krosnick and Alwin (1987), some respondents *satisfice*; that is,

they use a general response strategy to provide a *plausible* response. Other respondents follow a more comprehensive process to arrive at an *optimal* response. Another “two-track” theory (Strack and Martin, 1987) applies specifically to attitude measures, proposing one strategy for responses based on an existing judgment and a second strategy for new judgments derived at the time of questioning.

These various models of survey response are important because they address the unobserved *cognitive* aspects of the process and provide a framework for improving the design of survey questions. In the next section of this paper we review the four components of the Tourangeau *et al.* model of survey response, considering some of the empirical evidence and underlying theory, and the implications for questionnaire design.

Comprehension

Comprehension is the process by which a survey respondent attends to the question and instructions, represents the logical form of the question, identifies the information sought, and links key terms to relevant concepts.

Intuitively, we can see that a question is a request for information. More specifically, a question is a linguistic object that defines an *uncertainty space*. As an example, the simple question “Guess what?” defines a potentially infinite uncertainty space. In forming a question, the survey designer has some notion—explicit or implicit—of the boundaries of the uncertainty space. The wording of the question must communicate those boundaries to the respondent. Definition of the uncertainty space helps the respondent develop a retrieval strategy (step two in the four-step model). As we will see, most of the problems that occur in terms of comprehension are related to misspecification of the uncertainty space.

The comprehension step requires an initial reading of the question in linguistic terms. According to Rips (1995), this involves forming a *representation of the question* as well as a *representation about the question*. The representation of the question is a function of the linguistic structure of the question (and should be more or less consistent across competent speakers of a language). All languages have some way of indicating that a linguistic object (a word, phrase or sentence) is a question, as well as a means of indicating the missing information that is requested. English, French, Spanish, and Italian (as examples) typically use a combination of subject-verb inversion and *wh-* words (who, what, where, etc.) to create an interrogative sentence. The *wh-* word implies another word that would be located elsewhere in a declarative sentence. For example: “Who is that?” is an inversion of the declarative: “That is [who].” In an age when many surveys are administered on a global basis, it is important to understand that other languages may have different structures for questions. For example, in Japanese a question is indicated by adding a marker (“ka”) to the verb, which always appears at the end of the sentence or clause. That is, there is no subject-verb inversion when asking a question in Japanese.

The *representation about the question* involves inferences based on the sentence plus other information. While we may be able to recognize the form of the question without considering contextual cues, we often rely on such cues to resolve any ambiguous elements in the question. In order to arrive at this representation, the respondent must understand the semantic elements as well as the structural elements.

The *representation about the question* defines the uncertainty space. Factors that lead to “incorrect” definition of the uncertainty include *presupposition*, *vagueness*, and *cognitive complexity*. Presupposition (and a related factor, *unfamiliarity*) leads to an uncertainty space that is not *exhaustive*. Leading questions are a form of *false* presupposition, as in this example: “Which baseball team do you hate more, the Yankees or the Red Sox?” The question implies that *hate* is the only emotion that is attached to these two teams, which is not the case. Another form of presupposition occurs in questions where the respondents do not have the information needed to answer the question. When the respondent is unfamiliar with the required information, the uncertainty space defined by the question does not overlap the respondent’s *search space*.

Vagueness and ambiguity create an uncertainty space that is not *exclusive*. Common problems include conceptual ambiguity or vagueness, as in these adjectives that might be used to describe a fashion brand, “engaging,” “sophisticated” and “stylish.” Vague quantifiers can affect the validity of measurement scales. A common scale for measuring future intention uses quantifiers such as “Definitely will,” “Probably will,” “Might or might not” and so forth. While “definitely” might be unambiguous, “probably” and “might or might not” are vague quantifiers.

Attempts to counter problems stemming from presupposition and vagueness can result in cognitive complexity. Consider these two examples:

Thinking of the hotel brands that are in your usual price range, and assuming that they were all available and equally convenient in their location, which one would be your first choice to stay in for a future business trip?

Thinking of your three most recent hotel stays, to the nearest \$10, what is the average room rate you have paid per night, that is, excluding other charges like taxes, phone calls, room service, etc.?

In these two examples, subordinate clauses have been used to set specific boundaries around the uncertainty space. This may be necessary in some cases, but researchers should at least consider the increased complexity as they craft survey questions.

Retrieval

The retrieval step encompasses the cognitive processes of searching memory for the requested information. While a review of the current research and understanding of memory and retrieval is beyond the scope of this paper, a few key points will help survey researchers appreciate the role that retrieval plays in survey response.

Most survey questions require some retrieval of past experience from memory. Various theories have been proposed to explain *episodic memory*. Tulving (1983) proposed a relatively simple unstructured memory for specific events in which episodes are retrieved in their entirety. Others have proposed various structures for episodic memory. Kolodner (1985) proposed that events are organized by distinctive properties resulting in memories about general classes of events. At a second level are memories about event details. Conway (1996) proposed a three-layered model consisting of lifetime periods, general events, and event-specific knowledge.

All three of these theories may be correct to some extent, for different types of events. In general, it appears that survey questions that provide retrieval cues that are consistent with memory encoding and retrieval strategies are more likely to result in responses that contain the desired information. Thus, structuring questions about past experience in terms of general

lifetime periods, then classes of events, then specific event details may result in better quality data. All too often, questions about past experience deal only with the specific event details.

Questions about the timing of event occurrence, as well as the frequency and duration of events prevent other problems for retrieval. Biases include *temporal compression*, in which events are recalled as occurring more recently than they actually occurred, and *telescoping*, which results from the respondent's uncertainty about the boundaries of the relevant time period. For example, if asked about purchases in the past 12 months, the respondent might recall events in the past 15 or even 18 months. This is known as *forward telescoping*, when events that occurred before the reference period are recalled. In *backward telescoping*, the respondent erroneously omits events that occurred in the reference period.

Current thinking about cognition (such as described by Page and Raymond, 2005) presents a model of the mind consisting of three core “modules”—knowledge, emotion, and action. Cognitive elements that get into the “mental workspace” have knowledge, emotion and action “tags.” An important aspect of this model is that cognitive elements have to compete to get into the workspace. This implies that retrieval of items from memory may be stochastic rather than deterministic.

Judgment

Once the survey respondent has retrieved some information that might be relevant to answering the question, the *judgment* step begins. Judgment involves processing, combining, averaging or otherwise summarizing the information that has been recalled. Phenomenologically we may not be aware of the difference between or the transition from retrieval to judgment, but we do need a process to move from retrieval of various memories to generation of an answer to the question. Consider this question from a home use test for a fast moving consumer good:

How often do you think you would purchase the product you tried?

Never
Less often than once every 6 months
Once every 4-6 months
Once every 2-3 months
Once per month
2-3 times per month
Once a week
More than once a week

Here's a possible scenario for one respondent.

- recall last time a product in the category was purchased
- recall the next to the last time the category was purchased
- estimate an average purchase frequency from these memories
- recall experience (i.e. satisfaction) with recent category purchases
- compare recalled experience with the new product experience
- determine if the new product is better than the previous products
- average across these comparisons

- estimate a substitution rate (in simple terms, such as “none”, “some” or “all”)
- apply this substitution rate to the estimate of category purchase frequency
- find the response in the above list that is closest to this answer.

This is a hypothetical simplification, and not meant to represent the respondent’s conscious thought processes. However, this scenario does provide an idea of the amount of mental work that can be required to generate a response to a single survey question. Judgment comes into play in the third step in this scenario, where the respondent extrapolates from a limited set of representations to a larger class of representations (in this case, from two purchases, to all purchases in a category). The respondent does not find the answer to the purchase frequency question in memory. After all, only consumers whose behavior is invariant would have a single representation of purchase frequency. Instead, the respondent finds some number of relevant representations and then processes those representations in some way to generate the answer to the question.

Tversky and Kahneman (1974) have made major contributions to our understanding of judgment under uncertainty. They identify three key heuristics that individuals employ when making these judgments. It seems likely that these same heuristics come into play when survey respondents attempt to arrive at an answer from the various elements they have retrieved from memory.

The first is a *representativeness* heuristic, which individuals appear to use when making judgments about probabilities. Consider a survey question that asks the respondent to indicate whether each of several statements “describes” a particular brand. For statements where the respondent is uncertain, the respondent might compare each statement to her “stereotype” for the category, or perhaps compare the similarity of that statement to other statements about which she is more certain.

The second heuristic is *availability* in which judgments about the frequency or probability of an event are a function of the relative ease of retrieving instances of the event from memory. For example, if consumers are asked which brands they have purchased in the past year, those brands that more readily come to mind are more likely to be cited. This can become a problem when other factors, such as the amount of advertising for a brand, have an effect on the ease of retrieval.

Tversky and Kahneman call the third heuristic “anchoring and adjustment.” Using this heuristic, respondents would arrive at a judgment by adjusting an initial value to arrive at a final estimate. The key here is that the final judgment is affected by the “location” of the initial value. For example, if consumers are asked to indicate their overall satisfaction with a series of service encounters, their final judgments might be skewed by some extreme experience. Asked how satisfied you are with a particular airline and you have had generally positive experience but on your most recent trip your luggage did not arrive at the same time as you, your overall evaluation may be shifted in the direction of this extreme anchor.

Other judgment heuristics may come into play as well. For example, comparative judgments might rely on a “points of difference” heuristic that considers the number of differences, or perhaps a “critical difference” heuristic that looks for differences on one or two key aspects of the items being compared.

Attitude judgments are of special interest to market researchers. Many marketers, for example, have attempted to segment customers based on attitudes towards a product category, attitudes towards brands, and more general attitudes. The most commonly used method for measuring attitudes is the Likert scale (Likert, 1932), which is usually implemented as a five or seven-point “agree-disagree” scale. Here are a few attitude statements from a survey on public transportation:

Buses give me the flexibility to travel when and where I want.
I feel safe when traveling by bus.
People like me ride the bus.

In order to respond using the categories on a Likert scale, most respondents will retrieve various memories of their bus-riding experiences. Additionally, for most respondents, these memories will be a *sample* of all the possible representations stored in memory. Tourangeau, Rips and Rasinski (2000) have proposed a *belief-sampling* model for attitude judgments. In this model, the retrieval step yields an assortment of material or *considerations*. These considerations vary in accessibility, and some will be more relevant to the question than others. Additionally, time constraints and motivation may impact the assortment of considerations. According to these authors, an *attitude* is seen as a “kind of database consisting of feelings, beliefs, and knowledge about an issue.” The most important aspect of this model for this discussion is the potential variability in the assortment of considerations for any one respondent. Tourangeau, *et al.* argue that this may be a factor driving instability in attitude measurements over time.

Response

The final step in the four-step model of the survey process involves locating the internally generated answer within the *response space* specified by the survey questionnaire. An open-end verbatim question represents an unrestricted response space (within the boundaries of the uncertainty space defined by the question). A closed-end question restricts the response space to those categories or values that are defined within the response set.

Because the response step is observable, and because many closed-end response sets lend themselves to quantitative analysis, there is considerable research on some aspects of the response process (see the paper by Chrzan and Malcolm in this volume for an example). However, much of this research is “atheoretical” in that it does not test hypotheses about the judgment process nor the way in which judgments are matched to responses.

It is likely that the four-step survey response process does not proceed in a strictly linear fashion. For most survey questions, the respondent will know the available response categories at almost the same time that representations of and about the question are formed. The response set helps the respondent to define the uncertainty space and to formulate retrieval and judgment strategies. For example, a list of usage occasions may help the respondent identify the general classes of events to search in memory. A Likert scale provides some information about the type of judgments that might be needed.

Closed-end responses have a number of advantages for the survey researcher and scales—which return ordinal or interval values—have proven especially appealing due to the mathematical operations that can be performed on the data. However, the convenience of such scales may lead to indiscriminant use of particular scales. One common problem is a mismatch

between the dimensionality of the scale and the dimensionality of the answer space. For example, *attitude* and *belief* (“Brand A is made by a reputable company”) statements may have an “agree-disagree” dimensionality, while *frequency* statements (“I often do X”) do not, yet agree-disagree scales are often paired with frequency or simple occurrence statements.

IMPROVING SURVEY QUALITY WITH THINK ALOUD PRE-TESTING

At this point, you might be tempted to throw up your hands in despair. After all, it is not easy to observe the comprehension, retrieval, judgment and response processes, and it may be simpler to discard the obviously problematic respondents and hope for the best with the remainder. Fortunately, *think aloud* pretesting offers a means of improving the quality of survey response through identification of potential problems in comprehension, retrieval, and judgment.

Think aloud pretesting, as described by Bolton and Bronkhurst (1996) and Willis (2004) is a form of *concurrent protocol analysis*. In concurrent protocol analysis, a respondent provides information about conscious cognitive activity as it happens. We might also call it *stream-of-consciousness* pre-testing. Concurrent protocol analysis is one of many techniques that have been developed over the years to aid in instrument development.

In a think aloud pretest, a small number of individuals representative of those who will be sampled are asked to verbalize their thoughts as they complete the survey in the presence of an interviewer. The interview is typically unstructured; the interviewer prompts the respondent to “think aloud” as needed.

Think aloud pretesting is especially suited for identifying problems of comprehension—presupposition and vagueness, for example, as well as problems in forming the representations of and about the question—and misalignment between the response space (defined in the survey) and the answer space (the result of the judgment process).

A number of alternatives to the think aloud method have been employed to pretest market research survey questionnaires. These include what might be called a “retrospective” think aloud in which respondents described how they arrived at their answers to specific questions after the fact, either following the question or after completing the survey. A variety of approaches using follow-up questions have been used, including “confidence” ratings (respondents indicate their level of confidence in their answers), respondent paraphrasing of the question, and specific follow-up probes.

Implementation Guidelines

Our experience is comprised of think-aloud pre-tests conducted face-to-face, usually at a central location interviewing facility. Think aloud pretests for telephone surveys can be conducted over the telephone. For web interviews, a combination of telephone (for audio) and internet (for observing the respondent’s answers) may be an alternative to in-person interviewing.

The number of interviews needed to surface problems in a survey will vary depending on the complexity of the survey and the number of targeted subgroups or sampling quota groups planned for the survey. We typically conduct between 10 and 12 pre-test interviews for surveys with one or two sample subgroups.

We recommend conducting the interviews over a two-day period. This enables making changes to the survey after the first day of interviews so that the changes can be tested on the second day.

A few other recommendations:

- Respondents should be recruited using the criteria that will be used to screen respondents for the full survey.
- The think aloud pretest should be conducted using the mode of data collection that will be used in the full survey (i.e., an internet survey should be tested using a web or CAPI survey; a telephone survey should be pretested by telephone).
- The in-person think aloud will take longer than the actual survey; we generally schedule one-hour sessions to pre-test questionnaires that will take 20-25 minutes to complete in the full survey.
- For multilingual studies, we recommend conducting a think aloud pretest with individuals from each language group.

Preparation is key to maximizing the utility of think aloud pretesting. A think aloud pretest is an excellent opportunity to test alternative ways of asking key or potentially problematic questions before the study is fielded. We recommend planning probes around potential problem areas, which can be identified in advance using the four step response model as a guide. For example, if you think there might be a comprehension issue, you can probe for the meanings respondents assign to terms and concepts.

A FEW THINGS LEARNED FROM THINK ALOUD PRETESTS

Think aloud pretests are conducted to improve survey quality for a specific instrument. However, over the course of many think aloud pretests, we have observed some commonalities in *online* questionnaires that can be helpful in designing future studies.

For example, we've observed that respondents often read the responses before they read the questions. Respondents may infer the uncertainty space of the question from the available responses. This suggests that we should pay special attention to constructing the response sets for closed ended questions.

When respondents do read the questions, they often need to read the questions more than once. This is especially true for questions that are higher in cognitive complexity. Their think aloud verbalizations indicate that respondents may have trouble determining which aspects of a complex question are relevant to forming their answers.

We've also observed that respondents expect navigation features in online surveys to be the same as those they encounter on other websites (for example, respondents often expect to be able to access additional information simply by scrolling over an item).

Choice-based conjoint has become an increasingly popular technique for measuring buyer preferences over the last several years. Our think aloud pretesting of surveys that include choice-based conjoint exercises has revealed some things that can be used to improve the quality of responses to choice-based conjoint tasks. We have observed, for example, that without specific cues or instructions, respondents may not notice variation from one choice task to the next in an

online survey. Learning effects that have been inferred from other empirical analyses are apparent in think aloud pretests of conjoint exercises; respondents may complete three or more tasks before they realize they are being forced to make trade-offs. We have also observed behavior suggesting that respondents often use non-compensatory strategies in making their choices. The main question addressed via think aloud pretesting is whether such behavior is a true reflection of a respondent's decision process or an artifact of the way the task is structured and presented.

Think aloud pretesting can be invaluable when conducting research in different languages in different geographic regions. Translation of surveys can range from "acceptable" to "bad." Conducting a think aloud pretest with a native language interviewer and simultaneous translation for observers can reveal problems with the translation. Particularly in emerging markets, a think aloud pretest provides a reality check on the concepts included in the survey (problems of presupposition and unfamiliarity). For example, in a study concerning high definition television, the questionnaire asked respondents to indicate which of a list of television related products or services they owned or subscribed to. Chinese consumers had no understanding of "satellite TV" since no satellite services were offered in China (at least at that time).

FINAL THOUGHTS

Survey questionnaire design will always require some combination of empirical understanding of what works and what does not plus intuition and craftsmanship. One obstacle to improving the quality of survey questions lies in the different perspectives of question writer and respondent. The question writer typically "samples" an uncertainty space and then tries to write a question that encompasses those sampled points. The respondent, on the other hand, sees only the question and has to infer the writer's notion of the uncertainty space from the question. Without an opportunity for dialog between the writer and the respondent, there is no opportunity to clarify misunderstanding or incorrect inferences about the uncertainty space.

The four-step model of the survey response process proposed by Tourangeau *et al.* is a useful starting point for improving survey design. Each potential question can be evaluated against each step to determine if there might be problems of comprehension, problems in retrieval, problems at the judgment step, or problems in matching internal answers to survey response categories. Think aloud pretesting has proved to be an effective tool in identifying these problems.

REFERENCES

- Bolton, R.N., & Bronkhorst, T. M. (1996). Questionnaire pretesting: Computer-assisted coding of concurrent protocols. In N. Schwartz and S. Sudman, *Answering questions: Methodology for cognitive and communication process in survey research* (pp 37-64). Jossey-Bass.
- Conway, M.A., (1996). Autobiographical knowledge and autobiographical memories. In D.C. Rubin (Ed.), *Remembering our past* (pp 67-93). Cambridge, England: Cambridge University Press.
- Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87, 215-257.
- Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Kolodner, J. (1985). Memory for experience. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 1-57). Orlando, FL: Academic Press.
- Krosnick, J.A., & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 209-219.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.
- Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

EVALUATING FINANCIAL DEALS USING A HOLISTIC DECISION MODELING APPROACH

PAUL VENDITTI¹, DONALD PETERSON², MATTHEW SIEGEL³
GENERAL ELECTRIC

ABSTRACT

At GE Global Research we have developed a holistic⁴ decision modeling approach to evaluate deal approval likelihood for structured finance products. To develop this approach we (1) interviewed subject experts to determine the most important attributes in financial deals, (2) for each attribute ensured quality data using a Six-Sigma approach, (3) developed the expert-based model using adaptive conjoint, and (4) validated the model using empirical data. When we tested the model results using actual data from deals and compared the results to several years of historical deal data, we found the holistic decision modeling approach to be 87% accurate. A decision support tool was developed from this approach and is currently being used by the GE Energy Finance business. The purpose of the tool is to increase the number of successful deals with faster customer response times.

EXPERT-BASED DECISION SUPPORT SYSTEM (DSS) BACKGROUND

There is ample research dating back to the 80's and 90's regarding expert-based decision support systems⁵. Green, Johnson and Neal [2] cite many examples pioneered by John Little and his colleagues at MIT. GE Research has also fielded numerous expert-based decision support applications working with many GE businesses. One recent example was the automation of long-term care and life insurance applications at Genworth Financial [3]. A common set of characteristics for expert-based DSS includes:

- High-stake decisions
- Data availability issues
- Reasonably-well structured problems (e.g. insurance underwriting)
- Amenable to specialized software development

A stated novelty is the model builder's use of subjective estimates provided by expert decision makers, given the absence of hard data. Quaile *et al.*, [4] developed several expert-based

¹ Paul Venditti, Decision Science Researcher, Risk and Value Management Laboratory, GE Global Research Center, Niskayuna NY

² Donald Peterson, Senior Vice President, Strategic Marketing for GE Energy Financial Services, Stamford CT

³ Matthew Siegel, Managing Director and Chief Marketing Officer for GE Energy Financial Services, Stamford CT

⁴ Our definition for this term was derived in part from Simon and Fletcher [1]. We have used this term to refer to the comprehensive analysis that (a) attempts to integrate qualitative (art) and quantitative (science) attributes, (b) makes executives' prior knowledge and intuition explicit, (c) embodies analytical techniques used to insure data quality, (d) provides for empirical validation, (e) attempts to link all qualitative and quantitative evidence into a coordinated "story" that helps the users make more informed, evidence-based decisions.

⁵ A coordinated collection of data, system tools, and techniques with supporting software and hardware by which an organization gathers and interprets relevant information from business and the environment and turns it into a basis for making management decisions. 2. (models definition) A system, usually based on a model and computer software package, that describes the implications of specific decisions and/or recommends specific actions, using a set of input information. This information may either reside permanently in the DSS or be input for the particular scenario of interest (or both). The information can consist of primary information. An important aspect of many decision support systems is the facilitation of "what if" analyses; i.e., the sensitivity of optimal strategy to the assumptions in the input information.

DSS applications with few experts in support of internal decision-making actions (e.g. credit risk assignment). Quaile’s approach like others cited was designed for predictive accuracy in situations where there is quantitative and qualitative attributes and possibility for missing data.

In this paper, a holistic method for developing an expert-based DSS is proposed that seeks to balance the needs for a simple, realistic assessment of business transactions while also having a firm basis onto which mathematical models and DSS are developed.

This paper is unique because it (1) provides a practical approach for studying and addressing data quality issues (missing data and attribute ambiguity) prior to eliciting expert responses, (2) provides a modified approach to traditional adaptive conjoint based on issues reported in past research [5], and (3) provides examples of novel visualization techniques that can be utilized to further support cognitive decision-making processes and long term DSS viability.

Emphasis is placed on decisions with the following characteristics:

- Complex internal decision (e.g. large number of attributes, mix of quantitative and qualitative considerations).
- Several attribute choices difficult to capture in traditional relational databases (e.g. commodity risk).
- Data to explain variation of decision-outcomes are sparse, and subject to quality issues (missing data and data that are sometimes difficult to decipher).
- Reliance on experts with significant prior knowledge, experience and intuitive skills.
- Suitable to small or large number of experts.
- Homogeneity across experts more prevalent than heterogeneity.

HOLISTIC DECISION MODELING APPROACH

The below figure illustrates the four main activities for developing and validating the expert-based model discussed in this paper.

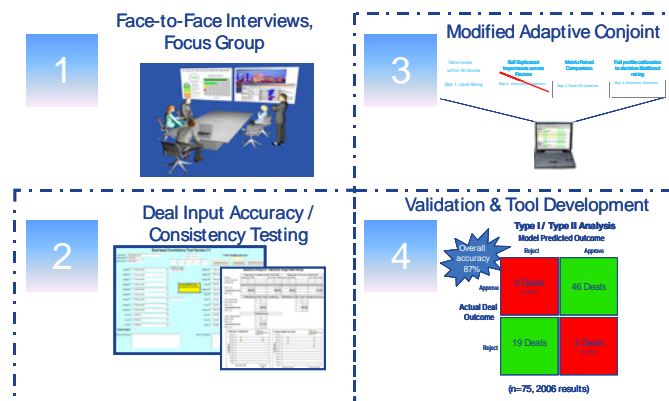


Figure 1:
Holistic Decision Modeling Process for Evaluating Financial Transactions

INTERVIEW STAGE:

Past research on the psychology of decision-making [6] shows that experts are superior in two out of three of the key decision-making activities. These activities are summarized in table 1 below.

Table 1:
Stages of Decision Making

Decision Activity	Superior Approach
- Identify relevant attributes (e.g. Return on Equity)	Expert Intuition
- Identify values for attribute levels (e.g. 25% to 15%)	Expert Intuition
- Integrate the individual attributes into overall evaluation (E.g. figure out relative importance of return on equity relative to other attributes like credit strength or commodity risk)	Model Derived

In the first stage of our approach, we extracted attributes from experts using (1) detailed qualitative interviews, (2) focus group sessions, and (3) observation of numerous deal approval discussions.

The face-to-face tasks and focus group sessions resulted in the identification of a large number of relevant attributes and appropriate values for each attribute.

To ensure understanding of attributes-in-context of financial deals, the model development team observed financial deal discussions between executive decision-makers and deal originators/underwriters. This took place over a several month time period.

At the conclusion of this stage, we identified and defined 25 attributes and associated levels (taking advantage of expert strengths as past research supports).

DATA INPUT ACCURACY / CONSISTENCY STAGE:

Our objectives for the second stage of the approach were to measure the accuracy and consistency of deal team member's ability to classify financial transactions using the 25 attributes and levels defined earlier.

This experiment was conducted on a total of 49 financial deals over a four-month time period.

The experiment consisted of the following steps.

Step 1 - Prior to financial deal discussions with executive decision-makers:

The deal team would post documents in a central repository the day prior to deal discussions with executive committee. This document would serve as the single-source for extracting attribute level information.

The data input experiment consisted of a senior marketing member and deal team member. Each of these individuals would **independently** use a customized excel tool (Figure 2) to input the 25 attribute levels for each financial deal to be reviewed with executive decision-makers. The data input tool included a help function that provided definitions for each of the attributes. The

drop down box for each attribute listed all available attribute levels (including provisions for data quality issues... e.g. unknown).

Figure 2:
Data Input Tool (Actual data not shown due to proprietary reasons)

Step 2: Determination of “gold standard” for each financial deal:

After financial deals were reviewed with executive-decision makers, a committee of experts (Risk & Marketing) along with individuals who provided inputs into the Excel tool would convene to determine the correct selections for each of the 25 attribute levels (per deal). This would be defined as the “gold standard” for that potential transaction. The agreements reached during these sessions were by consensus.

Step 3 – Utilize six sigma methodology to measure/analyze/improve and establish control procedures:

This paper will not attempt to serve as a full review of six-sigma methodology since this task is beyond scope of this paper. Suffice to say that the six-sigma process steps of define, measure, analyze, improve and control provided the approach for meeting data input quality objectives.

The philosophy behind Six Sigma is that if you measure how many defects are in a process, you can figure out how to identify root cause issues. You can then implement improvements to eliminate the defects and get as close to perfection as possible. In order to achieve Six Sigma, a process cannot produce more than 3.4 defects per million opportunities, where an opportunity is defined as a chance for nonconformance.

For the measure portion, we defined our opportunities and defects as follows:

Opportunity: Each attribute level for a financial deal (25 per deal per input person)

Defect: Each selected attribute level that was not in agreement with gold standard selections (consensus based decision arrived during step 2).

Total opportunities for study: = $t * n * d$, where t = number of input persons, n = number of attributes input per deal and d = number of deals included in study. For our study, this translated to: $2 * 25 * 49 = 2450$ total opportunities.

During the analysis phase, data input defects were investigated using analysis tools and techniques such as Pareto charts, chi-square tests and fishbone diagrams (Figure 3).

Improvement and control plan strategies were then developed from working sessions with the development team. Many of the improvements were implemented iteratively by periodically updating the data input tool (a total of 11 revisions were made over a 4 month time period).

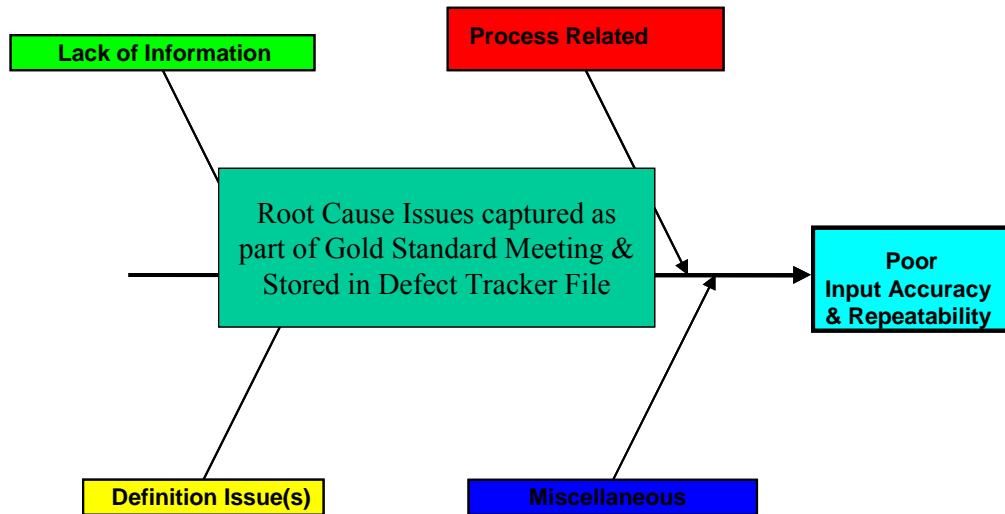


Figure 3:
Illustration of Fishbone Diagram used to analyze data input defects
(Actual data not shown due to proprietary reasons)

MODIFIED ADAPTIVE CONJOINT STAGE

The first two stages of this approach provided us a firm understanding of attributes-in-context of financial deal discussions.

For this stage of our approach, we provide discussion and comments of pertinent aspects that shaped rationale for the conjoint design we utilized.

The two main challenges we faced with our non-conventional application of conjoint were the need to handle a large number of attributes and provide reliable part-worth estimates at an individual level (tailored to 3 executive decision-makers).

Past research [2] suggests two primary approaches for handling the problem of large number of attributes and reliable individual-level part-worth utilities: (1) the self-explicated approach, (2) hybrid conjoint analysis (ACA is considered a hybrid approach because it contains self-explicated and decompositional⁶ tasks).

⁶ Conjoint analysis is often referred to as decompositional method of preference estimation, because rather than directly ask respondents to indicate their preferences for attributes and levels, these are statistically deduced (decomposed) from the overall product evaluations of conjoint profiles (tasks).

Self-explicated approach (compositional approach)

In the self-explicated approach, the respondent first evaluates levels of each attribute on a desirability scale where the most desired level receives the higher score and the less desirable level receives a lower score (e.g. on a scale of 0-7). The respondent is then asked to evaluate the relative importance of attributes by a task resembling a chip allocation task (e.g. allocating 1000 points across attributes based on relative importance). Part-worth utilities are then obtained by multiplying the importance weights with the attribute desirability ratings.

The primary advantage for this approach according to Green is:

- Ability to handle large number of attributes
- Less cognitive strain and fatigue on respondent
- Speed, simplicity and lower cost for data collection

The primary disadvantages (Green, Srinivasan, 1990) include:

- Potential for double counting (e.g. a paired comparison task incorporating orthogonal & efficient pairings would enable respondent to realize redundancy among attributes and respond accordingly to overall preference rating task)
- Over-simplification of desirable ratings for quantitative attributes (e.g. responses for quantitative attributes would become linear where intermediate levels would be selected in middle of rating scale)
- Respondent does not evaluate any full profile concepts

Traditional ACA Approach (Sawtooth Software)

In the default ACA approach using Sawtooth Software, the respondent (1) rates levels within attributes (unless best to worst ranking can be assumed), (2) evaluates the relative importance of attributes one at a time by comparing the range of levels for that attribute, (3) evaluates metric paired comparison questions and (4) evaluates a set of several full concept profiles. Part-worth utilities are then calculated using OLS techniques that combine utilities derived thru the priors section (Steps 1 and 2) with pairs section (Step 3) and calibration based on purchase intention (Step 4).

The primary advantage for this approach is:

- Ability to handle large number of attributes
- Greater realism among paired choices
- Greater chance of detecting real importance weights
- Less likelihood for double-counting

The primary disadvantages include:

- Length of survey time, respondent fatigue
- Issues with default importance section reported in previous research

Self-explicated approach versus traditional ACA approach:

In theory, it should be difficult for respondents to provide judgments, but empirical evidence suggests that they are quite accurate. A comparison study performed by Sattler and Hensel-Borner [7] of 23 studies using both self-explicated and various forms of conjoint produced mixed

results with a majority of studies 57% showing no statistical difference between approaches. Out of the 23 studies, only 22% (5 out of 23) show significantly better results (in terms of reliability or predictive validity) for conjoint measurement compared to self-explicated approaches ($p < 0.10$ level). While these results were of interest, none of these studies dealt with a comparable number of attributes (these 23 studies dealt with less than 10 attributes, and most concepts considered were straightforward consumer purchases).

Additional research suggests self-explicated methods have proven powerful when used in conjunction with decompositional methods. Green used self-explicated methods effectively in hybrid conjoint analysis – a method in which respondents’ self-explicated partworths modify overall partworths that are estimated. ACA is considered a hybrid approach in which self-explicated and metric partial profile data are combined to enhance accuracy.

We considered the pros and cons of both techniques and proceeded with selection of ACA design with the caveat that we would attempt to address some of the known ACA shortcomings and where possible incorporate the simplicity of the self-explicated approach.

Issues with traditional ACA approach:

According to Orme, the default stated Importances section in ACA has received some attention in recent Sawtooth Software conferences, and potential weaknesses have been noted:

- Respondents often tend to say that most everything is important (scale use bias), which tends to "flatten" final derived importances and can reduce predictive validity.
- Respondents don't use the importance scale in a "ratio" sense, even though OLS estimation within ACA assumes (for example) that a "6" on the scale is twice as important as a "3".
- Respondents may rate an attribute they see early on with the top scale point before realizing that another more important attribute is still to come.

Modified ACA design

As stated earlier, our design philosophy was to adopt the best aspects of ACA and self-explicated techniques. Furthermore, we wanted to incorporate knowledge gained from the interview and data input consistency stages of the methodology.

To accomplish our objective, we designed the modified ACA interview as follows:

1. Defined 22 attributes and levels (results from data input consistency necessitated in a reduction of attributes from 25 to 22). Re-confirmed with executives’ unacceptable ranges by attribute. These values were excluded from survey.
2. Incorporated prohibition rules (important aspect of realism). There were a total of 10 prohibition rules implemented across attribute levels. Attribute level ranking questions were asked for 3 out of the 22 attributes, all other attributes had a best to worst a priori ranking specified.

- Redesigned ACA importance question (via customization performed by Sawtooth Software) that allowed for a constant sum task allocation to be input in ACA survey as free format question⁷. Figure 4 shows an Excel input worksheet used by executives. These Excel files were submitted to the model development team prior to executives taking the web survey.

In this table, we've arranged numerous deal factors into several categories for your evaluation. The placement of these like factors into defined categories is meant to help you better evaluate factors relative to one another (within same & different categories).

For each detailed factor shown, please allocate points to the deal factors that you believe are most important to deal approvability (the total points assigned must equal 1000). The grand total is shown at the very bottom of the table. Please consider the full range (As good & as bad as) when considering a deal factor. If you feel one factor is twice as important as another, please assign it twice as many points. If the deal factor has no importance to you, then assign a zero value. You will be required to provide a value for each factor.

Deal Economic Factors	As Good As...	As Bad As...	Your Allocation of Points
Profitability 1	Attractive Description	Unattractive Description	96
Profitability 2	Attractive Description	Unattractive Description	57
Profitability 3	Attractive Description	Unattractive Description	41
Deal Structure Factors	As Good As...	As Bad As...	Your Allocation of Points
Structure #1	Attractive Description	Unattractive Description	65
Structure #2	Attractive Description	Unattractive Description	42
Structure #3	Attractive Description	Unattractive Description	31
Structure #4	Attractive Description	Unattractive Description	29
Structure #5	Attractive Description	Unattractive Description	10
Deal Underwriting Factors	As Good As...	As Bad As...	Your Allocation of Points
Underwriting 1	Attractive Description	Unattractive Description	60
Underwriting 2	Attractive Description	Unattractive Description	200
Underwriting 3	Attractive Description	Unattractive Description	47
Underwriting 4	Attractive Description	Unattractive Description	6
Underwriting 5	Attractive Description	Unattractive Description	4
Deal Asset Factors	As Good As...	As Bad As...	Your Allocation of Points
Asset 1	Attractive Description	Unattractive Description	137
Asset 2	Attractive Description	Unattractive Description	41
EFS Deal Factors	As Good As...	As Bad As...	Your Allocation of Points
Other 1	Attractive Description	Unattractive Description	56
Other 2	Attractive Description	Unattractive Description	29
Other 3	Attractive Description	Unattractive Description	1
Other 4	Attractive Description	Unattractive Description	25
Other 5	Attractive Description	Unattractive Description	18
Other 6	Attractive Description	Unattractive Description	2
Other 7	Attractive Description	Unattractive Description	3
TOTAL (Must Equal 1000)			1000

Figure 4:
Illustration of Constant Sum Excel Worksheet Task used in place of traditional ACA importance tasks (actual data not shown due to proprietary reasons)

- Executives taking the Modified ACA survey were required to re-submit their Excel worksheet inputs into free format question that was customized for this study by Sawtooth Software (required as part of utility weights for priors section)⁸.
- Adjusted ACA setting to only carry a maximum of 13 out of the 22 total attributes into pairs section (reduced number of attribute levels from 66 to 39 resulted in an increase of pairs contribution to final utilities by 43%)⁹. By reducing the number of levels, we were able to increase the emphasis on refining attribute utilities that were most important to decision-makers.

⁷ During pre-testing, executives indicated a preference to have this task 1st be given to them via an Excel file.

⁸ Sawtooth Software has developed new functionality that allows one to specify within ACA settings to “Skip default importance question” and also allow “setting prior importances based on other questions”. Version 6.2 also allows user data to be directly submitted into survey (based on user password). These new features and capabilities would further simplify tasks we had performed as part of this study (e.g. executives would not have had to re-enter their off-line constant sum task values).

⁹ Priors contribution (PC) = $n/(n+t)$ where n = total number of levels used in pairs section and t = number of pairs questions answered by respondent. For our study, Priors Contribution = $39/59 = .66$. Pairs contribution = $t/(n+t) = 20/59 = .33$. If all 66 levels would have been carried into pairs section then Priors contribution = $.77$; Pairs contribution = $.23$. Hence by reducing the number of levels from 66 to 39, the pairs contribution increased by 43%.

6. Based on pre-testing results of user fatigue, we specified 20 pairs questions with 3 attributes in each task. For each of the 20 pairs questions, we used the question: “All other attributes of the deal *being the same*, which financing transaction do you prefer?” A nine-point scale was used varying from 1 = strongly prefer deal on left, to 9 = strongly prefer deal on right.
7. Four calibration questions were asked in order to calibrate final utilities onto a more meaningful scale (Deal Approval Likelihood). For the calibration questions, we used the question: “What is likelihood you would approve this deal? Please type a number between 0 and 100 for each deal where 0 means, “definitely would not approve” and 100 means “definitely would approve.” Partworths were obtained based on ordinary least squares.
8. At the end of the survey, executive decision-makers were asked to answer four questions about the conjoint tasks. The question posed was “How much do you agree or disagree that this questionnaire was...” realistic, not tedious, not confusing, enjoyable. The scale range was 1 (strongly disagree) to 5 (strongly agree).

VALIDATION STAGE:

How can we be sure that the conjoint model we have constructed accurately represents the realities of the executive committee preferences and actual decision-outcomes? There are two primary tests we used to validate our model:

We were predominately focused on the empirical accuracy of the model to future financial deals (e.g. how well does the deal approval likelihood score predict actual deal approval).

While we were also concerned with expert coherence and performed several tests and analysis to gauge internal respondent consistency, this paper will not review those tasks since our focus here is on the applied managerial summary aspects most important to business decision makers.

Confusion matrix:

The primary method for assessing the empirical accuracy of our model was done by comparing the accuracy of the predicted decision outcome. The confusion matrix would display each predicted decision outcome¹⁰ in relation to the actual decision outcome¹¹. Accurate predictions were plotted in the diagonal going from the top left to the bottom right of the table. Values in cells below or above this diagonal represented classification errors (alpha and beta errors). Figure 5 illustrates a confusion matrix.

The error rates depend on the quality of the model and on the cut-off point used to classify the financial deals. One problem is that increasing the cut-off point in order to reduce the number of false negatives will usually generate an increase in false positives. For our study, we derived 4 classification categories¹² taking into account attribute level probabilities based on a historical deal set sample along with sensitivity analysis of the deal approval likelihood model. We

¹⁰ Predicted decision outcome here means that the deal approval likelihood score for a future financial deal would suggest an outcome such as “Approve Deal” or “Reject Deal”

¹¹ Actual decision outcome here means what actually happened with financial deal (e.g. Deal was approved or deal was rejected)

¹² Four classification categories were broken down into 2 approval classes (90-100 and 75-89) and 2 rejection classes (55-74 and less than 55)

accomplished this by performing a Monte Carlo analysis that is detailed in appendix A. This was not intended to be an exact science but rather a decision support aide to establish reasonable thresholds for future financial deals.

Seventy-five financial deals were used for in-sample validation. These deals were randomly selected during normal the course of business in 2006. The attribute levels for these deals were validated by a committee of experts (e.g. adhered to “gold standard” process).

Subsequently, an external-sample of 41 deals was evaluated in August of 2007¹³. When we re-tested the predictive accuracy of our model, we performed no adjustments to our previously validated model.

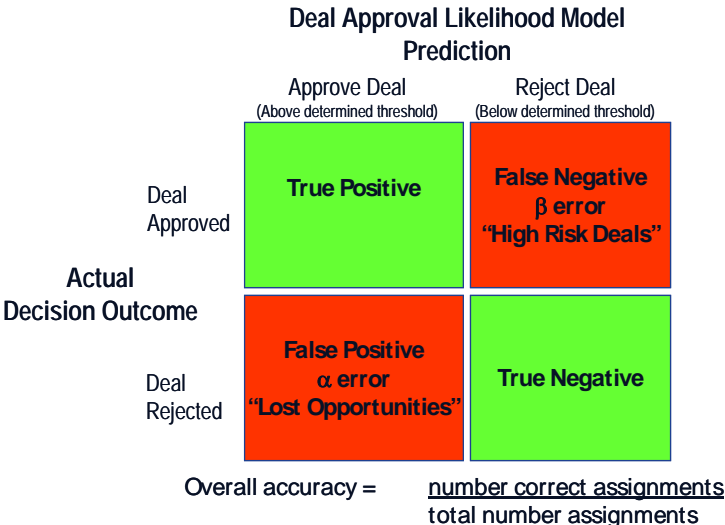


Figure 5:
Illustration of Confusion Matrix

Blink scores:

The “Blink” score exercise was comprised of comparing the “deal approval likelihood score” from the tool and from the 3 credit committee members taken immediately after each stage of a transaction review (preliminary, proposal, approval). “Blink” scores were taken during a six-month period in 2006 at the twice-weekly transaction review meetings. Figure 6 illustrates the scorecard used for capturing each executive “Blink score”.

¹³ This external sample did not adhere to “gold standard” process as that experiment and study had concluded a year earlier. The production process used in 2007 had all improvements implemented from data input consistency work.

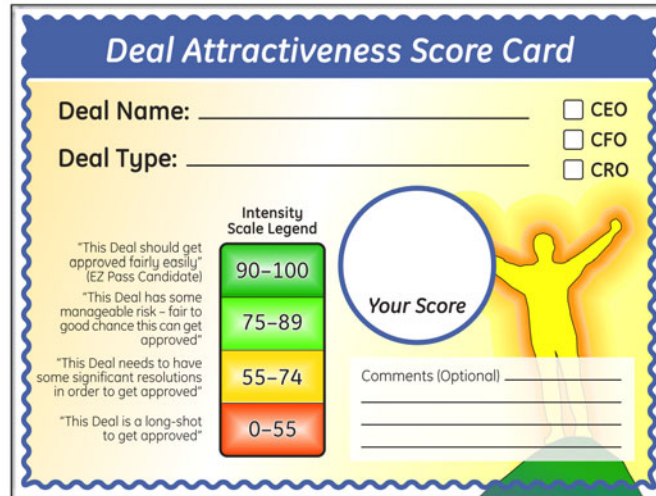


Figure 6:
Illustration of Blink Scorecard

RESULTS

Confusion matrix:

We show the results of the in-sample validation task for the 75 deals in table 2. We note that the accuracy decreases as the score declines. In reviewing errors on the two extremes (30% of total classification errors), it was discovered that reasoning used for decision outcomes on those transactions was of strategic nature and not typical.

It is not surprising that the majority of classification errors occur as the score nears 75 (cut-off used between approval and rejection). We had arbitrarily attempted to use different thresholds but were not capable of gaining additional accuracy.

An external sample (n=41) of financial deals in 2007 has yielded predictive accuracy of 85% indicating on-going performance of model.

Table 2:
In sample of decision outcome predictive accuracy

Predicted Class	Correct Classifications		Incorrect Classifications		Accuracy
	True Positive	True Negative	False Positive	False Negative	
90-100	20	N/A	1	N/A	95%
75-90	26	N/A	4	N/A	87%
55-74	N/A	15	N/A	3	83%
0-54	N/A	4	N/A	2	67%
TOTAL	46	19	5	5	87%

(90-100 and 75-89 are thresholds determined by development team to represent thresholds for “Deal Approval” Prediction. 55-74 and 0-54 are thresholds determined by development team to represent thresholds for “Deal Reject” Prediction. N/A means not applicable. False Positive indicated a model prediction of approval when in actuality deal was rejected. False Negative indicated a model prediction of rejection when in actuality deal was approved.

Blink scores:

Prediction of deal approval was compared between the holistic decision modeling approach and an intuition based score provided by decision makers. After one preliminary deal review, the intuition-based approach was found to be 65% predictive to eventual outcome. The holistic decision modeling approach was found to be 83% predictive. As deals progressed through numerous reviews, the accuracy results converged to nearly identical. We have inferred from this experiment that visualization and concise communication of decision-makers’ important attributes are essential during early stages of deal reviews. Past research [6, 9] supports our findings on models being able to outperform experts’ intuition when it comes to accurately integrating attributes into an overall evaluation. We note that this is especially important earlier in a decision-process.

Table 3:
Blink score comparison to model predictive accuracy

	Accuracy of model based score	Accuracy of intuition based scores (average of 3 executives)
Early stage investment meeting	83% (n=41)	65% (n=21)
Proposal stage meeting	82% (n=16)	70% (n=9)
Final stage approval meeting	81% (n=52)	80% (n=40)

ADDITIONAL RESULTS

Table 4 shows results of predictive accuracy broken down by individual executive for self-explicated, Modified ACA and the final validated model. We note the average of the executive decision-making committee accuracy compared to individual-level accuracy. This suggests the value of consensus-based decision-making [10].

Table 4:
Individual level predictive accuracy

	In-sample Outcome Accuracy Rates (n=75 from 2006)		
	Self-explicated Model (Rating + Constant Sum)	Modified ACA (Priors + Pairs + Calibration)	Final Model + Non-compensatory adjustment
Executive # 1	72%	77%	83%
Executive #2	69%	71%	80%
Executive #3	75%	77%	81%
Committee Average	76%	81%	87%

Data input consistency results:

Discrete choice experiments used to measure data input quality identified 644 defects out of a total of 2,450 opportunities. Root cause analysis enabled us to categorize 66 different types of user input errors.

Improvements were identified and implemented to address issues impacting user input consistency. These improvements included:

1. Revisions to attribute and attribute level definitions (updates to 11 of the attributes addressed a majority of the defects).
2. Consolidation of several attributes (study resulted in reduction of attributes from 25 to 22).
3. Addition of prohibition logic to prevent illogical combinations (10 prohibition rules created).
4. Establishing predefined choice selection for attributes deal team members could not be expected to reliably access (e.g. legal assessment of a particular country)

A new sample of 10 deals was re-tested with new process improvements. There were zero defects for the 440 total opportunities¹⁴. Quantifying the impact of data quality on model predictive accuracy was not in scope of this project, but due to the significance we found, we do plan on doing more detailed analysis in this area.

These improvements were also incorporated into DSS in manner that mistakes would be eliminated or at least minimized.

Finally, we show executive decision-maker feedback on survey tasks. These results are shown in table 5. We have determined that while efforts to make the survey realistic have been somewhat successful; we still have to further consider ways to reduce cognitive strain and the burden of survey tasks.

Table 5
Executive perception of conjoint tasks

Task was...	Executive # 1	Executive # 2	Executive # 3	Average
Realistic	3	4	4	3.7 (~agree)
Not Tedious	2	3	2	2.3 (~disagree)
Not Confusing	4	4	4	4 - Agree
Enjoyable	3	2	4	3 - Neutral

(Responses were on 5-point scale (1 being strongly disagree, 3 being neither agree nor disagree and 5 being strongly agree))

DISCUSSION - DECISION SUPPORT SYSTEM USABILITY:

Perhaps the most intriguing aspects of our approach are the Deal Heatmap (figure 7) and Tornado Graph (figure 8). A “Heatmap” is a one-page visual representation of deal attribute levels. The attribute levels are represented by different colors that correspond to executive

¹⁴ This is based on 10 deals, 2 respondents and 22 attributes per deal.

preferences derived from conjoint tasks. When used in DSS, the heat map aides the decision-making process by combining the transaction evidence into a coordinated “story” that is more explainable and transparent.

The Tornado Graph is an extension of the Heat Map that provides a greater degree of model transparency in that the factors are dynamically ordered based magnitude of part-worth utility weights. At the two extremes of the tornado graph are the factors that are most positively and negatively impacting decision-outcome. Another feature of the Tornado Graph is the ability to adjust attribute level settings for any of the attributes (e.g. realtime what-if sensitivity analysis) and receive immediate feedback as to that adjustment’s impact on deal approval likelihood score (based on the purchase likelihood model).

During our validation work, we encountered a few examples of non-compensatory decision processes. For example, a deal in a particular country was considered “too risky” based on one or two important attributes but more than adequate on all other measures. The sixth factor in figure 7 would be an example of this (indicated by red color).

Thru empirical testing, we were able to discover common attributes that were at lowest levels when decisions were non-compensatory. Since these examples were the exception rather than the rule, we decided to provide a visual indication for the possibility of a non-compensatory decision process. This was accomplished by adding Red Flag indicators to the side of the Heat Map. If any of these red flag items were red, then there would be a corresponding attribute(s) that would be red as well. This would require immediate attention and discussion.

It is important to note that a presence of a red flag on a heat map doesn’t necessarily mean the deal will be rejected. What it does suggest however, is emphasis on those considerations early in the dialogue regarding deals. If those considerations are truly not a deal killer (non-compensatory), then the heat map in total along with the deal approval likelihood score are prescriptive to decision outcome as intended.

By reviewing the heat map in this manner, we were able to generally avoid calibrations and model manipulations to accommodate the occasional non-compensatory decision-making processes.

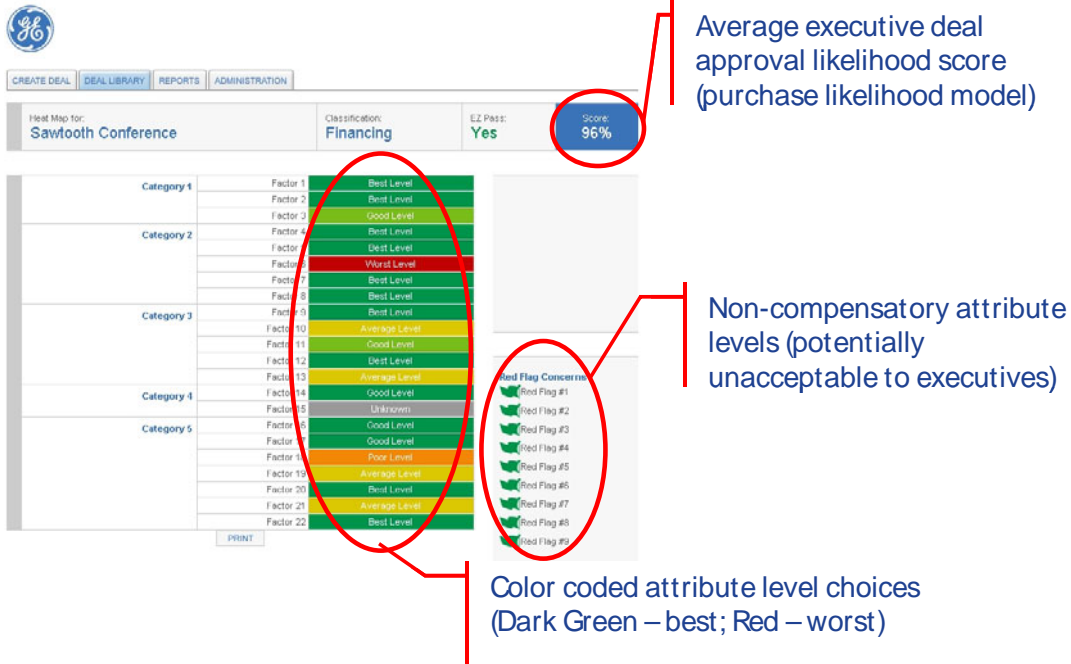


Figure 7:
Heat Map illustration (Actual data not shown due to proprietary reasons)

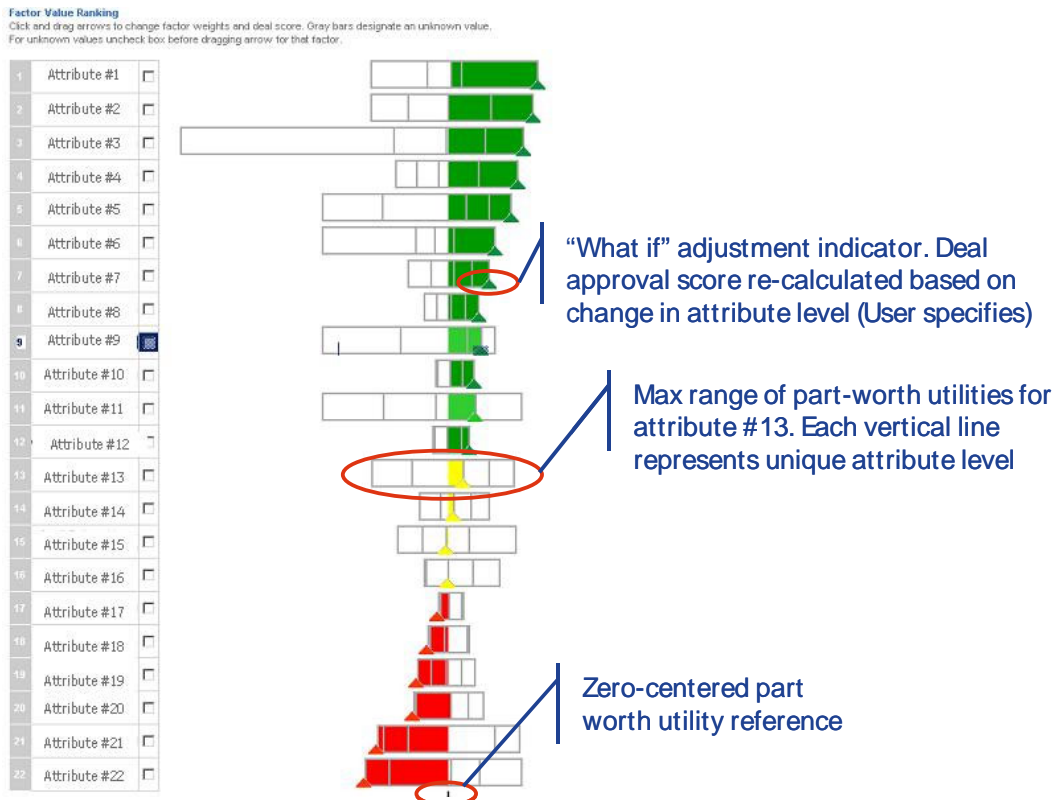


Figure 8:
Tornado Graph (Actual data not shown due to proprietary reasons)

CONCLUSION

We have presented a holistic decision modeling method that successfully demonstrated the feasibility of predicting deal approval outcomes for a relatively complex decision process with few decision-makers.

The results have shown that individual-level weights can be predictive to decision-outcomes as long as the analyst is diligent with respect to:

1. Handling of non-compensatory levels (e.g. unacceptables or levels that create substantial non-linearity to decision outcome).
2. Being explicit as to the meaning of attributes and levels and carefully using prohibitions when presenting concepts for trade-off (properly balancing realism versus choice independence).
3. Thoroughly testing for data input consistency. This is vital if the decision-support system requires input from inexperienced users.
4. Validating and fine-tuning as necessary based on many historical and current transactions. For example, adjusting the return level thresholds as market conditions change.

Visualization techniques greatly aided in demystifying the “black box” associated with a model predicted score. Visualization techniques can lead to extensions of this work wherein real-time feedback can be captured from decision-makers. Such an approach could facilitate a consensus-based feedback process, thus increasing the likelihood of the DSS long-term viability.

A second application of holistic decision modeling approach has recently been completed with similar results.

We intend to do more detailed analysis of results with particular emphasis on:

- Performing posterior decomposition of accumulated deal data and decision outcomes
- Exploring techniques aimed at maintaining model predictive accuracy over time
- Exploring conjoint design options that further increase realism while reducing cognitive strain

APPENDIX A

MONTE CARLO SIMULATION APPROACH USED FOR DETERMINING DECISION-OUTCOME THRESHOLDS

In the body of this paper, we showed how the deal approval likelihood score mapped to prediction of approval or rejection. In case the reader is interested in more details concerning the Monte Carlo¹⁵ simulation method used to derive decision outcome thresholds, we provide the following procedure & discussion comments:

MOTIVATION

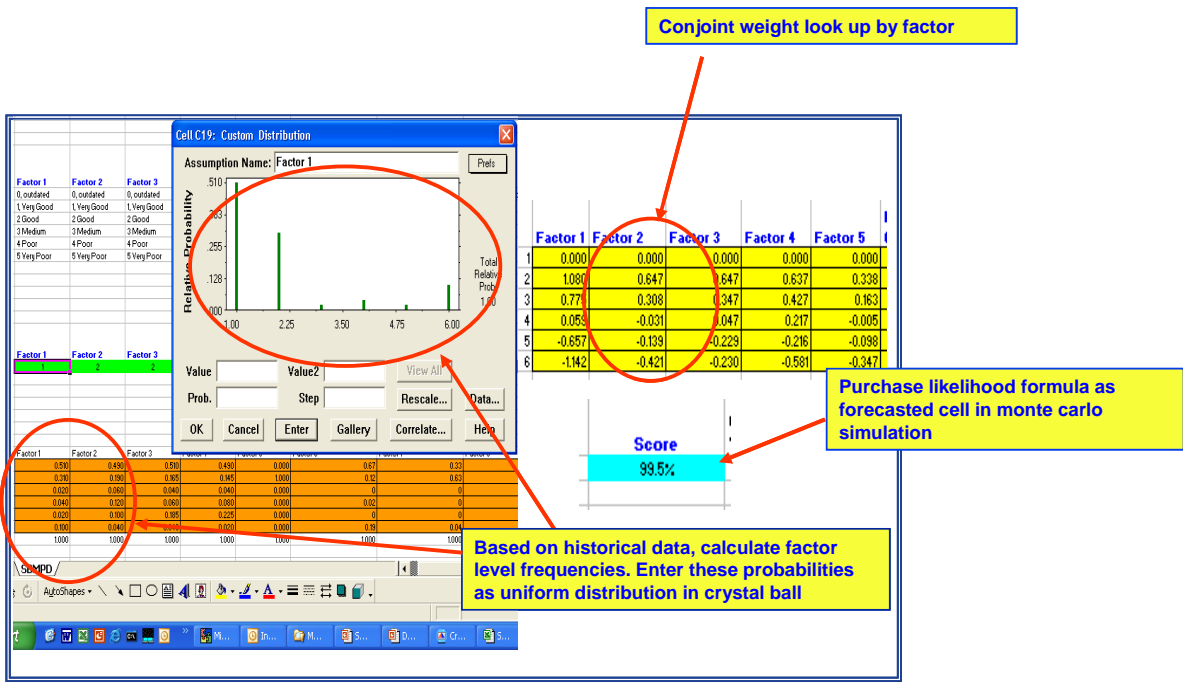
Our primary motivation for using this using this approach was to derive empirically logical threshold values that could be used to test whether deals used for validation purposes should be approved or rejected. Another objective was to have a better understanding of likely score distributions we would expect to witness into future based on larger number of deals being scored.

PROCEDURE

1. Develop spreadsheet model using deal approval likelihood model¹⁶ using lookup table for conjoint weights for each attribute level. The forecast cell is the resulting model score.
2. From a random draw of deals, determine relative probability for each attribute level.
3. Each attribute level has a uniform distribution assumption defined with unique probabilities as determined from step2.
4. Lastly, a look up function is added such that when the simulation runs a simulated deal scenario, the corresponding conjoint weight for an attribute level is used in calculating deal approval score.

¹⁵ A system that uses random numbers to measure the effects of uncertainty in a spreadsheet model

¹⁶ The equation transforming the product utility to relative deal approval likelihood is as follows $D_i = (100) * e^{U_i} / (1 + e^{U_i})$ where U_i is the utility for deal i . Deal approval predictions are averaged across the 3 executive decision-makers in order to reflect an average deal approval likelihood for executive committee.



DISCUSSION

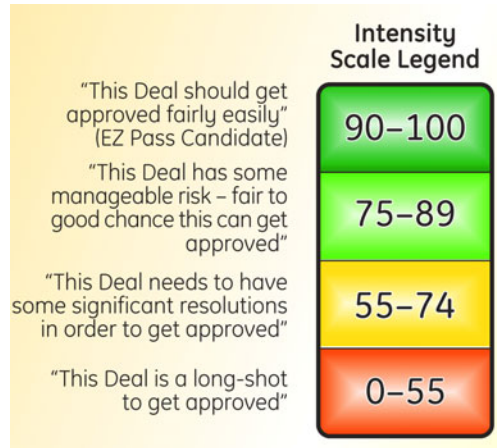
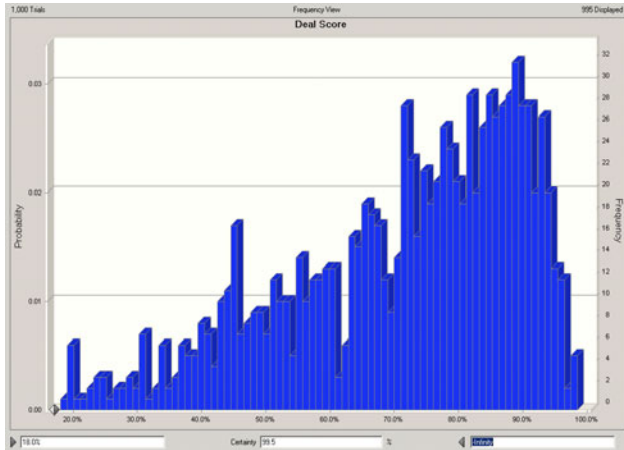
The Monte Carlo simulation produced a simulated set of 995 financial deals taking into account the executive committee’s conjoint part-worth utilities (average of 3 executive decision-makers) along with relative probabilities of historical deal attribute levels.

The below illustrates a negative skew, suggesting an expectation that a majority of deals to be scored should be in higher range. This was an intuitive result to the GE business as the process and people generally pursue deals that have a reasonable chance of getting approved.

We suspect the predicted lower range of scores might be slightly overstated because we did not take into account correlation among attribute levels. Wherein reality, deal teams that are considering deals with some poor attribute levels tend to have offsetting attributes that make deal somewhat compelling to GE executive team members (compensatory behavior).

In addition to being intuitive at a macro level, a second benefit from this work comes in being able to define threshold ranges into any number of equally probable bins. For purposes of our project, we decided to use four equally probable bins. Due to higher concentration of deals predicted on higher score range, the upper bin contains a narrow range of scores (e.g. 90-100) as compared to the lower bin (e.g. 0-55) even though according to the Monte Carlo simulation, a simulated deal is equally probable to occur in either bin. Another rationale for having multiple bins as opposed to only two is that the additional bins might differentiate deals substantially enough allowing for faster decisions. (Deals in the most upper bin should in theory be far superior to deals in bottom two bin categories).

One drawback to this approach is arbitrary assignment of approval threshold and number of bin categories. (We determined that 75 was the cut-off for approval). Our preferred methodology would entail deriving binning assignments directly from executive committee member responses. We are currently considering additional research that would allow us to directly capture thresholds limits for each individual respondent.



One option considered was to generate scores for historical deals with known outcomes using the deal approval likelihood model. We could then proceed to rank the deal scores from highest to lowest while also appending decision outcomes. The threshold level for approval and rejection could then have been optimized thru one of several mathematical or qualitative approaches. This option would have resulted in a binary threshold level which would have been predominately influenced by the sample deal set.

ACKNOWLEDGMENTS

We would like to acknowledge the following individuals for their support and helpful advice:

Bryan Orme and Justin Luster from Sawtooth Software for design support and customized coding of the ACA survey, Jarrid Hall for leadership of data input consistency six-sigma project, Tom Repoff for all his statistical analysis in support of this project, Mike Clark and Peter Kalish for countless insights and advice they both provided over the course of this project.

REFERENCES

- [1] Smith, D., Fletcher, J., The art and science of interpreting market research evidence. John Wiley & Sons Inc.
- [2] Wind, Yoram, Green, Paul E., (2005) Market Research and Modeling: Progress and Prospects. International series in quantitative marketing
- [3] Aggour, K., Bonissone, P., Cheetham, W., Messmer, R., (2006) Automating the Underwriting of Insurance Applications. American Association for Artificial Intelligence ISSN 0738-4602
- [4] Quaile, J., Bollapragada, S., Ramanathan, K., Osborn, M., Ganti, P., GE's energy rentals business automates its credit assessment process. Interfaces Vol. 33, No. 5, September-October 2003, pp. 45-56
- [5] King, W. C., Hill, A., Orme, B., (2004) The 'Importance' question in ACA: Can it be omitted?. Sawtooth Software: Sequim WA.
- [6] Hoch, S., Kunreuther, H., Gunther, R., Wharton on Decision Making. Ch.5 Combining models with intuition to improve decisions. John Wiley & Sons Inc.
- [7] Sattler, H., Hensel-Borner, S., A Comparison of Conjoint Measurement with Self-Explicated Approaches. Springer.
- [8] Orme, B., (2005) Getting started with conjoint analysis. p 117. Research Publishers, LLC. Madison, WI.
- [9] Hammon, K., Coherence and Correspondence Theories. Cambridge Series on Judgment and Decision Making, 2003.
- [10] J. Surowiecki, The Wisdom of Crowds. Doubleday Press, Random House Publishers, USA, 2004.

ISSUES AND CASES IN USER RESEARCH FOR TECHNOLOGY FIRMS

EDWIN LOVE

UNIVERSITY OF WASHINGTON SCHOOL OF BUSINESS

CHRISTOPHER N. CHAPMAN

MICROSOFT CORPORATION

INTRODUCTION

We describe aspects of market research conducted in technology companies that make technology product research different from that of traditional consumer products. These include differences in both the products themselves and the business environments in which they are developed. Although many traditional market research techniques are appropriate in technology companies, they may be poorly utilized due to these structural differences. We explore these factors and offer brief case studies from projects at Microsoft. Finally, we provide certain recommendations for the implementation of preference modeling in the new technology environment.

TECHNOLOGY PRODUCTS ARE UNLIKE SHOES, SODA, AND DETERGENT

We characterize traditional marketing research as occurring when three conditions are met: (1) products are well-defined in their real or potential characteristics; (2) research methods allow marketers to gauge interest in those products for potential markets; and (3) business structures enable development decisions to be made based on this research. Such conditions occur to a reasonable degree in most consumer product categories. Among traditional consumer goods, specifying potential product configurations is not difficult (at least in principle) because the universe of product attributes is largely contained. For instance, a new running shoe product has a limited array of potential materials, colors, general shapes, and sizes. Similarly, the general composition and manufacturing methods for sodas, detergent, clothing, televisions, automobiles, and most other consumer products are well understood.

However, technology products often are not so well-specified. Even their general characteristics (e.g., the product category) may be poorly-defined, and researchers may not know which features of a product are feasible. Even where these characteristics are well understood by the firm, they may not be understood among consumers. This exposes the firm to considerable research risk where understanding of product characteristics differs from respondent to respondent.

Nevertheless, quantifying consumer interest in innovation can be especially valuable for firms selling technology products. Since new technology products and new features are very costly to develop in time and money, early market intelligence can help firms to allocate effectively resources between development projects. Also, small differences in given features may have a large impact on cost-of-goods, particularly when dealing with such features as onboard memory, image resolution, and screen size. Further, interaction effects between features may be substantial. The value of image resolution, for example, will depend substantially on the viewable image size. Finally, technology products typically follow highly truncated lifecycles,

typically no more than 5 years and sometimes much less. This puts additional pressure on firms to develop the correct set of features in the initial version.

TECHNOLOGY PRODUCTS CAN BE DIFFICULT TO RESEARCH

Traditional marketing research techniques rely upon consumers' capability to understand new products, but discontinuous innovations presented by technology products often cannot be understood by consumers until the product is actually created. Since the outcome of such research is, or should be, a primary input into the product definition and development, this creates something of a chicken/egg paradox for technology firms. Product features cannot be well understood by research subjects without some degree of experiential context, and such a context cannot easily be provided without at least a working prototype available. In most cases, high cost and development time preclude firms from developing prototypes without some degree of confidence that the product will ultimately reach the market.

Consider for instance the concept of a digital pen, which captures ordinary handwriting with an ink pen on paper and makes it digitally available to various PC applications (for classroom notes, sketches, appointments, etc.). When Microsoft researched this concept in 2004, customers' perceptions of the product's capabilities were highly inaccurate, and this led to extreme variance in their appraisal of its value. After 1.5 hour focus groups discussing the concepts, respondents were enthusiastic and stated high levels of value (up to \$500). However, after using such a product and becoming familiar with its limitations (such as poor handwriting recognition and the need to use special paper), no customer stated that they would pay more than \$25. Microsoft opted not to pursue commercialization of that digital pen. Other firms have released similar products, although apparently without great success.

Because the level of consumer interest has often been unknown (and in many cases unknowable) at critical points in product development, business decisions for technology projects are often based on "macro" adoption models, such as the well-known Bass model (c.f. Bass 1969). These models are highly sensitive to assumptions about market size and adoption coefficients related to the behavioral characteristics of likely purchasers. These coefficients are themselves difficult to estimate *ex ante*, and a common approach is to estimate by analogy, i.e., by using the adoption coefficients from what are believed to be similar products that have already reached the market.

Marketers have a strong tendency toward choosing analogous products that have achieved some degree of commercial success, especially where similar products that have not succeeded are difficult to identify.

This general approach can bias organizational thinking in favor of products that are entrenched in engineering or management, even when evidence suggests that customers do not want such a product. A common retort is, "Of course they don't want it because they've never seen anything that can do what this will do. Once it's real, they will want it!" Lacking effective customer-level research methods, the only way to disprove this is to develop and manufacture the product. In such an environment, the role of marketing becomes one of developing channel and promotional strategies for products that are defined by engineers.

Consumer Reference Price Effects

It has long been held in marketing literature that consumers evaluate the attractiveness of a product's price relative to some reference price (Niedrich *et al.* 2001). Where a product is priced below the reference price, the price tends to be deemed attractive. Where it is priced above the reference price, it is considered unattractive. This effect is asymmetrical; the negative impact of prices above the reference price is generally greater than the positive impact of prices below the reference price.

Two general categories of reference prices have been modeled: memory-based and stimulus-based. Each has been shown to be predictive of consumer preference.

Memory-based reference prices: These references, sometimes referred to as internal reference prices, are developed based on the consumer's previous purchases and other experience within the product space (Kalyanaram and Winer 1995; Monroe and Lee 1999; Vanhuele and Dreze 2002). In essence, the consumer develops an estimate of the dollar value of the product prior to entering the purchase process.

Stimulus-based reference prices: Stimulus-based reference prices are formulated at the time of purchase. These are references are highly dependent on cues, such as the observed prices for comparable products, and context, such as the store environment.

In general, researchers have investigated the influence of either one or the other of the reference prices. These reference prices are sometimes modeled as single values such as the weighted average of observed prices. More recent research, however, has modeled reference prices as being drawn from a distribution of values at the time of purchase (Kalwani *et al.* 1990; Sherif M 1958), and hybrid memory-based/stimulus based reference price models are being developed (Park *et al.*).

When a consumer encounters a really new product, it is believed that they seek out one or more exemplar product(s) upon which to base a reference value (Mao and Krishnan 2006). Where no environmental stimulus is available, this exemplar will be sought in the individual's memory. It will therefore be highly dependent on the individual's experience and will vary from person to person. Where the product is extremely new, this variance can result in dramatic differences in perceived value. Should researchers provide cues as to the appropriate reference set, then they may bias their results.

We have found that technology products are in fact quite sensitive to consumer experience with a related product. In studies using webcams and digital music players as the product stimuli, we found that current owners of products respond in systematically different ways than customers who intend to buy the product but do not yet own one. In general terms, it appears that owners rate features highly when they extend current use cases or apply to core product experience features, whereas intenders rate features highly when they provide new use cases. Intenders do not distinguish as clearly as owners do between crucial features and relatively insignificant features. This suggests that product experience may be informative regarding the true comparative value of the product features.

There is also good reason to believe that reference feature levels exist. These may be construed as expected levels of performance (Love and Okada, In progress). Where product features do not meet their expected level of performance, then demand drops off dramatically. When a product feature meets the expected level of performance, however, there is diminishing

sensitivity among consumers to further improvements to that aspect of the product. As with reference prices, these levels of performance expectation vary between groups of consumers.

The Preference/Experience tradeoff

Features may also appeal to users even when they would have no effect on the underlying product experience. Consider, for example, that consumer webcams today offer a maximum resolution of 2.0 megapixels. In a small study with 40 respondents, we found that higher resolution (in megapixels) was the most strongly preferred of 14 potential webcam features. However, this improvement would be largely irrelevant in a new product, because Internet bandwidth, PC processing power, and USB interface bandwidth cannot stream such a video signal.

This creates a serious resource allocation problem for the firm. Should it choose to develop features that are preferred by respondents but add no experiential improvement, the product will provide lower long-term satisfaction and value. Should it choose to ignore stated preferences and develop features that it believes will most enhance product value, then the product may be perceived as deficient by consumers. Developing both types of features will make the product noncompetitive in yet another dimension: price.

In the case of the webcam (as in many other cases), the megapixel number provides a convenient measure of comparison between products. A consumer choosing between a 2 megapixel camera and a 4 megapixel camera may be assured that they are getting more megapixels with the latter than the former, which they implicitly associate with greater image detail and higher overall quality. The user preference must be reinterpreted: users are communicating that they want better video experience than current products offer, and the feature score is a measure of that desire. As researchers, we are challenged to make the correct inferences from the data.

ADAPTING MARKET RESEARCH TO NOVEL TECHNOLOGY PRODUCTS

The solution to these problems, we believe, is found in the combination of traditional market research techniques with the methods of user research (aka usability engineering). User research focuses on how to understand users' needs behaviorally and bring that understanding into an engineering process. In this process, product concepts begin as vague ideas and gradually become more and more specific until the final product is delivered. This approach is generally engineering focused and may occur in relative isolation from executive decision making and consumer research.

We have found that when choice-modeling is carefully integrated into user research and matched with other market research methods, product viability may be determined much more rapidly and product feature sets may be specified with greater confidence.

Furthermore, insights gained from this approach may then be rolled into macro models that may provide much more realistic adoption estimates. Where these models are developed in combination with reliable demographic, psychographic, and behavioral survey information, it becomes possible to create product performance scenarios and to test the sensitivity of such scenarios to different product, market and promotional characteristics. It also becomes possible to make predictions regarding competitive response.

Choice-based research techniques have proven highly informative in our research by helping to narrow the field of possible products and features of those products. For instance, scaling methods (such as MaxDiff) can be used to quantify customers' stated needs, dissatisfactions with current products, interest in general product concepts, and usage cases. Instead of asserting that a concept is "exciting" on the basis of speculation or focus group discussion, it can be objectively evaluated against other concepts and user needs. However, these techniques must be applied carefully; because features of new products are not well understood, customer data can be unstable and easy to misinterpret.

In one example, we conducted an online survey of 1008 respondents that asked about a novel webcam feature that would enable new use cases. We used both a conjoint format and MaxDiff format. Respondents showed significant interest in the feature in conjoint analysis ($p < .01$), as shown in Chart 1.

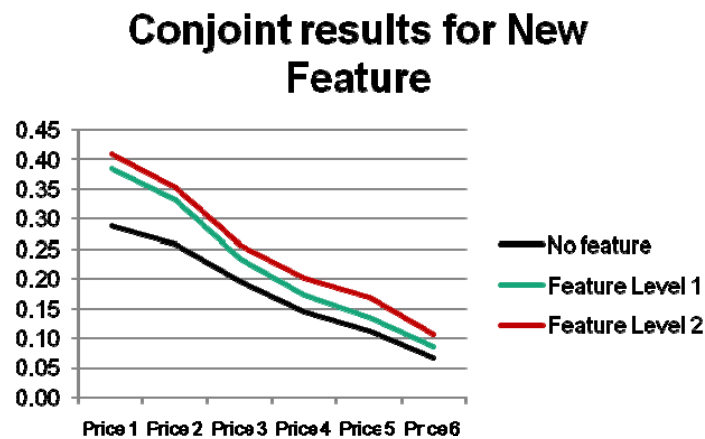


Chart 1

However, that same feature ranked 8th of 22 features in the MaxDiff exercise, as shown in Chart 2, scoring no better than would be expected from a random response set with 22 items (average scaled score of 4.9).

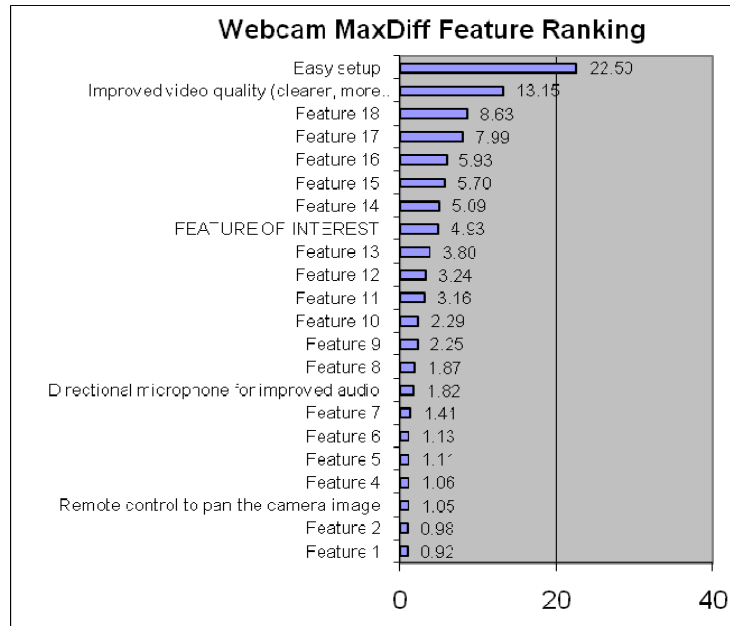


Chart 2

Such a result can appear puzzling, but is explained by four factors. First, they assessed a slightly different set of attributes between the conjoint and MaxDiff, and we may not be able to assume independence to the non-included alternatives. Second, because MaxDiff items can be non-specific they may have been so vague as to skew the results. In the present case, the top-scoring MaxDiff concept was “Easy Setup”. When a concept applies to every possible customer and use case, it is not surprising that it would score well and consequently depress the scores of more specific product concepts.

Uncertainty in the product feature set implies uncertainty in the value product value proposition. Therefore, pricing research must be conducted with caution. In general, we opt to include a broad range of price levels that provide information regarding the relative impact of changes in other features. When price is considered in this way as a relative attribute, the results are not necessarily indicative of exact market pricing.

Clearly, such effects may lead to pitfalls in the interpretation of choice-based research. We therefore suggest the following general rules for applying such methods in early product research:

- Assess concepts and features in multiple ways, using multiple methods. We have found that qualitative and quantitative methods can be used together in order to achieve results that are both interpretable and actionable. Multiple methods are particularly important where features are not well understood.
- Choice-based research in the early part of lifecycle must be combined with careful behavioral analysis. Features must be rationalized with regard to real product experience, which users may not be able to anticipate. Selecting features to maximize choice preference may yield an experientially inferior product, and a “superior” product may not be maximally preferred. Make inferences regarding stated preferences where necessary.

- Careful attention must be paid to systematic differences that affect specific population subgroups. In particular, for technology products, customers who have experience in a product space may make markedly different choices than potential customers without such experience.
- Eliminate vague or poorly defined features from quantitative analysis. Such features may skew respondents' evaluations of other features. Also, response to such features may create the illusion of specific value where none exists.
- If price is used as a product attribute, determine early whether it will be used purely as a relative gauge of value, or whether it will be viewed as corresponding to actual product pricing. Investigating actual pricing of new technology products requires careful design and we believe it generally benefits from multiple measures, involving multiple methods of measurement (conjoint, willingness to pay methods, auction or allocation methods, etc.)

FUTURE WORK

The issues discussed here present several opportunities for future work. Some of these areas are suitable for research projects conducted as part of work for individual clients or projects, while others would benefit from exploration across multiple projects and clients. We suggest the following areas for investigation:

- How to apply methods for rapid prototyping in order to obtain more realistic consumer evaluations. There are several approaches available for creating consumer prototypes of technology products, including: (a) evaluation of similar existing products; (b) early engineering prototypes of a product in development; (c) creation of visual demos such as interactive user interface models; (d) construction of prototypes from platform development products (such as mocking up a mobile product using a mobile PC); and (e) information acceleration alternatives (e.g., creating 3D virtual representations of products and purchasing or usage environments (Urban *et al.* 1997)).
- Exploration of market segmentation models with regard to product familiarity. If, as we suggest, consumer choices are markedly affected by general technology enthusiasm as well as direct experience in a product space, then such factors should be taken into account in market segmentation. One research question here is whether such factors apply in general across many product spaces, and if so, how they may best be characterized.
- Integration of market exploration in a systematic fashion with iterative behavioral research. We postulate that market research in technology companies should be integrated with behavioral design at the earliest time of a design process. However, to date, there is no systematic model for such integration. This is the subject of an upcoming work (Chapman and Love 2008).

REFERENCES

- Bass, F. M. (1969), "New Product Growth for Model Consumer Durables," *Management Science Series a-Theory*, 15 (5), 215-27.
- Chapman, C and E Love (2008), "Quantitative Early-Phase User Research Methods: Hard Data for Initial Product Design," *Hawaii International Conference on System Sciences (HICSS)*, January 2008, Waikoloa, HI
- Kalwani, M. U., C. K. Yim, H. J. Rinne, and Y. Sugita (1990), "A Price Expectations Model of Customer Brand Choice," *Journal of Marketing Research*, 27 (3), 251-62.
- Kalyanaram, G. and R. S. Winer (1995), "Empirical Generalizations from Reference Price Research," *Marketing Science*, 14 (3), G161-G69.
- Love, Edwin and Erica Okada (In progress), "Enhancing Product Features: Diminishing Sensitivity and the Acceptability Threshold " Working Paper.
- Mao, H. F. and H. S. Krishnan (2006), "Effects of prototype and exemplar fit on brand extension evaluations: A two-process contingency model," *Journal of Consumer Research*, 33 (1), 41-49.
- Monroe, K. B. and A. Y. Lee (1999), "Remembering versus knowing: Issues in buyers' processing of price information," *Journal of the Academy of Marketing Science*, 27 (2), 207-25.
- Niedrich, R. W., S. Sharma, and D. H. Wedell (2001), "Reference price and price perceptions: A comparison of alternative models," *Journal of Consumer Research*, 28 (3), 339-54.
- Park, Joo Heon, Douglas MacLachlan, and Edwin Love "Optimal New-Product Pricing Under Customer Anchoring Mechanisms," Working Paper.
- Sherif M, Taub D, Hovland CI. (1958), "Assimilation and contrast effects of anchoring stimuli on judgments," *J Exp Psychol.*, Feb;55(2), 150-5.
- Urban, G. L., J. R. Hauser, W. J. Qualls, B. D. Weinberg, J. D. Bohlmann, and R. A. Chicos (1997), "Information acceleration: Validation and lessons from the field," *Journal of Marketing Research*, 34 (1), 143-53.
- Vanhuele, M. and X. Dreze (2002), "Measuring the price knowledge shoppers bring to the store," *Journal of Marketing*, 66 (4), 72-85.

MINIMIZING PROMISES AND FEARS: DEFINING THE DECISION SPACE FOR CONJOINT RESEARCH FOR EMPLOYEES VERSUS CUSTOMERS

*L. ALLEN SLADE
COVENANT COLLEGE*

INTRODUCTION

One of the keys to successful discrete choice modeling (DCM) research is careful definition of the decision space, i.e., setting the attributes and levels to be investigated. Greater interdependence between the client organization and the conjoint survey respondent complicates the definition of the decision space in employee research. Marketing research respondents are largely independent of the research client, while employee research respondents are very interdependent with the research client.

Researchers should assess the degree of interdependence between the client organization and the respondents. A case study of a conjoint survey of rewards and employee turnover at Microsoft will be discussed. Suggestions are given on how to define the decision space to minimize counterproductive employee expectations and maximize research value.

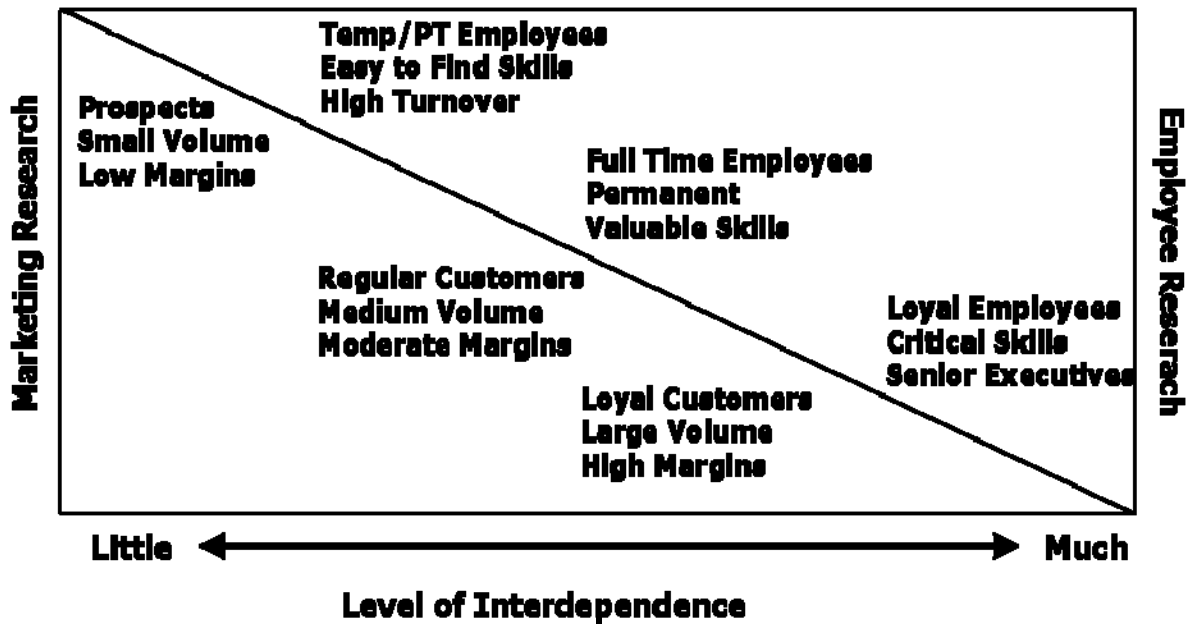
EMPLOYEE RESEARCH VS. CUSTOMER RESEARCH

There are many differences between employee research and customer research. One of the key differences is the level of interdependence between the client organization and the survey respondent. A higher level of interdependence complicates the definition of the decision space in employee research. Marketing research respondents are largely independent of the research client, while employee research respondents are very interdependent with the research client. In comparison to customers, employees have a longer term commitment to their employer, they depend on the employer for their livelihood, and they spend their workdays thinking about and acting on behalf of their employer.

The very nature of a choice survey tends to involve the respondent in a “what if” mode of cognitive processing while taking the survey. Making choices between jobs with different rewards in a DCM survey is more involving than simply clicking on strongly agree in a traditional employee opinion survey. This “what if” processing can be harmful when valued attributes such as pay or career opportunities are at stake, creating fear of cuts or false promises of increases.

A simple dichotomy of customers versus employees is an oversimplification. A more useful continuum of interdependence is illustrated below. It is often the case that employees are more interdependent with the research client than customers. Yet, there are situations where employees have less interdependence and customers have more interdependence. For example, a loyal customer who is a small business owner whose long term success and survival depend on the product may be more interdependent with the research client than an employee who is a high school student working a summer job. Current customers, part-time employees, contract

employees and temporary employees are at varying degrees of interdependence. Researchers should assess the degree of interdependence between the client organization and the respondents.



PROLIFERATION AND THE ENGAGING NATURE OF DCM RESEARCH

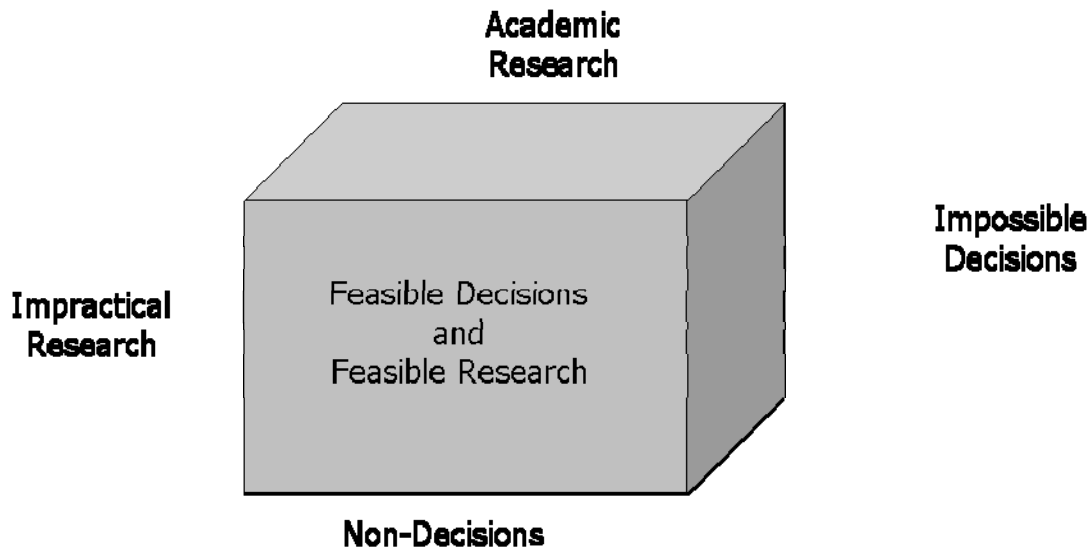
Discrete choice modeling (DCM) research is engaging for clients, for participants and for researchers. While the engaging nature of DCM research is a good thing, it is also a potential problem in that it can lead to expanding research scope. Expanding research scope is a problem with all applied research. For DCM, it often shows up in proliferating attributes and levels. The problem of expanding research scope may be magnified for DCM research with employees by creating false promises or unreasonable fears

Besides interdependence, researchers using DCM must also consider the proliferation problem. “Let’s play!” is a problem with all applied research. Successful researchers engage their clients with promises of practical problem solving. We often sell our research projects so well that the clients want to add more questions to the research. For DCM, the problem shows up most often in proliferating attributes and levels. On a Likert type survey, one more question is (usually) not a big deal. But with DCM, one more attribute or even one more level within an attribute substantially expands the number of possible combinations of decision choices, leading to impossibly long surveys or incomplete designs. Another problem associated with proliferation in DCM research with employees is the creation of false promises of rewards or unwarranted fears of takeaways.

Conjoint research may amplify the impact of interdependence between the client organization and the respondent. The very nature of a choice survey tends to involve the respondent in a “what if” mode of cognitive processing for the survey respondent. This “what if” processing can be harmful when valued attributes such as pay or career opportunities are at stake, creating fear of cuts or false promises of increases.

DEFINING THE DECISION SPACE

The decision space for an applied research project consists of the decisions that both the client is ready to make and that the research will help the client make. The decision space is the intersection of feasible decisions with feasible research. Excluded from the decision space for applied research are decisions which are impossible or non-decisions (i.e., where there is no choice to be made). Impractical research is excluded from the decision space, as is research which is mostly academic (designed for theory testing or methodological development). The diagram below illustrates the decision space for applied research.



In DCM research, defining the decision space consists primarily of setting the attributes and levels to be investigated. Other research decisions are also part of defining the decision space¹, but the point of this paper is to figure out when and how to limit the attributes and levels of a DCM project. Because of the problem of proliferation, the most common problem in DCM research is defining a decision space that is too large. It is possible that the decision space will start too small. If the decision space is too small, the researcher can help expand the decision space by helping the client to increase the number of alternatives with creative problem solving, mining past opinion surveys, employee focus groups on the topic, etc. However, it is more likely that the decision space will be too large because of the engaging nature of DCM research.

Here are three questions that researchers can use with their clients to limit the decision space in DCM research:

1. **“Would we be willing to do this?”**
2. **“How does this intervention compare to the others we are considering?”**
3. **“How would an employee react to taking this survey?”**

The first two questions deal with the feasibility of the decisions or interventions being researched. Question 1 (“Would we be willing to do this?”) deals with pragmatic issues and organizational politics. First, the decision space is constrained by the probability of the research

¹ For example, in applied research it is usually best to use “real people” as subjects (Gordon, Slade and Schmidt, 1986).

team recommending an intervention. If the team knows it will not propose doing something, then it should not research that intervention. Second, the decision space is constrained by the probability of getting approval for intervention. Impossible decisions should not be in the decision space either.

If the first question does not limit the decision space enough, the research team should ask question 2 (“How does this intervention compare to the others we are considering?”) to further limit the decision space. One way to compare interventions is to consider the cost/benefit ratio of each decision. Interventions with the highest cost benefit ratio should be dropped from the decision space first.

In the initial stages of planning the research, it may seem premature to rank interventions. After all, we are doing the DCM research to answer the question of how the interventions compare to each other. How can we compare the interventions when we don’t have the data yet? In my experience, the research team can and should use reasoned judgment and debate at this stage to narrow the decision space. This discussion should include researchers and content experts. In employee research on rewards, for example, the research team should include compensation experts, employee relations managers and training experts. Because we are prioritizing research questions (NOT making the final decision), we can trust our judgment.

It is valuable to cost the alternative rewards at an early stage of the research project. The costs should not just be a crude estimate – the costs should be an agreed upon commitment to deliver these alternatives at the stated cost. For example, if a training intervention is being considered to improve management effectiveness, the company’s training function should consider the exact nature of the training and what it would cost to deliver it. It is useful to get hard cost estimates at this stage, because interventions with higher absolute costs are riskier regardless of the projected cost/benefit ratio. These costs will also be useful if a simulator is built using the choice preference data.

The third question (“How would an employee react to taking this survey?”) is concerned with the impact of the research on employees. It is my contention that employee researchers should minimize promises and fears. Interdependence makes this question vital. Attitudinal research changes attitudes (e.g., the Hawthorne effect documented by Roethlisberger, F. J., 1939; cf. Bramel & Friend, 1981). If an employee completes a DCM survey with new hope of an implicitly promised reward or new fear of an implicitly promised cut, then that DCM survey has changed the employee’s attitudes. The hope of a new reward which is never actually delivered can turn to disappointment, disillusionment and possibly even the employee’s departure. The fear of a new cut in rewards can also lead to disappointment, disillusionment and departure even if the cut never actually happens. While DCM research cannot totally avoid creating promises and fears, the research team should limit the promises and fears to interventions that are in the realm of the feasible. Asking “How would an employee react to taking this survey?” can be a useful and non-threatening way to raise the issue of creating promises and fears among employees.

The greater interdependence of employees magnifies the problem of promises and fears. While DCM research with customers may also increase promises or fears, the impact of those new attitudes for customers is not as great as for employees.

MICROSOFT'S DCM STUDY OF REWARDS AND TURNOVER

Slade, Davenport, Roberts and Shah (2002) report a study of rewards and turnover at Microsoft that used Adaptive Conjoint Analysis (ACA). This study was intended to identify reward strategies that would help retain employees who were being attracted away from Microsoft.

In the Microsoft ACA study, employees were asked to choose among baskets of rewards based on a reward matrix similar to the diagram below.

Element	Level 1 (Low)	Level 2 (Medium)	Level 3 (High)
Annual Base Pay	No change in current annual base pay (with ongoing merit opportunity)	10% more than the current annual base pay (with ongoing merit opportunity)	20% more than current annual base pay (with ongoing merit opportunity)
Internal Job Market	No change in current internal job market practices (i.e., online career center, current transfer policy)	You may apply for other internal positions without manager's permission	Managers are allowed to actively recruit employees from other departments
Manager Effectiveness	No change in your manager's effectiveness	Organization invests to ensure that your manager is exceptional at delegating, motivating, being fair, and empowering	
Learning Opportunities	You negotiate training opportunities with your manager	Managers are held accountable for ensuring that employees receive at least 40 hours of formal training per year	Assurance of at least 60 hours of formal training per year, plus participation in a mentoring program
Health Care	You pay a total monthly health care premium between \$25-\$50 for all dependent coverage	No change to current health care program	You receive cash for waiving portions of health care coverage

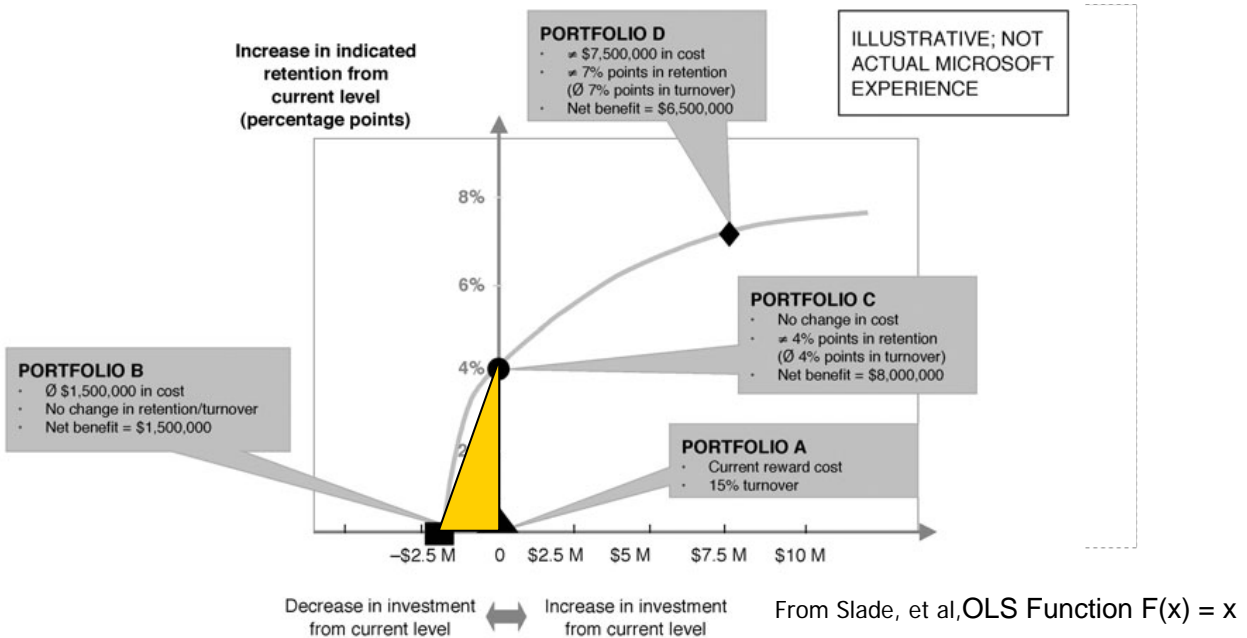
Current reward level
 Level not used

Note: This is a partial reward matrix. The rewards and levels shown here are similar, but not identical, to those analyzed by Microsoft.

Employees were asked to choose the basket of rewards most likely to motivate them to stay at the company for another four years. Slade *et al.* (2002) report some of the findings and applications of the research.

TAKEAWAYS AND THE SEARCH FOR THE GOLDEN TRIANGLE

The research team helped to sell this research project with “the search for the golden triangle” – that is, finding reward strategies where Microsoft could have the same turnover at less cost or less turnover at the same cost. The diagram below shows the optimum reward curve. The golden triangle is the area between Portfolios A, B and C. Portfolio A is the client’s current combination of reward strategy and the resulting turnover. Portfolio B is a superior to Portfolio A because the cost of rewards is reduced by \$1.5M with no change in turnover. Portfolio C is also superior to Portfolio A because turnover is reduced by 4% with no change in the cost of rewards. Portfolio D may also be a viable rewards strategy, but D requires an increased net investment. Portfolios B and C are “free,” in the sense that the organization can get better outcomes or reduced costs with no new net investment.



This golden triangle is a very attractive concept. It is a persuasive argument in funding a DCM research study on employee turnover. However, the search for the golden triangle requires less valuable rewards to be traded for more potent rewards. Potential reward takeaways could include pay cuts or reduction in benefits. Investigating takeaways can produce fears. More potent rewards might include pay increases and new or enhanced benefits. Investigating more potent rewards can produce false promises. It is not possible or desirable to totally avoid false promises or fears, but the research team should seek to minimize promises and fears by reducing the decision space as much as possible.

MINIMIZING PROMISES AND FEARS BY REDUCING THE DECISION SPACE

The research team at Microsoft included a reward attribute of change in base salary. For the attribute levels, the team considered a range of changes in pay, from a large pay increase to no change in pay to a pay cut. A very large pay increase was considered, and a moderately sized pay decrease was considered. One of the sponsors of the research was especially intent on having a large range of pay options. In the spirit of question 3 above (“How would an employee react to taking this survey?”), the team asked the sponsor several questions, such as “When employees start e-mailing Bill Gates about plans to cut their pay, what will you say to Bill?” and “Do we need to brief HR managers before the survey goes out on how to handle employee concerns?” This led to a more modest level of pay cut in the actual survey. In the spirit of question 1 above (“Would we be willing to do this?”), the team asked if we would ever actually consider the very large pay increase. When the sponsor said, “No, but I would like to see what the data say,” we explicitly discussed the problem of raising false promises. This discussion led to a more reasonable (but still healthy) pay increase as the highest level of the base salary attribute.

The team also considered changes in health benefits. Microsoft had a very generous set of health benefits, so enhancing benefits was not a reasonable option. The team investigated a

number of potential cuts in health benefits. Some very extreme ideas were proposed for the DCM survey. The research team asked question 1 “Would we be willing to do this?” When we determined that, no, there were certain reductions we would not be willing to do, we reduced the number and size of the proposed reductions in health benefits in the actual DCM survey.

Even with the reduced number of levels in the reward attributes of base salary and health benefits, there were too many attributes and levels. The team asked itself Question 3, “How would an employee react to taking this survey?” We found that merely reviewing the list of attributes/levels did not capture the experience of actually completing the survey. Instead, the research team took the survey repeatedly. We determined that the survey was too long and would undermine responses to the MS Poll, which is Microsoft’s premiere annual employee opinion survey.

To reduce the number of reward attributes and the number of levels within attribute, the team reviewed the entire reward matrix, asking Question 2 repeatedly, “How does this intervention compare to the others we are considering?” The team requested cost data on all the proposed attributes and the level of each attribute. The departments who would deliver the new or enhanced rewards were asked to provide commitments, not merely estimates. These commitments were in the form of both a budget and a plan to actually accomplish the new level of the reward. These commitments were relatively straight forward for certain proposed changes (like base pay or changes in health benefits) but more complex for softer reward attributes like increasing the quality of management. Some proposed rewards were eliminated because the department responsible for delivering that reward could not make a specific commitment. Some levels of rewards were eliminated because the absolute cost was too high, making the reward strategy too risky. And other rewards or levels were eliminated because the cost/benefit ratio was too high. We eventually ended with a reward matrix that represented a reasonable decision space. The reward attributes and the levels were feasible for Microsoft to implement. And, as much as possible, we minimized promises and fears that would have an adverse impact on Microsoft employees.

To confirm the judgments of the research team, we pilot tested the survey with a small sample of employees. We actually observed people taking the survey, and then got their reactions. The pilot test with employees confirmed that the revised survey was not too long and did not unduly increase implied promises or fears of reward cuts.

CONCLUSION

Care in defining the decision space is always worthwhile to maximize return on research investment. Extra care should be taken as the interdependence between the research client and the research subject increases. Here are three questions that researchers can use with their clients to limit the decision space in DCM research:

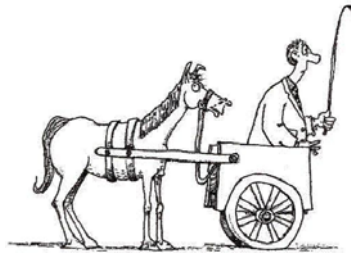
1. **“Would we be willing to do this?”**
2. **“How does this intervention compare to the others we are considering?”**
3. **“How would an employee react to taking this survey?”**

REFERENCES

- Bramel, D. & Friend, R. (1981). Hawthorne, the myth of the docile worker, and class bias in psychology. American Psychologist, 36, 867-878.
- Gordon, M. E., Slade, L. A., & Schmitt, N. (1986). The 'science of the sophomore' revisited: From conjecture to empiricism. Academy of Management Review, 11, 191-207.
- Roethlisberger, F. J. (1939). *Management and the worker; an account of a research program conducted by the Western electric company, Hawthorne works, Chicago.* Cambridge, Mass.: Harvard University Press, 1939.
- Slade, L. A., Davenport, T. O., Roberts, D. R., & Shah, S. (2002). How Microsoft optimized its investment in people. Journal of Organizational Excellence, 22, 43-5.

A CART-BEFORE-THE-HORSE APPROACH TO CONJOINT ANALYSIS¹

ELY DAHAN
UCLA ANDERSON SCHOOL



This research proposes a new method of measuring and estimating individual preferences for product attributes. The primary advantages include faster, lower cost preference measurement with approximately equal predictive accuracy. Additional advantages arise in specific situations such as too many attributes or the possibility of hybrid utility functions combining compensatory and non-compensatory preference with or without attribute interactions.

While mostly conforming to traditional conjoint analysis methodologies, the conjoint adaptive ranking database systems (CARDS) differs from traditional conjoint in important ways as shown in Table 1.

Table 1 Comparison of Traditional Conjoint Analysis vs. CARDS

<u>Traditional Conjoint Method</u>	<u>CARDS Method</u>
<ul style="list-style-type: none"> • Collect Data first, then analyze it 	<ul style="list-style-type: none"> • Build utility <i>database</i> first, then collect data
<ul style="list-style-type: none"> • Respondents must rate products they don't like 	<ul style="list-style-type: none"> • Respondents focus on preferred choices (CBC)
<ul style="list-style-type: none"> • Number of stimuli > Number of parameters 	<ul style="list-style-type: none"> • Number of responses < Number of parameters
<ul style="list-style-type: none"> • Inconsistency errors are common 	<ul style="list-style-type: none"> • No inconsistency

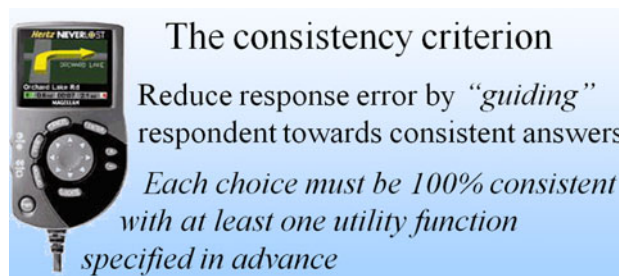
While traditional conjoint analysis collects data from respondents before estimating utility function parameters, CARDS starts with a database of possible utility functions in advance of collecting any respondent data. (The exact construction of the utility function database depends on the researchers' goals, and will be discussed shortly.) Traditional conjoint requires respondents to evaluate preferred attribute bundles as well as those that would be completely rejected, while CARDS places much greater emphasis on preferred profiles. While CARDS

¹ This research was voted as Best Presentation at the 2007 Sawtooth Software Research Conference in Santa Rosa, CA.

depends on ranking profiles, the respondent experiences a survey methodology that appears more like a choice-based conjoint task.

Due to its database structure and underlying assumption of accurate and consistent responses, CARDS does not necessarily require a greater number of responses than the number of utility function parameters being answered as traditional conjoint commonly requires. The key assumption underlying CARDS is that respondents answer in a consistent fashion, i.e. that the decision rules they use apply consistently throughout the survey process and that such errors can be reduced by simplifying the questioning task. Traditional conjoint assumes that inconsistency errors are common, and measures them through metrics such as violated pairs.

Figure 1
Helping respondents “navigate” towards their preferences



The underlying assumption of internally consistent responses, highlighted in Figure 1, is embedded in the CARDS method and is implemented using a database of utility functions that represent all the preference possibilities being considered by the researcher. This list of utility functions can span the entire continuous space of preference possibilities (in a discrete way at various levels of “resolution”), or can be narrowed based on prior knowledge about the problem at hand. The process of helping respondents “navigate” the database towards their individual utility is similar to a navigation system helping a user enter the desired destination.

For example, if searching for “Main Street” on a navigation system, the user would first type “M”. At that point, the virtual keyboard would eliminate most letters of the alphabet other than A, C, E, I, O, U, and Y, since no street name in its database has any of the other 19 letters as the second letter of its name. This process of elimination utilizes a database in the storage area of the navigation device. The database was set up before the user even entered the “M” in Main Street.

Likewise, once a respondent chooses his or her favorite full-profile conjoint card out of N full profile cards, very few of the remaining N-1 cards can be his or her second favorite one given that the respondent is navigating towards a particular utility function. Here, the analogy is between the letters of the street name and the full conjoint profiles within a perfectly consistent rank ordering. For N conjoint full product profiles, there are N! possible rank orders, but only a few of them are perfectly consistent with a utility function, typically fewer than 1-2% for the 18-24 profiles in most conjoint studies. Of course, one could probably devise unusual utility functions consistent with almost all N! possible rank orderings, but most researchers impose some limitations on the form of utility that is considered reasonable and informative. One implication here is that in traditional conjoint analysis, in which any rank ordering is permissible,

there is a high probability of at least some degree of internal inconsistency. In fact, this is exactly what we observe in practice.

The validity of the consistency assumption is empirically testable, for example by comparing the hit rates of CARDS versus traditional conjoint analysis in predicting holdout choices and ranks. For this purpose, we employ the example of smart phones shown in Figure 2, similar to the attributes tested in Yee, *et al.* 2007 (“Greedoid paper”). A second example of iPod music players is also tested. The data analyzed were collected over one week in June, 2005.

Smart Phone Example



Figure 2
Seven (7) attributes of smart phones for the empirical test

The conjoint profiles used in both empirical tests combine visual and verbal cues so that each attribute bundle appears to be a “real” product. In the case of iPods, the empirical test included incentive compatibility in that respondents chosen through a lottery received \$400 each towards the purchase of the real iPod fitting their holdout responses. Respondents received their first-choice iPod and kept any difference between \$400 and the price of that iPod. They were also given the option of keeping the \$400 and receiving no iPod.

Figure 3
Two products x Two methods

Empirical Test (June 3-10, 2005)
(*within subject*)

- Two examples: Phones & iPods
- Four cells:

		First Product Tested	
		Phones	iPods
First Method Tested	Ranked Conjoint	1 <i>n</i> = 92	2 <i>n</i> = 87
	CARDS	3 <i>n</i> = 99	4 <i>n</i> = 97

- Each respondent saw both methods & products
- Order was randomized

16

As shown in Figure 3, each respondent was exposed to both smart phones and iPods, and to preference measurement methods, cards and rank-order conjoint analysis. The ordering of method and product was randomized, resulting in four experimental cells. It was later determined that neither method nor product ordering affected the results, so the data were pooled.

% Correct pairs of 12 in Holdout Ranking Tasks

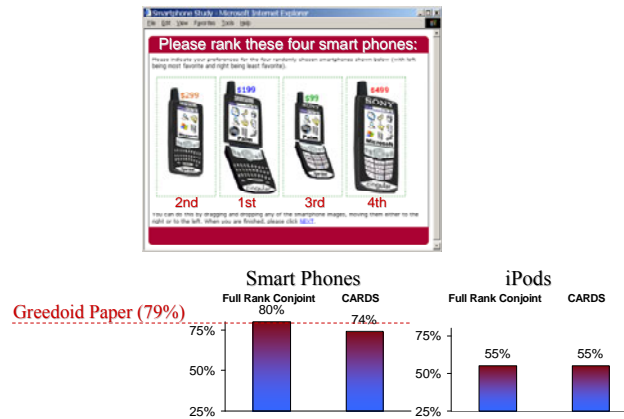


Figure 4

For both methods, respondents identified their consideration sets out of the 16 smart phone profiles and 18 iPod profiles using the method described in Yee, *et al.* (2007). Respondents then ranked the cards within their consideration sets, and then those that they had rejected. Of course, in the CARDS method, some profiles disappeared during the ranking task since they were not consistent as the next possible choice for any utility function in the database.

The results for two four-profile holdout tasks, summarized in Figure 4 and Figure 5 show that the CARDS method performed almost, but not quite as well as traditional ranked conjoint and the Greedoid methods for smart phones. While the iPod results were significantly less predictive,

CARDS performed as well as traditional conjoint at predicting the ordering of the four holdout and slightly better at predicting first choice out of four.

Hit Rate: First choice out of four (Two Holdouts)

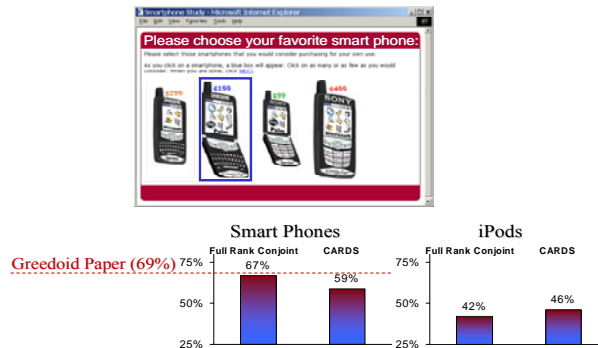


Figure 5

The database used in the smart phone test actually came from the Yee, *et al.* Greedoid empirical test. It consisted of the best fit utility functions from the approximately 500 respondents to their study. In essence, this database navigated each respondent in the present research towards the closest prior respondent with similar preferences. While this is clearly a risky approach when the prior respondents come from a different market segment than the one being tested, it worked here because of the similarity between the two test populations (MBA students). A database based on prior respondents corresponds to the fourth part of Figure 6, “Existing segments with known preferences.”

Figure 6

Four Potential CARDS Database Designs

- Span all possibilities (discretize the space)
 - The finer the “resolution” of the grid, the larger the database
 - But even very tight grids eliminate 98+% of possible rank orders due to consistency*
- Look at real world product options
 - Utility functions are matched to preferences for known products
- Use prior knowledge of utility (HB)
 - Database would follow the population distribution
- Existing segments with known preferences
 - The database would “guide” respondents towards known *segments*

The iPod test relied on a very different source for its list of possible utility functions. As depicted in the “Look at real world product options” part of Figure 6, one can create a set of possible utility functions to answer the question, “Which of the real product options that exist in the marketplace will this respondent prefer?” Since the number of real products that exist is typically finite, as was the case with the eight different iPods available at the time of the experiment, the number of possible utility functions needed to identify a respondent’s choice was also limited. Thus, the iPod utility database included enough options to explain virtually any preference ordering for the real products in existence, but no more than that.

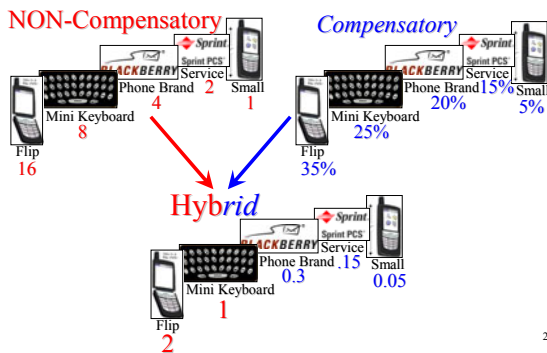
Other database design options might include knowledge of population preferences based on Bayesian approaches, or would be completely agnostic and span the entire universe of possible utility functions using a discrete representation of continuous space.

More sophisticated utility functions may also be captured by a CARDS database. For example, as illustrated in Figure 7, both compensatory and non-compensatory utility functions can be stored in the database, as can hybrids of the two. Similarly, interactions between attributes can also be included to supplement main effects models. One virtue of the cards approach is that the researcher need not commit to one form of utility modeling. Multiple underlying models can be included in the database, and respondents’ actions can dictate which model best explains those actions. Of course, it might be possible that two different models map to the same exact rank ordering of conjoint profiles, but we would expect this outcome to be extremely rare, especially with a sufficient number ($N \geq 16$) of full profiles.

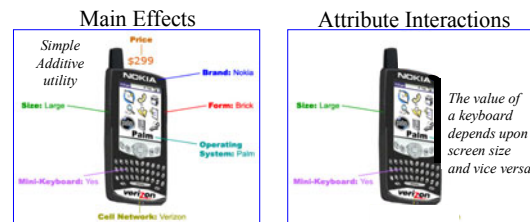
Figure 7

More sophisticated utility functions may be included in the database

- Mixtures of non-compensatory vs. compensatory



- Mixtures of main effects vs. interactions



The CARDS database can include **multiple** types of complex utility functions

Beyond producing reasonable predictive accuracy, CARDS dramatically reduced respondent effort, and could therefore reduce the cost of market research. In the case of smart phones, the number of clicks required was reduced from 16 clicks (in 3.9 minutes) in the traditional conjoint ranking task to a mean of 5 clicks (in 2.5 minutes) for the CARDS version. Similarly the iPod study improved speed from 18 clicks (in 3 minutes) to 6 clicks (in 2 minutes).

Table 2 summarizes a few of the advantages of the CARDS approach while also pointing out its deficiencies. We have seen that predictive accuracy is reasonably good while respondent effort is reduced, and one can imagine that developing relatively sophisticated utility function databases will become easier as storage technology improves. An extra benefit is that estimates of utility may be possible by looking at the possibilities not yet eliminated at any point in the

process. And CARDS tasks typically end well before respondents have to express preferences between product profiles that they have completely rejected (although the method could easily be modified to explore both extreme regions of preference first and leave the “middle ground” to be explored last).

Table 2
Effects of Enforcing Consistent Responses

Positive Effects	Negative Effects
<i>Good</i> : Predictive Accuracy	Early answers matter a lot
<i>Fast</i> : Minutes for resps.; quick analysis	Upfront work is greater
<i>Cheap</i> : Pack more into the same study	Need prior knowledge
<i>Easy</i> : 30% to 70% effort reduction	No real error theory
Scalable with storage tech	
Utility Scores as you go	
Emphasizes likes	

The potential downsides of CARDS include the importance of early responses. Were one to be looking for “Main Street” but accidentally type “N” rather than “M” as the first letter of the street name, it would be tough getting to the correct destination. A potential solution to this problem might be to ask a few confirmatory questions early on.

CARDS also requires a bit more advance work to construct the database before the first respondent even sees the task. That may require more prior knowledge of the product attributes, market population, and existing product space than has been common in traditional conjoint analysis. On the other hand, such knowledge may already exist for many firms in many long-running markets. And, finally, CARDS is not as theoretically satisfying as other approaches featuring sophisticated theories of error and response. As such, we would expect this to become a tool appreciated by practitioners more so than by academics, but even academics may appreciate the flexibility it affords them to simultaneously test out competing models of hybrid forms of utility.

We hope that future research explores the potential of CARDS to improve the efficiency of preference measurement, increase the number of attributes that can be explored, and identify more sophisticated forms of utility function that better capture to nuances of decision heuristics being employed when decision makers make choices.

Demonstrations of the CARDS approach described in this paper may be found at: http://www.webconjoint.com/cards_login.php?sid=13 for the smart phone example and at http://www.webconjoint.com/cards_login.php?sid=16 for the iPod example.

REFERENCES

Yee, M., E. Dahan, J.R. Hauser, J. Orlin (2007), “Greedoid-Based Noncompensatory Inference,” *Marketing Science*, 26: 4, pp. 532-549.

TWO-STAGE MODELS: IDENTIFYING NON-COMPENSATORY HEURISTICS FOR THE CONSIDERATION SET THEN ADAPTIVE POLYHEDRAL METHODS WITHIN THE CONSIDERATION SET

STEVEN GASKIN

APPLIED MARKETING SCIENCE, INC. (AMS)

THEODOROS EVGENIOU

INSEAD

DANIEL BAILIFF

AMS

JOHN HAUSER

MIT

1. INTRODUCTION

In most product categories, consumers simplify their choices by forming a “consideration set” of products (or services) that they will seriously evaluate before making a final choice (Hauser and Wernerfelt 1990, Roberts and Lattin 1991). There is evidence that simply knowing which products are in the consideration set can explain 80% of the uncertainty that could be explained with a logit-based model (Hauser 1978). This two-stage process is well-established in the academic literature as a realistic description of the process by which consumers make decisions (Payne 1976). Indeed there has been recent interest in analytic models in which the consideration stage is unobserved, but inferred from final choices (Gilbride and Allenby 2004, Jedidi and Kohli 2005).

The consideration set is often motivated by recognizing that it is optimal for consumers to balance search costs (evaluating all products) with opportunity costs (evaluating only those products most likely to be chosen). Because the products identified in the first stage (consideration) will be evaluated again in the second stage (choice), it is not unreasonable that consumers use heuristic processes in the consideration stage (possibly in the choice stage, too) that focus on a relatively few important features and do so in a simple (“first cut”) non-compensatory manner (Payne, Bettman and Johnson 1988, Gigerenzer and Goldstein 1996). This is particularly true when there are a large number of alternatives in the first-stage consideration decision (Payne, Bettman and Johnson 1993). There is evidence that such heuristics might be more efficient and lead to better selections than more-complex heuristics, particularly in situations similar to those that consumers face in real markets (Gigerenzer, Hoffrage and Kleinbölting 1991).

In this paper, we explore a two-stage consider-then-choose model that is grounded in this theoretical and empirical literature and attempts to mirror the purchasing process more naturally than the one-stage compensatory choice-only models typically used. In particular, rather than going straight to a choice-based conjoint (CBC) design we first ask respondents to indicate which profiles they would consider. We use these data to infer the heuristics that best explain each respondent’s consideration decision. Based on these heuristics, we identify the set of features that each respondent is likely to use for choosing from within the consideration set in the

second stage. We then generate an adaptive choice-based design to estimate the second-stage (compensatory) decision process.

We test the proposed two-stage model with new data on consumer consideration and choices of Global Positioning Systems (GPS), testing the proposed model against a traditional CBC design in which we use hierarchical Bayes methods to estimate partworths (HB/CBC).

We posit that the two-stage model will more accurately reflect consumer decision making and, hence, be more accurate. Moreover, because the first-stage consideration task is quick and easy for the respondent, we hope to be able to collect data more efficiently. We seek to make that data collection even more efficient with a display tool that enables us to present a large number of potential features from which the respondent can choose to use in the first stage of his or her consider-then-choose decision process.

2. ILLUSTRATIVE EXAMPLES OF MANAGERIAL RELEVANCE

Automobiles. Separating the steps of consideration and choice can provide important insights. Take, for example, automobiles. A consumer shopping for a new vehicle has a choice of hundreds of makes and models from which to choose. Because it is time-consuming and expensive to seriously evaluate every make-model combination (not to mention combinations of features within a make-model offering), the average consumer evaluates in detail far fewer make-models than the 300+ on the market – well under ten make-model combinations for the typical consumer. From a manufacturer’s standpoint, an automobile cannot be sold unless it is considered. The value to the manufacturer of getting its make-model considered is tremendous, not unlike reducing its odds of selling from worse than 1-in-300 to better than 1-in-10. If we can identify the features by which the consumers screen automobiles for their consideration sets, the manufacturer can assure that those features are available and prominent in any marketing communications.

Global Position Systems (GPSs). GPSs have long been used in navigation, but in the last five years they have become popular for the use in automobiles and, when handheld, in outdoor activities. However, they can be complex with many features such as accuracy, reception, weight, display resolution, etc. Furthermore, because of their relative novelty consumers are still becoming familiar with the meaning of these features for their own use of GPSs. For example, the REI web site provides a virtual advisor to help consumers select a GPS (Figure 1, <http://www.rei.com/rei/gearshop/advisor/gps.html>). It allows the user to shop by price, or by projected use (in an automobile, for fitness training, in the outdoors). By the use of filtering questions, such as, “Do you want a GPS Unit with a Quadrifilar Helix Antenna which may provide better reception in densely covered areas?,” the virtual advisor narrows down the choices to a small set of acceptable options which the consumer can examine in more detail.

Figure 1
GPS Finder Web Page at REI.com

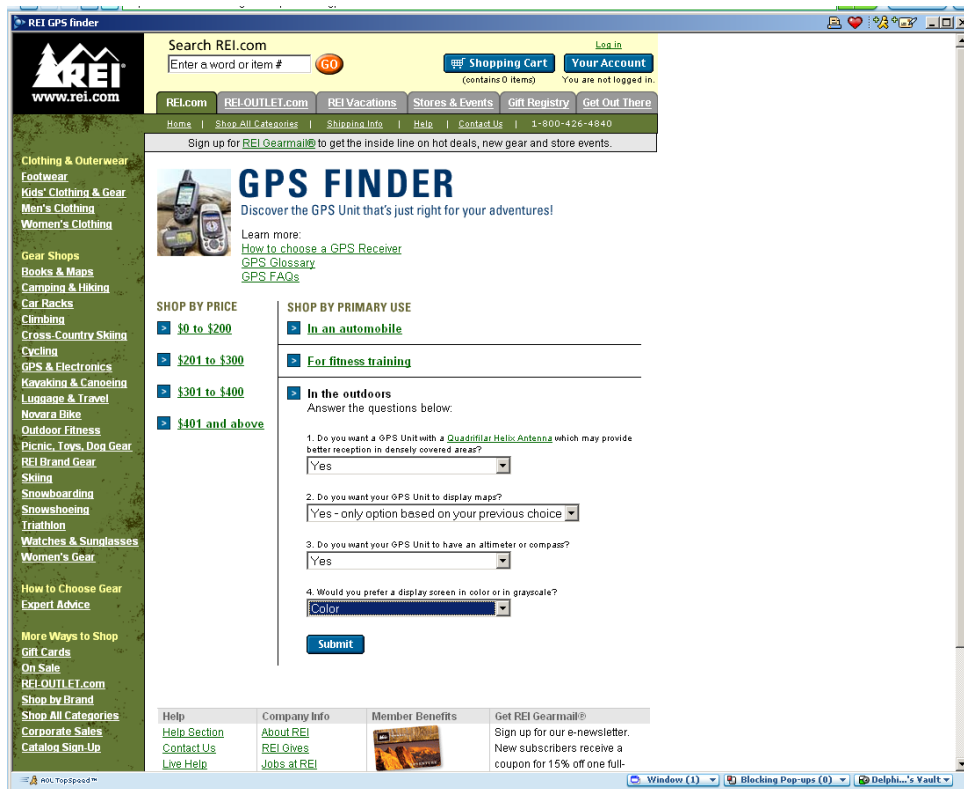


Table 1 summarizes sixteen important features of GPSs that were determined from qualitative research and pretests. The features can be represented by images and icons to provide visual cues which enable respondents to quickly evaluate profiles within a choice set (illustrated in subsequent figures in the paper).

To illustrate the difficulty of the consumer's decision process, consider Figure 2 which shows 32 such profiles – fewer than are available on the market. It would be unlikely that a consumer would use a compensatory process to choose a profile from the set of 32 profiles. More likely the consumer will simplify the decision process.

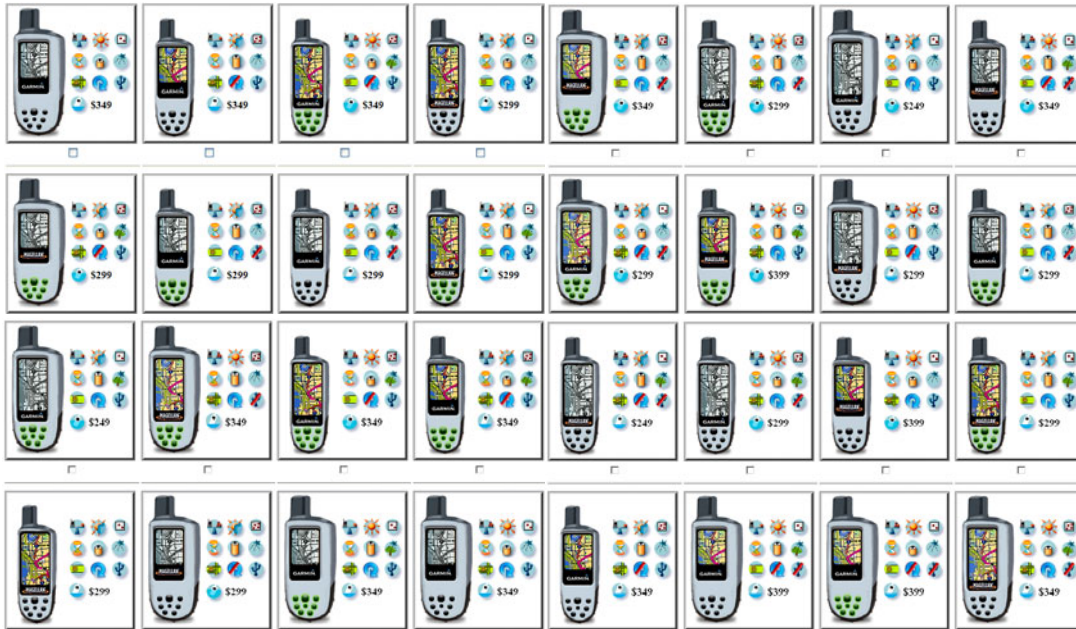
Table 1
Important Features for Handheld GPSs

Features 1 – 8*		Features 9 – 16*	
Level 1	Level 2	Level 1	Level 2
Color screen	Monochrome screen	Average reception	Reception under trees
Large screen	Small screen	Accuracy to within 50 ft.	Accuracy to a few feet
Garmin brand	Magellan brand	Track log	No track log
4 oz. weight	7 oz. weight	Mini-USB port	No port
normal display	Extra bright display	Floats on water	Does not float
High display resolution	Low display resolution	Large size GPS	Small size GPS
2 sec. acquisition time	10 sec. acquisition time	Backlit keyboard	Normal keyboard
30 hr. battery life	12 hr. battery life	Price increment (\$150)**	No price increment

* For many of these features, either level can be preferred by the respondent.

** The price that is shown to respondent is based on price increment and the cost of features rounded to one of four levels (\$249 to \$399). The rounding rule is chosen to approximate level balance.

Figure 2
 Illustrative Choice From Among 32 Profiles



Moreover, a one-stage compensatory model may lead to choices that may never have happened under a two-stage decision process with a non-compensatory consideration (first) stage. For example consider the two profiles in Figure 3 and suppose that the six features listed are the only features that matter to the respondent. (This alone is a simplifying heuristic.) But even with six features a partworth-based compensatory choice process might not capture a screening rule. For example, if the partworths are as shown, the respondent would choose the right-hand profile – its “utility” equals 30 “utils” compared to 26 “utils” for the left-hand profile. However, if, when faced with a large choice set, the respondent screens for a small handheld GPS with a color display, he or she will consider the left-hand profile and never even consider the right-hand profile. With such a screening rule, the respondent will never even evaluate the other features. If this were the true process that the respondent used when screening a GPS for final evaluation, the one-stage compensatory model would predict the wrong profile as chosen. When this is the case, we expect that knowing the process and the screening features is important to managerial decisions. If we were to assume a compensatory process or if we were to simply give the respondent a choice-based task in which there were never a large number of profiles, we would estimate the wrong model and make the wrong managerial decisions.

Figure 3
Comparing Two GPS Products – Compensatory Partworths



Color = 10	B&W = 3
Small = 7	Large = 1
\$349 = 7	\$249 = 15
Bright = 1	Extra bright = 6
Accurate = 1	Very accurate = 5

3. MEASUREMENT CHALLENGES

Figures 2 and 3 illustrate that if we are to identify the true underlying process we must design our measurement carefully. If the task does not have high fidelity with the environment faced by the consumer, then we may not capture the true process. In this paper we illustrate one attempt to mimic the environment faced by the consumer. We believe it is innovative and worth testing, but we do not claim that it is yet the best task we can develop. Our more-modest goals are to improve upon the standard choice-based task. If the task and analysis we test outperforms the current “gold standard,” hierarchical Bayes (one-stage) choice-based conjoint analysis (HB/CBC), then we will know we are on the right path, or at least on one of the right paths.

Consideration task. Our first goal is to represent the consideration task. One potential display format is to have a large number of profiles on the screen, as in Figure 2, but this is likely too difficult for the web-based respondent. Thirty-two profiles with sixteen features leads to images and icons that are too small for the computer displays we expect for most web-based respondents. As a compromise, we used eight profiles per screen and presented the thirty-two profiles in four sets. Figure 4 illustrates the consideration task. Respondents found this task natural and felt that it reflected the way they would select GPSs in a real market environment.

Figure 4
 Consideration Task with Eight Profiles per Screen

Please examine the items below and check the box next to the items that you would consider purchasing. If you would not consider the item for a purchase leave its checkbox blank.



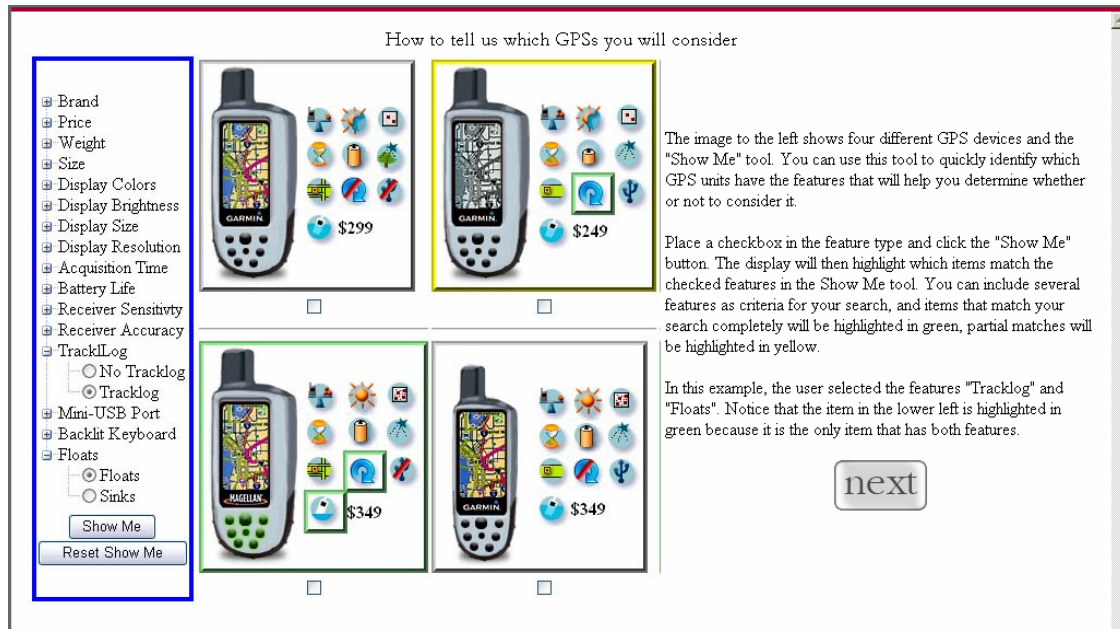
Even with this display we were concerned that the respondents might use the more-visual features, such as the size of the GPS, more often than those features that were represented simply by icons. (This would not affect internal validity testing, such as holdout tests or even validation tasks, because the same images would be used in both the estimation and validation tasks. However, a focus on visual features might affect external validity. At present this is a hypothesis to be tested.)

We also wanted a task that would scale well to a large number of features. For example, there could be fifty or more features in automobile choice. So that we might scale to a large number of features, we developed a means by which the respondents could highlight which profiles had which features. We felt that this process mimicked well the environments consumers face on the web (such as Figure 1) or in the store where the store manager chooses how to display items and the salespeople choose which features to highlight. We also felt that the tool should be entirely optional. The respondents could choose to use it or not and, if they choose to use it, use it at whatever depth (number of features) that they wanted.

The method we test in this paper is the Show-Me™ Tool (Figure 5). It is based on the “Christmas Tree” status board on U.S. submarines. The “Christmas Tree” status board has an array of colored lights. If they are all green, that means that all the doors and valves are closed before submerging. It has proven to convey information quickly, easily, and dependably. The Show-Me™ Tool allows respondents to use a pick list to specify some of the features they think they must have in order to consider a product. The chosen features are outlined in green. Those with some but not all of the “must have” features are outlined in yellow. Those with none of the “must have” attributes are not outlined at all.

There are other methods to display a consideration task.¹ For example, in a parallel study, also with GPSs, researchers at MIT are experimenting with five other tasks.² These tasks vary on whether the respondent can select the next profile to evaluate for consideration (from a “bullpen”) or whether the next profile is presented randomly. The tasks also vary on whether the respondent must evaluate every profile for consideration, just indicate consideration, or just indicate rejection. A final format tests text vs. icons.

Figure 5
The Show-Me™ Tool

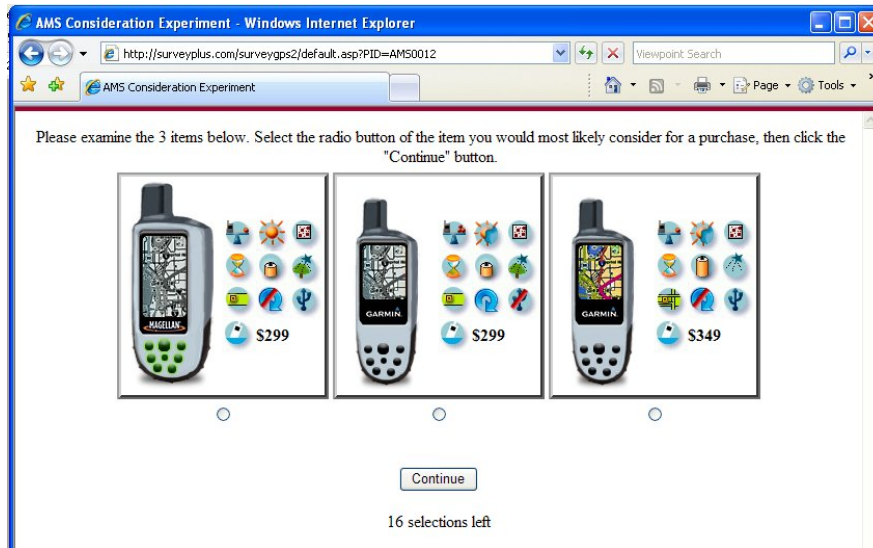


Second-stage choice task. After respondents complete the consideration task, we automatically infer the features that they used in the consideration task. Details are given below. Basically, if a feature is in every considered profile but no not-considered profile, then it is likely to be a non-compensatory feature and important to the respondent. We do not need further information on its partworth. Similarly, if a feature has no effect on which profiles are considered then it is likely unimportant and we do not need further information. The other features (not “must have,” not unimportant) identify a set of features for which we need to estimate partworths from the second stage of the decision process. We collect data with which to estimate partworths by using a standard choice task as illustrated in Figure 6. The profiles in the second-stage are chosen dynamically to gather as much information as feasible (details below). Figure 6 illustrates a choice task with three profiles; some choice tasks have more profiles and some have fewer profiles as dictated by the adaptive algorithm. (In the second stage the non-compensatory features do not change across the profiles since we do not need further information on them.)

¹ Future research can test whether the Show-Me™ tool encourages more or less heuristic processing or whether it just mimics the respondent’s natural purchasing environment. For example, when the MIT research is complete, we will have a baseline of non-compensatory processing against which to compare any effects of the Show-Me™ tool.

² The team includes Rene Befurt, Theodoros Evgeniou (INSEAD), John Hauser, Clarence Lee, Daria Silinskaia, Olivier Toubia (Columbia University), and Glen Urban.

Figure 6
Second-Stage Adaptive Choice-Based Conjoint Task



One-stage choice-based conjoint task (the benchmark). We compare the two-stage data collection and analysis to a standard HB/CBC analysis. To make the comparison as fair as possible, we use the same basic choice-task format as in Figure 6, except that the profiles are chosen from a standard experimental design where, unlike the second stage of the two-stage model, all features vary across profiles. In the CBC task, there were always three profiles per choice task.

4. ANALYSIS OF THE CONSIDERATION STAGE TO IDENTIFY PROCESS HEURISTICS

To analyze the first-stage of the consider-then-choice two-stage decision process we use the greedoid dynamic program (GDP) developed by Yee, Dahan, Hauser, and Orlin (2007) and summarized in a previous volume of the Sawtooth Software Conference Proceedings (Hauser, Dahan, Yee, and Orlin 2006).

The GDP is an efficient way to identify a heuristic that best fits the data. Although Yee, *et al.* (2007) tested it for full-rank and consider-then-rank data, it can be used for consideration-only data. The GDP searches efficiently over the set of all possible lexicographic heuristics to find the specific ordering of features that best groups profiles into those that are considered and those that are not. In a full-rank lexicographic rule, a respondent first selects one feature, say color display, and ranks all profiles with color displays before those with B&W displays. He or she next selects another feature, say brand (Garmin vs. Magellan in GPSs), and ranks profiles according to color-Garmin, color-Magellan, B&W-Garmin, and B&W-Magellan. The respondent continues choosing features until all profiles are ranked. While the theory is easy to understand for full-rank, it can also be applied to considered-vs.-not-considered data.

For such data the GDP considers all pairs of profiles and treats the data as if all considered profiles are ranked ahead of not-considered profiles. Any given lexicographic ordering ranks all profiles, but we can evaluate a heuristic on only those pairs for which we know the rank

(consider vs. not-consider). We pick the heuristic for which the predicted ranks violate as few of the actual observed pairs of ranks as possible.

In theory, the GDP can ultimately evaluate all features and place them into a lexicographic ordering. However, from the perspective of a two-stage model, we are only interested in the first few features. Furthermore, for consider-only data, there is a non-uniqueness issue. For example, if a respondent will only consider color-Garmin GPSs, then two lexicographic orders, (1) color then Garmin and (2) Garmin then color, will explain the data. This is not an issue for a two-stage analysis because we are only interested in the set of features used in the first stage, we need not (and cannot from the data) identify the ordering within such “must have” features.

We apply the GDP to the consideration data. The GDP provides a ranking of features based on the lexicographic hierarchy that is most consistent with the profiles the respondent considers. It is a bit more complex for multi-level features. The GDP actually works with aspects rather than features. An aspect is a binary description: color vs. B&W. For multi-level features, we code the feature as multiple aspects. (For more details see Yee, Dahan, Hauser, and Orlin 2007.)

The highest ranked aspects identified by the GDP are set aside as “non-compensatory” if they appear in 90% of the considered profiles. Our reasoning is that a “must have” aspect should show up in nearly every considered profile, subject to respondent error. Similarly, a “must-not have” aspect should show up in very few, if any, of the considered profiles. We selected the 90% cut-off rule based on judgment. Others rules might apply to different data.

We isolate these non-compensatory aspects and save them for later use in the prediction of consideration in the holdout sample. It would have been feasible to include price levels (e.g., \$199) as non-compensatory aspects and the GDP sometimes identifies a price level as such. However, we decided it would be more realistic to the respondent if we included price levels in the second-stage adaptive CBC exercise. Thus, we move price to the second stage whether or not it is a “must have” feature.

After setting aside the non-compensatory “must have” (and “must-not-have”) aspects, we select the top remaining features (based on the GDP rankings of aspects) for the second-stage CBC analysis. Because we sought a parsimonious description of the respondent and because we wanted a relatively simple CBC design in the second stage, we selected at most six features on which to collect tradeoffs in the second stage. This judgment was based on our experience in the category. It was also informed by the parallel MIT research on alternative data collection and analysis methods. The selection of six features is, naturally, subject to future research as is the possibility of including the lowest-rank “must have” features in the second-stage CBC exercise.³

There were some exceptions. If there were more than six obviously “must have” aspects, we included them all. If there were six or more non-compensatory aspects, we used the next two aspects in the lexicographic order (plus price) in the second-stage CBC exercise.

Summary of the consideration-stage analysis. Based on the consideration task in which the respondent identifies which of 32 profiles he or she would consider, the GDP identifies two sets of important aspects (features). The first set are the “must have” aspects that are set aside to determine which validation profiles are considered. The second set is the clearly unimportant features that are removed. The third set are the next highest in the lexicographic ordering of

³ We are grateful to Prof. Ely Dahan of UCLA for this suggestion.

aspects. These aspects are likely compensatory and likely to be used by the respondent in the second stage of the consider-then-choice process. Only these last aspects (levels of features) plus price are moved forward to the second-stage data collection – the adaptive CBC exercise.

5. THE SECOND-STAGE: ADAPTIVE CHOICE-BASED CONJOINT ANALYSIS

An important feature of the two-stage analysis is that the features in the second-stage CBC task are chosen specifically for each respondent based on that respondent's answers to the first-stage consideration task. Indeed, even the number of features in the second-stage CBC task is tailored to the respondent. Such customization requires that the CBC choice sets be generated “on the fly” for each respondent.

In addition, our overall goal is to develop questioning tasks that are perceived as realistic and put as little burden on the respondent as feasible. Having already asked the respondent to complete a first-stage consideration task, we wanted the second-stage CBC task to be as efficient as feasible – asking only as many questions as are needed.

We draw on the fast polyhedral adaptive conjoint estimation (FastPace) techniques developed at MIT, in particular, the adaptive CBC version (Toubia, Hauser and Simester 2004). The FastPace CBC technique (FPCBC) recognizes that each choice made by a respondent imposes constraints on the set of feasible partworths as illustrated in Figure 7. The green polygon represents the set of feasible partworths as determined by previous questions. (We have shown only two of the partworths – the actual set of feasible partworths forms a high-dimensional polyhedron.) If we ask a respondent to choose among two profiles, and if the profiles are chosen judiciously, the choice reduces the set of feasible partworths by approximately 50%. For example, if the respondent chooses profile 1, then the set of feasible partworths becomes the dark green region; if the respondent chooses profile 2 it becomes the light green region. More profiles in a choice set mean more cuts. For example, four profiles in a choice set divide the region into four sub-regions, each corresponding to the choice of one of the four profiles (Figure 8).

Figure 7
 Choices in CBC Tasks Shrink the Set of Feasible Partworths
 (Adapted from Toubia, Hauser and Simester 2004)

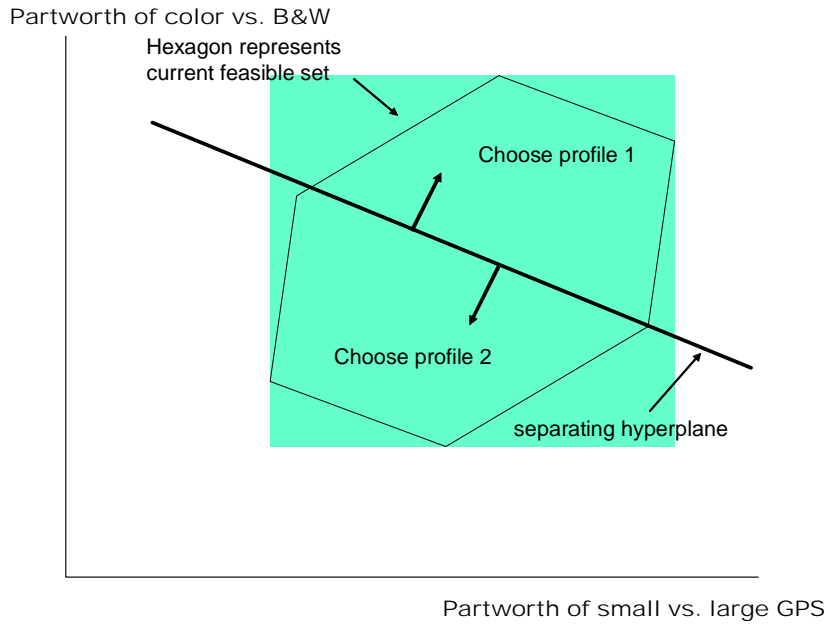
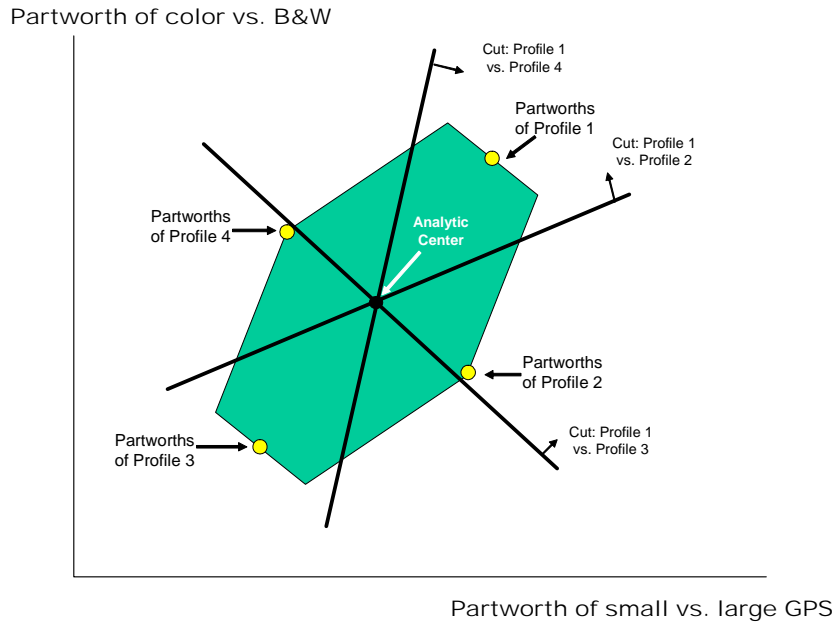


Figure 8
 Illustration of the FPCBC Algorithm for Four Profiles
 (Adapted from Toubia, Hauser and Simester 2004)



FPCBC chooses the profiles for the choice set to reduce the region of feasible partworths as rapidly as possible. This usually means that the regions are of roughly equal size and as close to

symmetrical as feasible. The mathematical details of the algorithm are beyond the scope of this paper and are described in Toubia, Hauser and Simester (2004). Basically, the algorithm first finds an ellipsoid that approximates the set of feasible partworths, then finds the longest axes of the ellipsoid and selects cuts that are perpendicular to the longest axes. Finally, it chooses profiles for the choice set by solving the consumer's budget problems (for target partworths) such that the choices among the profiles imply the selected cuts. The algorithm has proven to recover partworths (synthetic data) accurately and, in most cases, more accurately than alternative question-generation methods. Empirically, it performs as expected, reducing the choice set rapidly and achieving maximal-information choice balance. Recently, it has been improved to incorporate measurement error and managerial priors (Toubia, Hauser and Garcia 2007). In this paper we test the basic version recognizing that the performance of a two-stage model will improve when we move to the probabilistic version.

In our application, FPCBC generated choice sets that the respondents perceived as realistic and the set of feasible partworths converged rapidly to partworth estimates. One managerial advantage of FPCBC is that the partworths are estimated automatically during the questioning process and are immediately available for either managerial analysis or for further adaptive questioning. For example, although we did not test such adaptive branching in our surveys, one can imagine a series of adaptive open-ended questions to query a respondent on why some features are "must have" features (consideration stage) and why other features are important in choice (second, choice stage).

In this paper, we use the automatically generated estimates that are based on the "analytic center" of the region of feasible partworths that remain after the CBC questions are answered. However, one can also use these questions as input to a standard HB/CBC estimation. For comparative testing see Toubia, Hauser and Simester (2004) and Toubia, Hauser and Garcia (2007).

Summary of the two-stage analysis. After the respondent has completed both the consideration task and the FPCBC task, we have identified three sets of features: (1) "must have" features that the respondent used to make consideration decisions, (2) compensatory features that are important in the second-stage choice decision, and (3) unimportant features. For the compensatory features we have also estimated FPCBC partworths to describe the second-stage choice decision.

To make forecasts, we use the "must have" features to predict consideration sets. We then use the compensatory partworths to predict choice within the consideration set. Any profile or product that is not predicted as considered, we assume will not be chosen.

6. EMPIRICAL TEST OF THE TWO-STAGE CONSIDER-THEN-CHOOSE TASKS

To test and refine our methodology, we developed two web-based surveys. The first survey used both the consideration task in Figure 5 and the choice task in Figure 6. The second (control) survey was a traditional choice-based conjoint survey that used standard CBC tasks similar to those in Figure 6.

Both surveys used the same images (jpegs) and icons to represent the 16 features (including price). Five features (brand, size, display size, display color, and backlit keyboard) were represented in the images themselves and eleven features (weight, display brightness, display

resolution, acquisition time, battery life, receiver, accuracy, track log, mini-USB port, floating ability, and price) were represented by icons. We felt that the price should be realistic to the respondent, yet we wanted to manipulate price in the experimental design. We achieved both goals by defining a “base price manipulation (\$0 or \$150)” that was added to feature-based prices to get a total price. The total price was then “rounded” to one of four levels with the rounding rules chosen for level balance. For the consideration task we chose 32 profiles from an orthogonal array.⁴ The FPCBC profiles were chosen by the polyhedral method described above. The CBC profiles and choice sets (second survey) were chosen by a randomized design using all features which we created using Sawtooth Software.

The surveys were programmed in ASP. Both the GDP and FPCBC were coded in Matlab, and then compiled into DLLs using Microsoft ASP.NET. The database for the survey was coded in SQL. The HB/CBC estimation (CBC control) was programmed in Matlab using code developed by Toubia, Hauser and Simester (2004). This code was checked by those authors to give the same estimates as the Sawtooth Software HB/CBC module.

The surveys included an introduction, a description of handheld GPSs and their features, and either the consider-then-choose tasks (two-stage survey) or the standard CBC tasks (control). Following some questions about GPS usage we used “puzzler” questions to cleanse the mental palette (Frederick 2005). After the cleansing task, respondents completed a holdout task in which they were given eight GPS profiles (chosen from a master experimental design, different from the design used in the calibration consideration task). Respondents indicated which profiles they would consider and then ranked all eight profiles. (Although we do not need the consideration data to test the CBC analysis in the control survey, we included the task to make the holdout tasks the same in both the test and control surveys. Finally, respondents completed a battery of self-explicated importance questions and questions about the survey itself.

The sample frame was current or prospective users of handheld GPSs. Sample was provided from the Internet panel maintained and operated by Survey Sampling, Inc. We used the standard incentives provided by Survey Sampling, Inc. The number of completed interviews was 291 in the two-stage survey and 265 in the CBC control survey.

7. RESULTS: Comparison of Two-Stage Tasks to Traditional CBC

A minimum requirement for estimating a two-stage model is that the two-stage task be seen as interesting and enjoyable by respondents. Table 2 summarizes the results that we obtained. The two-stage survey was seen as significantly more interesting (48.8% vs. 35.8% top-box, $p = 0.012$) and more enjoyable (26.8% vs. 23.4%, $p = 0.29$, not significant) than the traditional CBC survey where both constructs were measured with five-point scales. We expect further refinements of the tasks to increase both interest and enjoyment of the two-stage survey.

⁴ The features were based on qualitative interviews and surveys to identify the features that were likely to be perceived as important by the target sample. The images and icons were developed jointly by Applied Marketing Science, Inc. and MIT and generated by Limor Weisberg (LimorDesign at www.limor.com). MIT is using the same images and similar icons with five alternative data collection formats and to evaluate improved algorithms based on disjunctions of conjunctions. Rene Befurt of MIT was instrumental in managing the production of images.

Table 2
Comparison of Interest and Enjoyment

	Two-stage Consider-then-choose		One-stage CBC Task	
	Percent top box	Percent top 2 box	Percent top box	Percent top 2 box
Interest	48.8%	84.5%	35.8%	78.1%
Enjoyment	26.8%	67.0%	23.4%	61.1%

We next turn to validity testing. We compared the hit rate (most preferred profile) between the two-stage consider-then-choose and the one-stage CBC surveys. The two-stage task/analysis (GDP/FPCBC) predicted 41.7% of the top-ranked holdout profiles correctly. The one-stage task/analysis (HB/CBC) model predicted 39.3% correctly. Although the results were not significantly different ($p = 0.66$), the two-stage model did show a slight advantage. At minimum we infer that the two-stage task/analysis is worth further testing; we expect its predictive ability to improve in future generations. It is encouraging that it does as well as traditional CBC, a method that has been developed over fifteen years and testing extensively.

We can also assess the two-stage model on its ability to predict consideration. We were able to predict 73.6% of the profiles correctly. We compare this to null model that predicts consideration randomly in proportion to the size of the consideration set. For the calibration data, respondents considered, on average, 5.6 profiles (17.5%) and for the holdout data they considered, on average, 20% of the profiles. This results in a null prediction of 69.5%. Thus, the actual prediction is significantly better than this strong random null model ($p = 0.00$). We can also test the GDP against a model based on the self-explicated importances.⁵ The GDP predictions are better, and significantly so than the 71.9% obtained with the self-explicated importances, $p = 0.09$ with a paired t -test.

⁵ We chose top box as the cutoff for non-compensatory importance for the self-explicated model based on the assumption that, if respondents judged a feature to be “must have,” they would likely give it a top box rating. This is, of course, subject to improvement and further testing.

8. SUMMARY AND FUTURE DIRECTIONS

Analyses to date suggest that two-stage consider-then-choose data collection and analysis have potential advantages:

- Two-stage data collection is based on respondent tasks that mimic those used by consumers to form consideration sets and make choices in market environments.
- Respondents find the tasks significantly more interesting and more enjoyable than the traditional choice-based conjoint task.
- The natural format and the increased interest and enjoyment are likely to enable us to handle a much larger set of features than traditional formats.
- Holdout hit rates are improved slightly with the two-stage analysis relative to one-stage (CBC) analysis. At minimum the two-stage process is at least as accurate as the traditional one-stage process – the next generation shows even more promise.
- The first-stage GDP predicts consideration sets significantly better than a strong null model and better than analyses based on self-explicated importances.

We view the survey tasks and analyses in this paper as a “proof of concept.” With further development we expect to improve choice hit rates, consideration hit rates, respondent interest and enjoyment, and realism. To the best of our knowledge, this was the first application of the Show-Me™ tool. We are working to improve its look and feel. This was also the first application of a combined GDP and customized FPCBC. We made a number of heuristic choices such as the 90% cut-off, the number of features advanced to the second stage, and the inclusion or not of “must have” features in the second stage. All of these assumptions are subject to test and improvement. Finally, we used the deterministic FPCBC as a first test. In future analyses we hope to experiment with HB/CBC to estimate partworths from the FPCBC-selected questions and we hope to experiment with the probabilistic versions of FPCBC.

Finally, we note that the MIT team is working on improved first-stage tasks and on analysis methods for predicting consideration. Early indications suggest that improved analysis of our first-stage data can increase holdout hit rates. We are optimistic for further improvements and for the future of two-stage consider-then-choose analysis.

REFERENCES

- Frederick, Shane (2005), "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*, 19(4), 25-42.
- Gigerenzer, Gerd and Daniel G. Goldstein (1996), "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review*, Vol. 1003, No. 4, 650-669.
- _____, Ulrich Hoffrage, and H. Kleinbölting (1991), "Probabilistic Mental Models: A Brunswikian Theory of Confidence," *Psychological Review*, 98, 506-528.
- Gilbride, T. and G. M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive and Compensatory Screening Rules," *Marketing Science*, Vol. 23, No. 3 (Summer), 391-406.
- Hauser, John R. (1978), "Testing the Accuracy, Usefulness and Significance of Probabilistic Models: An Information Theoretic Approach," *Operations Research*, Vol. 26, No. 3, (May-June), 406-421
- _____, Ely Dahan, Michael Yee, and James Orlin (2006), "'Must Have' Aspects vs. Tradeoff Aspects in Models of Customer Decisions," *Proceedings of the Sawtooth Software Conference* in Del Ray Beach, FL, March 29-31, 2006
- _____, and Birger Wernerfelt (1990), "An Evaluation Cost Model of Consideration Sets," *Journal of Consumer Research*, Vol. 16, (March), 393-408.
- Jedidi, K. and R. Kohli (2005), "Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers," *Journal of Marketing Research*, Vol. 42, No. 3, 483-494.
- Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search," *Organizational Behavior and Human Performance*, 16, 366-387.
- _____, James R. Bettman, and Eric J. Johnson (1988), "Adaptive Strategy Selection in Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 534-552.
- _____, _____ and _____ (1993), *The Adaptive Decision Maker*, (Cambridge, UK: Cambridge University Press).
- Roberts, John H. and James M. Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28, (November), 429-440.
- Toubia, Olivier, John R. Hauser and Rosanna Garcia (2007), "Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application," *Marketing Science*, 26, 5, (September-October), 596-610.
- _____, _____, and Duncan Simester (2004), "Polyhedral Methods for Adaptive Choice-based Conjoint Analysis," *Journal of Marketing Research*, 41, 1, (February), 116-131.
- Yee, Michael, Ely Dahan, John R. Hauser and James Orlin (2007), "Greedoid-Based Noncompensatory Inference," *Marketing Science*, Vol. 26, No. 4 (July-August), 532-549.

A NEW APPROACH TO ADAPTIVE CBC

*RICHARD M. JOHNSON
BRYAN K. ORME
SAWTOOTH SOFTWARE, INC.*

BACKGROUND

Choice-Based Conjoint (CBC) is the most widely used conjoint technique today. The marketing research community has adopted CBC enthusiastically, for several reasons. Choice tasks seem to mimic what actual buyers do more closely than ranking or rating product concepts as in conventional conjoint analysis. Choice tasks seem easy for respondents, and everyone can make choices. And equally important, multinomial logit analysis provides a well-developed statistical model for estimating respondent partworths from choice data.

However, choice tasks are less informative than tasks involving ranking or rating of product concepts. The respondent must examine the characteristics of several product concepts in a choice set, each described on several attributes, before making a choice. Yet, that choice reveals only which product was preferred, and nothing about strength of preference, or the relative ordering of the non-preferred concepts. Initially, CBC questionnaires of reasonable length offered too little information to support multinomial logit analysis at the individual level. More recently, hierarchical Bayes methods have been developed which do permit individual-level analysis, but interest has remained in ways to design choice tasks so as to provide more information.

Huber and Zwerina (1996) showed that choice tasks are more efficient (statistically) if the alternatives within a task are more nearly equal in utility, giving rise to the term “utility balance.” Such choice tasks cannot be designed without knowledge of the respondent’s utilities, which is not usually available until after the interview. This “chicken-and-egg” problem has led to several attempts at “adaptive” CBC questionnaires, where inferences from early choice tasks are used in an attempt to create greater utility balance in later choice sets. The authors have participated in three previous attempts to use adaptive principles to produce more efficient choice designs, but without consistent success, two of which were reported at previous Sawtooth Software conferences (Johnson, Huber and Bacon, 2003; Johnson, Huber, and Orme, 2004) and the third attempt reported at the joint Sawtooth Software/SKIM event in Berlin (Johnson, Orme, Huber, and Pinnell, 2005). Those first attempts relied on the assumption that respondents answered in a compensatory manner, consistent with the logit rule. We suspect that we were not more successful because respondents often use non-compensatory decision rules.

In recent years marketing researchers have become aware of potential problems with CBC questionnaires and the way respondents answer CBC questions.

- The concepts presented to respondents are often not very close to the respondent’s ideal. This can create the perception that the interview is not very focused or relevant to the respondent.
- Respondents (especially in internet panels) do choice tasks very quickly. According to Sawtooth Software’s experience with many CBC datasets, once respondents warm

up to the CBC tasks, they typically spend about 12 to 15 seconds per choice task (Johnson and Orme, 1996). For the CBC study presented in this paper, respondents spent about 18 seconds per task (on average across *all* tasks) even when considering 4 alternatives, each specified on 9 attributes. It's hard to imagine how they could evaluate four alternatives each specified on nine attributes in as short a time as 18 seconds (or fewer once warmed up). It seems overwhelmingly likely that respondents accomplish this by simplifying their procedures for making choices, possibly in a way that is not typical of how they would behave if buying a real product.

- To estimate partworths at the individual level, it is necessary for each individual to answer several choice tasks. But when a dozen or more similar choice tasks are presented to the respondent, the experience is often seen to be repetitive and boring, and it seems possible that respondents are less engaged in the process than the researcher might wish.
- If the respondent is keenly intent on a particular level of a critical attribute (a “must have” feature), there is often only one such product available per choice task. Such a respondent is left with selecting this product or “None.” And, respondents tend to avoid the “None” constant, perhaps due to “helping behavior.” Thus, for respondents intent on just a few key levels, standard minimal overlap choice tasks don't encourage them to reveal their preferences much more deeply than the few “must have” features.

Gilbride and Allenby (2004) and Hauser *et al.* (2006) used sophisticated algorithms to examine patterns of respondent answers, attempting to discover simple rules that can account for respondent choices. Both groups of authors found that respondent choices could be fit by non-compensatory models in which only a few attribute levels are taken into account.

We have done something much simpler, which also suggests that CBC respondents may make choices using simple screening rules:

For each respondent, compute a Kendall's Tau coefficient for each attribute level, to measure the relationship between presence of that attribute level and choice of an alternative.

Assume that the attribute level with highest Tau is one on which the respondent has screened concepts, and that it accounts for his/her answers for those choice tasks. Remove those choice sets from further consideration.

Repeat the process until no choice sets are left. Count the number of attribute levels that are required to account in this way for all of that respondent's choices.

We find that when choice sets are composed so as to have minimal overlap, most respondents make choices consistent with the hypothesis that they pay attention to only a few attribute levels, even when many more are included in product concepts. In a recent study with 9 attributes, 85 percent of respondents' choices could be explained entirely by assuming each respondent paid attention to the presence or absence of at most four attribute levels.

We also examined another CBC data set described in more detail below. This data set had 18 choice tasks, but respondents were given the option of choosing “None.” Respondents' answers other than “None” were as follows:

11% answered all 18 tasks by choosing 1 attribute level consistently.
34% answered by choosing at most 2 attribute levels.
80% answered by choosing at most 3 attribute levels.

Such results might lead us to conclude that CBC respondents behave in a way quite different from what we had expected, and contribute less information than we had hoped. And to make matters worse, respondents who apply consistent screening rules involving few attribute levels could easily apply those same rules to holdout choice sets. Thus, success at predicting holdout choices does not imply that respondents are providing informative and thoughtful answers to our questionnaires.

However, the meaning of these results may not be so clear as it appears. A respondent may apply a compensatory model, and yet produce results compatible with a simpler non-compensatory model. To establish this, we used a compensatory model to generate artificial responses to an 18-task CBC questionnaire and then analyzed those responses using the non-compensatory approach. We found that all answers could be accounted for by the hypothesis that the artificial respondent had paid attention to only two of 37 possible attribute levels. Thus, even if a respondent's answers can be explained by a simple non-compensatory model, we cannot be sure that his/her choice process was actually that simple.

Nonetheless, we find these results unsettling. Most CBC respondents answer more quickly than would seem possible if they were giving thoughtful responses with a compensatory model. Most of their answers can be accounted for by very simple screening rules involving few attribute levels. Combine those facts with the realization by anyone who has answered a CBC questionnaire that the experience seems repetitive and boring, and one is led to conclude there is a need for a different way of asking choice questions, with the aim of obtaining better data.

We believe CBC is an effective method that has been of genuine value to marketing researchers, but that it can be improved. And we believe the greatest need at this point is not for better models, but rather for better data.

A NEW APPROACH TO DATA COLLECTION

Like our previous papers on Adaptive CBC, the title for this paper contains the word “Adaptive.” However, this time our aim is not to design choice tasks with more statistical efficiency, but rather to acquire better data. We recognize that the respondent may employ screening rules, and we seek to recognize those rules, providing choices among products that pass such screening criteria. In this way we hope to help respondents make choices more thoughtfully, and in a way more like what they would in an actual purchase situation. Our objectives are as follows:

- Provide a more stimulating experience that will encourage more engagement in the interview than conventional CBC questionnaires.
- Mimic actual shopping experiences, which may involve non-compensatory as well as compensatory behavior.
- Screen a wide variety of product concepts, but focus on a subset of most interest to the respondent.
- Provide more information with which to estimate individual partworths than is obtainable from conventional CBC analysis.

The interview has several sections, with each section quite different from the previous (the interview is posted online at www.sawtoothsoftware.com/test/byo/byologn.htm). Throughout the interview we attempt to keep the respondent interested and engaged. The instructions appear on the screen in text, but as though they were spoken by a friendly and attractive female interviewer. Her pictures (images purchased from www.clipart.com) appear frequently at various places in the interview, from different perspectives and in different poses. She explains to the respondent that this is a simulation of a buying experience, and she gives a rationale for each interview section. For example, here is an introductory screen:



Now I'm going to present laptops to you as if you were visiting a store and I were the salesperson assisting you. We have many available configurations to choose from, and it's my job to help you find the right laptop computer.

To help you find the one that best suits you, I'd first like to ask you about the laptop you'd be most likely to purchase.

Next

BYO Section:

In the first section of the interview the respondent answers a “Build Your Own” (BYO) questionnaire to introduce the attributes and levels, as well as to let the respondent indicate the preferred level for each attribute, taking into account any corresponding feature-dependent prices¹. A typical screen for this section of the interview is shown below:

Feature	Select Feature	Cost for Feature	Help on this item
Screen Size and Weight	15 inch screen, 6 pounds \$750	\$ 750	
Brand	Dell +\$0	\$ 0	
Processor Speed (Intel Pentium)	Intel Core 2 Duo T7400 (2.16GHz) + \$300	\$ 300	?
Operating System	Vista Home Premium +\$50	\$ 50	?
Memory	1 GB +\$100	\$ 100	?
Hard Drive	100GB +\$50	\$ 50	?
Video Card	Select Feature 80GB +\$0 100GB +\$50	\$ 0	?
Battery	120GB +\$100 160GB +\$150	\$ 0	?
Office Software	Select Feature	\$ 0	?
Total:		\$ 1250	

Past research has shown that respondents enjoy BYO questionnaires and answer them rapidly, and that the resulting choices have lower error levels than repetitive choices from CBC questionnaires (Johnson, Orme, and Pinnell, 2006).

Based on answers to the BYO questionnaire, we create a pool of product concepts that includes every attribute level, but for which attribute levels are relatively concentrated around the respondent’s preferred attribute levels. Each concept in the pool is generated by altering 2, 3, or 4 attributes from the BYO-specified concept. These concepts are constructed so as to represent a nearly orthogonal design. For this study, we experimented with pools of 40 and 50 concepts.

Screening Section:

In the second section of the interview the respondent answers “screening” questions, where product concepts are shown a few at a time (we have used 5 at a time). Prices are determined by summing the costs of the features involved in the concept (per the BYO exercise) plus or minus 7% or 20%, and rounded to the nearest \$50. In the Screening Section, the respondent is not asked to make final choices, but rather just to indicate whether he/she would consider each one “a possibility.” We suggest that he/she narrow down the range of possibilities by retaining about

¹ We should note that our approach should also be able to accommodate projects for which some attributes do not involve price changes from the base product, or for projects that do not include price at all.

half of them, but the number retained is left to the respondent. A typical screen from this section of the interview is shown below:

Here are a few laptops you might like. Do any of these look like they are possibilities? It's helpful if you can keep about half of them for further consideration. But, it's up to you.

Size	14 inch screen, 5 lbs.	17 inch screen, 8 lbs.	15 inch screen, 6 lbs.	15 inch screen, 6 lbs.	15 inch screen, 6 lbs.
Brand	HP	Dell	Acer	Dell	Dell
Processor	Intel Core 2 Duo T7200 (2.00GHz)	Intel Core 2 Duo T7400 (2.16GHz)	Intel Core 2 Duo T7400 (2.16GHz)	Intel Core 2 Duo T7600 (2.33GHz)	Intel Core 2 Duo T7400 (2.16GHz)
Operating System	Vista Home Premium	Vista Ultimate	Vista Home Premium	Vista Home Basic	Vista Home Premium
Memory	1 GB	1 GB	2 GB	1 GB	1 GB
Hard Drive	100 GB	120 GB	160 GB	100 GB	100 GB
Video Card	128 MB Video card, adequate for most use	128 MB Video card, adequate for most use	128 MB Video card, adequate for most use	256 MB Video card for high-speed gaming	128 MB Video card, adequate for most use
Battery	4 hours	3 hours	3 hours	3 hours	6 hours
Productivity Software	Microsoft Office Basic (Word, Excel, Outlook)	Microsoft Office Basic (Word, Excel, Outlook)	Microsoft Office Basic (Word, Excel, Outlook)	Microsoft Office Basic (Word, Excel, Outlook)	Microsoft Works
Price	\$1,150	\$2,200	\$1,800	\$1,650	\$1,200
	<input type="radio"/> A possibility <input type="radio"/> Won't work for me	<input type="radio"/> A possibility <input type="radio"/> Won't work for me	<input type="radio"/> A possibility <input type="radio"/> Won't work for me	<input type="radio"/> A possibility <input type="radio"/> Won't work for me	<input type="radio"/> A possibility <input type="radio"/> Won't work for me

Must Haves:

After each group of concepts has been presented, we scan previous answers to see if there is any evidence that the respondent is using non-compensatory screening rules. For example, we might notice that he/she has expressed interest in only one level of some attribute, in which case we ask whether that level is an absolute requirement (a “Must Have”). Here is a typical screen for this question:

I don't want to jump to conclusions, but I've noticed that you've selected laptops with certain characteristics, shown below. If any of these describe what you **absolutely need**, it would be helpful to know.

If you'd like, please check the **one most important rule**, and I'll only show you laptops with that feature.

- At least: Intel Core 2 Duo T7200 (2.00GHz)
- At least: Vista Home Premium
- At least: 1 GB memory
- At least: 100 GB hard drive
- At least: Microsoft Office Small Business (Basic + PowerPoint, Publisher)
- None of the rules above are absolute requirements for my laptop.



Next

Past research with ACA has suggested that respondents are quick to mark many levels as unacceptable that are probably just undesirable. We considered that the same tendency might apply to “must have” rules. To avoid this possibility, we offer only cutoff rules consistent with the respondent’s previous choices and we allow the respondent to select only one cutoff rule on this screen. After each new screen of five products has been evaluated, the respondent has another opportunity to add a subsequent cutoff rule.

Unacceptables:

If the respondent has systematically avoided an attribute level, we ask whether that level would be completely unacceptable (“Unacceptables”). If the respondent identifies any “must have” or “must avoid” levels, then all further concepts shown will satisfy those requirements. The respondent has several opportunities to express such decision rules, with the result that the number of concepts actually presented to him/her is usually reduced. For this study, respondents needed to evaluate an average of 32 of the 40 product concepts (an average of 8 were automatically screened out due to confirmed decision rules).

Choice Tasks Section:

In the third section of the interview the respondent is shown a series of choice tasks presenting the surviving product concepts (those marked as “possibilities”) in groups of three, as in the screen below. In this questionnaire we asked for best and worst in each task, but it would also be possible to ask just for first choices.

Among these three, which is the best option? Which is the worst?
(I've grayed out any features that are the same, so you can just focus on the differences.)

(3 of 7)

Size	15 inch screen, 6 lbs.	15 inch screen, 6 lbs.	17 inch screen, 8 lbs.
Brand	Dell	Dell	Dell
Processor	Intel Core 2 Duo T7600 (2.33GHz)	Intel Core 2 Duo T7400 (2.16GHz)	Intel Core 2 Duo T7200 (2.00GHz)
Operating System	Vista Home Premium	Vista Home Premium	Vista Ultimate
Memory	4 GB	2 GB	1 GB
Hard Drive	100 GB	120 GB	160 GB
Video Card	128 MB Video card, adequate for most use	128 MB Video card, adequate for most use	128 MB Video card, adequate for most use
Battery	3 hours	3 hours	3 hours
Productivity Software	Microsoft Office Small Business (Basic + PowerPoint, Publisher)	Microsoft Office Small Business (Basic + PowerPoint, Publisher)	Microsoft Office Small Business (Basic + PowerPoint, Publisher)
Price	\$1,700	\$1,650	\$1,450
Best	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Worst	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next

At this point, respondents are evaluating concepts that are close to their BYO-specified product, that they consider “possibilities,” and that strictly conform to any cutoff (must have/unacceptable) rules. To facilitate information processing, we gray out any attributes that are tied across the concepts, leaving respondents to focus on the remaining differences. Any tied attributes are typically the most key factors (based on already established cutoff rules), and thus the respondent is encouraged to further discriminate among the products on the features of secondary importance.

The winning concepts from each triple then compete in subsequent rounds of the tournament until the preferred concept is identified.

Calibration Section (Optional):

The fourth section of the interview may be used to estimate a “None” parameter for the respondent. The section is introduced with this screen:



We are just about done! Now, I'd like to ask you a *different* question: **Would you actually purchase a laptop like those I've been showing you?** I'll ask you about just 5 laptops.

First, I'll show the laptop you originally configured. Next, I'll ask about other laptops you said were possibilities.

Last, I'll ask if you'd actually purchase the laptop you selected as best among those I showed you. I'm thinking you might like this last laptop best.

Next

The respondent is re-shown the concept identified in the BYO section, the concept winning the Choice Tasks tournament, and three others chosen from among those he/she has identified as worthy of consideration. We ask for each of those concepts how likely he/she would be to buy it if it were available in the market, using a standard five-point Likert scale, with a screen similar to the one below:

How likely would you be to purchase this laptop?

*** This is the original laptop you configured ***

Size	15 inch screen, 6 lbs.				
Brand	Dell				
Processor	Intel Core 2 Duo T7400 (2.16GHz)				
Operating System	Vista Home Premium				
Memory	1 GB				
Hard Drive	100 GB				
Video Card	128 MB Video card, adequate for most use				
Battery	3 hours				
Productivity Software	Microsoft Office Small Business (Basic + PowerPoint, Publisher)				
Price	\$1,550				
	Definitely Would Not	Probably Would Not	Might or Might Not	Probably Would	Definitely Would
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next

This section of the interview is used only for estimation of a partworth threshold for “None.” Partworths from other sections of the interview are used to estimate the respondent’s utility for each concept, and then a regression equation is used to produce an estimate of the utility corresponding to a scale position chosen by the researcher, such as, for example, somewhere between “Might or Might Not” and “Probably Would.” Within the market simulator, if the utility of a product concept exceeds the None utility threshold, it is chosen. (None of the simulations presented in this paper used the “None” utility threshold.)

The interview as a whole attempts to mimic the actual in-store buying experience that might be provided by an exceptionally patient and interested salesperson. For example, after the BYO section she explains that this exact product is not available but many similar ones are, which she will bring out in groups of five, to see whether each is worthy of further interest. The Choice Tasks section is presented as an attempt to isolate the specific product which will best meet the respondent’s requirements.

If the respondent has answered conscientiously, he/she will find that the final product identified by the salesperson as best is actually more preferred than the original BYO product. This occurs because the overall prices of the products generated in the product pool are varied as much as +/- 20% from the fixed BYO prices. Therefore, at least one of those (in our case, 40) product concepts will feature better features than the BYO product at the same price, the same features at a lower price, or a combination of these benefits. This makes it seem that the salesperson in our ACBC interview has actually done a good job finding a product that exceeds the quality of the BYO product and fits the needs of the respondent.

METHOD OF ANALYSIS

The data from the first three sections of the questionnaire can be analyzed with a multinomial logit model. Although the respondent was not actually completing conventional choice tasks in the first two sections of the interview, we can structure the data in synthetic choice tasks, as follows.

- The **BYO** section can be considered to produce one choice task per (non-price) attribute. Each task contains information for only a single attribute, and each alternative consists of a single level and an accompanying price.
- The **Screening** section can be considered to produce as many choice tasks as product alternatives that are screened. For each alternative, we compose a choice task that pairs the product alternative versus a constant alternative representing a threshold of acceptability.
- Suppose that c concepts are taken into the **Choice Tasks** section. Then the choice data can be arranged in $2 * c$ choice tasks, with half containing three alternatives and half containing two alternatives. If we had asked only for first choices, the number of choice tasks from this section would be c .

All of the above real (or synthetic) choice tasks can be combined² in one multinomial logit analysis³. The amount of information obtained is greater than from a typical CBC interview and may be enough information to permit estimation of individual partworths without having to “borrow” information from other respondents using HB analysis.

Sawtooth Software’s current HB algorithm assumes that respondent error is constant across the different kinds of synthetic choice tasks. There is empirical evidence that error levels are higher when more complex judgments are required. (For example, Johnson, Orme and Pinnell (2006) found that BYO data contained less error than CBC data.) Our analysis presented here assumes constant error levels in all questionnaire sections, but Thomas Otter has modeled these same data using modifications of the HB algorithm to permit varying error levels (Otter 2007). His findings confirm that the way we have used HB to estimate partworth utilities works quite well, but also suggest that perhaps even better results could be achieved with more appropriate models.

² To investigate the relative contribution of the three main ACBC sections, we omitted each section (retaining the other two sections) and measured the decrease in predictive ability (*vis-à-vis* holdouts) of the model relative to retaining all information. We found that the relative worth of the sections was, in rank order, 1) BYO, 2) Screening, and 3) Choice Tasks. (Note, in Appendix C, the screening section had the most worth in predicting holdouts in a second ACBC study.) Our procedure helped us get a rough assessment of the impact of the various sections, but we recognize Allenby *et al.*'s finding that deleting a previous section biases results based on later sections (Allenby *et al.* 2007).

³ Our approach to estimation does not treat “unacceptable” levels as “absolutely unacceptable under all conditions.” Each respondent’s data are consistent with never choosing a product concept that includes the unacceptable level. However, HB shrinks individual estimates toward population parameters, so the unacceptable utility value, while strongly negative, is not scaled so negatively that it becomes an absolute barrier to purchase irrespective of all other potential feature improvements.

AN EXPERIMENT

Early in 2007 we performed an experiment⁴ to compare this new type of adaptive CBC questionnaire (ACBC) with conventional CBC. The subject was laptop computers, described by 10 attributes with a total of 37 levels. The attributes and their levels are shown in Table 1.

Table 1
Attributes and Levels for Laptop Questionnaire

Screen Size/Weight:

14 inch screen, 5 pounds

15 inch screen, 6 pounds

17 inch screen, 8 pounds

Brand:

Acer

Dell

Toshiba

HP

Processor:

Intel Core 2 Duo T5600 (1.86GHz)

Intel Core 2 Duo T7200 (2.00GHz)

Intel Core 2 Duo T7400 (2.16GHz)

Intel Core 2 Duo T7600 (2.33GHz)

Operating System:

Vista Home Basic

Vista Home Premium

Vista Ultimate

Memory:

512 MB

1 GB

2 GB

4 GB

Hard Drive:

80 GB

100 GB

120 GB

160 GB

Video Card:

Integrated video, shares computer memory

128MB Video card, adequate for most use

256MB Video card for high-speed gaming

⁴ A few months later, we had the opportunity to field a second test of ACBC, this time as part of a real study for a client. The results of that test are reported in Appendix C.

Battery:

3 hour

4 hour

6 hour

Productivity Software:

Microsoft Works

Microsoft Office Basic (Word, Excel, Outlook)

Microsoft Office Small Business (Basic + PowerPoint, Publisher)

Microsoft Office Professional (Small Business + Access database)

Price:

\$1,000

\$1,300

\$1,700

\$2,200

\$2,800

Data were obtained from the Opinion Outpost Internet Panel. Respondents first answered a brief screener to ensure that they had at least moderate familiarity with the product category. Approximately 600 respondents were then divided randomly into two groups, with half participating in an ACBC interview, and half participating in a conventional CBC interview. Respondents in each group first received three holdout choice tasks, each consisting of four alternatives. A fourth holdout task was constructed for each respondent by combining the concepts preferred in the first three tasks.

Following the holdout tasks the ACBC respondents answered the questionnaire described above and the CBC respondents answered a conventional CBC questionnaire with 18 choice tasks, each with four alternatives plus a “None” alternative (see Appendix A for layout of CBC task).

Finally, all respondents received an identical set of questions in which they rated their interview experience on several qualitative aspects.

EXPERIMENTAL RESULTS

We observed that a few respondents in each group had unusually short interview times, and a few others in each group had very long times. We thought the fastest respondents had probably not taken the task seriously, and that the slowest ones might have been distracted and hence not given us their full effort. To minimize the possibility of including respondents of either type, we deleted the fastest 5% and the slowest 10% of each group. The partworth utility estimates and implied importances for the remainder of the respondents, (277 for CBC and 282 for ACBC) appear somewhat similar, as shown in Appendix B. However, one notes greater curvature (disutility for worst levels) for ACBC data, and more reliance on Brand to make product choices among the CBC respondents.

We recorded the interview time and also asked respondents some qualitative questions regarding their experience with the surveys. Here are the qualitative results:

Table 2
Qualitative Results Comparing ACBC with CBC

Median time to complete the CBC or ACBC sections (excluding the screener questions and post qualitative questions):

ACBC	11.6 minutes
CBC	5.4 minutes

How would you compare your overall experience with this survey compared to other internet surveys you have completed?

	ACBC	CBC
This survey was far better (5):	24%	15%
This survey was better (4):	47%	44%
This survey was about the same (3):	26%	35%
This survey was worse (2):	2%	4%
This survey was FAR worse (1):	0%	2%
Means:	3.93	3.66 (t = 4.1)

How much do you agree with the following statements about this survey?
(5=Strongly Agree, 1=Strongly Disagree; Top Box % shown beneath means.)

	ACBC	CBC	
Q1. The laptop configurations I was asked to evaluate seemed realistic.	4.4 54%	4.1 37%	(t=4.2)
Q2. This survey was at times monotonous and boring.	2.6 4%	2.8 6%	(t=2.3)
Q3. I'd be very interested in taking another survey just like this in the future.	4.3 52%	4.3 54%	(t=0.4)
Q4. The survey format made it easy for me to give realistic answers that reflect exactly what I'd do if buying a real laptop.	4.3 48%	4.1 37%	(t=2.7)
Q5. The way the laptops were presented made me want to slow down and make careful choices.	4.1 38%	3.9 27%	(t=2.9)

QUALITATIVE RESULTS:

The average ACBC interview took about twice as long as the average CBC interview. That may appear to be a disadvantage at first, but it seems less so when one realizes that CBC respondents spent an average of only 18 seconds per choice set, seemingly inadequate time to provide truly thoughtful answers.

ACBC had significantly more favorable answers than CBC on five of the six questions, despite its greater interview time. This suggests that we may have achieved our goal of providing a more stimulating experience to encourage more engagement in the interview.

HIT RATES:

Hit rates for the holdout tasks revealed interesting differences between groups. Recall that there were three holdout tasks each having four alternatives, and a final holdout task that was custom-made for each respondent, containing the winners from his/her first three holdout tasks. Prior to collecting the data, we hypothesized that ACBC would have an advantage over conventional CBC in predicting the outcome for the final holdout task (that presented the three winning concepts from the previous holdout tasks). We did not know what to expect for first three static holdout concepts.

Table 3
Holdout Hit Rates

	ACBC	CBC
First three holdouts	55.7%	57.0%
Fourth holdout	60.8%	50.0% (t = 2.54)

For the first three holdout tasks there is no significant difference, although for these samples of respondents CBC has a slight advantage. However, for the fourth holdout there is a large and significant difference in favor of ACBC.

Earlier, we presented evidence that CBC respondents may be using simplified strategies for responding to choice tasks, in which they may pay attention to only a few attribute levels. The first three holdout tasks, as well as the calibration tasks in the CBC questionnaire, were constructed with “minimal overlap,” with the alternatives in each choice set being as different from one another as possible. For example, with four brands and four alternatives in a choice set, there was always one alternative with each brand. Thus a respondent who happened to answer by always choosing a particular brand could answer every choice task consistently, and could also answer the holdout questions using the same strategy. Respondents behaving in this way could be expected to do well on the first three holdout tasks.

However, the fourth holdout task did not have this characteristic, since it was assembled from those alternatives previously preferred by each respondent. For example, if a respondent had consistently chosen a particular brand, then all three alternatives in the final holdout task would have featured that brand. If a respondent had answered the calibrating questions simply by choosing a preferred brand, his answers would contain no information with which to predict his choice in the fourth holdout. (Although, with HB estimation of partworths, the borrowing of information from other respondents would have provided some relevant information.) Thus, the fact that ACBC had a significantly better hit rate than CBC on the fourth holdout tends to confirm that some CBC respondents may have resorted to simplification of their decision processes, and that ACBC captures a greater depth of attribute processing that is more predictive of challenging choice scenarios where concepts are closer in utility and perhaps tied on key aspects.

Share Predictions:

The fourth holdout choice set was custom-made for each respondent, so it was not useful for share predictions, which require that the same choice sets be shown to many respondents. We thought it desirable to have more than three holdout choice sets for share predictions, and also thought it would be interesting to see how well our two treatment groups could predict holdout shares generated by an entirely different sample of respondents.

Accordingly, we used another group of 955 panelists who completed the same screener to assure familiarity with the product category, and who then answered 12 choice tasks (standard CBC format with 4 concepts per task, without a “None”) that were identical for all respondents. These were generated to have a modest degree of level overlap. We arbitrarily deleted the fastest 28 and the slowest 27 respondents, leaving a total of 900. These were divided into three groups of 300 on the basis of their times taken to answer the holdout questionnaire. Table 4 gives Mean Absolute Errors of share predictions for the CBC and ACBC respondents, when used to predict shares for all 900 holdout respondents, as well as each third of them based on holdout interview time. (For each prediction, we tuned the scale factor to minimize the MAE.)

Table 4
Mean Absolute Errors for Prediction of Holdout Shares

	Total Holdout Sample (n=900)	Fastest 1/3 (n=300)	Middle 1/3 (n=300)	Slowest 1/3 (n=300)
ACBC	4.52	5.24	4.72	4.42
CBC	4.49	4.88	4.95	4.98

We are not aware of good statistical tests for comparing differences in MAE for choice shares across 12 choice tasks, but the differences in Table 4 tell a consistent story. ACBC essentially matches the overall prediction accuracy of CBC, but excels in predicting the shares generated by the slower two groups of holdout respondents. In contrast, CBC has smaller prediction errors when predicting holdout shares generated by respondents who answered the holdout tasks most quickly.

These results also seem consistent with the hypothesis that some CBC respondents may use simple decision rules, such as choosing products that have a small number of critical attribute levels. It seems reasonable that holdout respondents who take longer with their choices may be using more elaborate and potentially different decision rules. To investigate this possibility, we used the Swait/Louviere test to assess whether the slow and fast responders to the 12 holdout questions differed significantly with respect to main effect parameters after controlling for scale (the design of the 12 holdout tasks supported main effects estimation). The test for difference in parameters was strongly significant, with $p < 0.001$. We also found that the slower group had a scale factor 40% larger than the faster respondent group, implying less error in their responses. Table 4 provides evidence favorable for ACBC. Despite the strong methods bias which should favor CBC in predicting CBC holdouts, ACBC can match CBC’s prediction accuracy overall. More importantly, ACBC produces better predictions of the shares generated by more thoughtful holdout respondents.

FURTHER EVIDENCE OF SIMPLIFICATION

At the 2006 Sawtooth Software Conference, Hoogerbrugge and van der Wagt (H&W) presented an interesting paper titled “How Many Choice Tasks Should We Ask” (2006). They re-analyzed a large number of CBC data sets with HB, in which they first estimated respondent partworths using only the first choice task, then again using the first two choice tasks, etc. For each re-analysis they used the estimated partworths to predict a holdout choice task, and they measured success with hit rates. They found that hit rates increased as the number of calibration choice tasks increased until about ten choice tasks, but from then on the hit rates were essentially flat. They concluded that there was often little reason to administer more than ten choice tasks to a CBC respondent. These results were surprising to many researchers, because general statistical experience has led us to expect that prediction will be better with more information. If respondents were using compensatory models to make their choices, one would expect that more information would indeed permit better predictions.

Our CBC respondents had a total of 18 calibration choice tasks plus four holdout tasks. We have duplicated the H&W analysis with our data, and we reach similar conclusions. Our hit rates increase gradually when using from 2 to 12 calibration tasks, after which there is no further systematic improvement.

If respondents are simplifying their decision processes by paying attention to only a few attribute levels, that pattern can be detected after relatively few choice tasks. Therefore, it may be that H&W have provided additional evidence that respondents are in fact simplifying their decision processes.

BENEFITS FROM INCREASED INFORMATION

We have pointed out that the ACBC interview should provide more information than a conventional CBC interview. This raises the question of whether ACBC data may be especially useful when the researcher is faced with small samples, or even for individual-level estimation.

To examine the effect of small samples, we drew 10 random samples of 25 respondents of each type and re-estimated individual partworths in each sample using HB. We reasoned that if ACBC provided more information, the estimates of population parameters should be more precise, leading in turn to better estimation of individual partworths. We used the partworths estimated from each sample to predict choice shares for the holdout respondents on the 12 holdout choice sets.

ACBC had an advantage with a mean absolute error of 6.29 share points compared to 6.91 for CBC. This difference of 0.62 share points may be compared to a corresponding difference of 0.03 share points in favor of CBC for estimates obtained when all respondents are used to estimate population parameters (see Table 4). Thus it appears that ACBC has an advantage over CBC when sample sizes are small.

It should be noted that ACBC’s superior performance occurs despite the disadvantage that the holdout responses being predicted are CBC responses. The presence of any “methods bias” would be a disadvantage for ACBC.

We have also used a simple monotone regression algorithm (Johnson, 1975) to estimate partworths. This approach makes no assumptions about error distributions. It simply seeks a set

of partworths that satisfy the inequality constraints implied by the data. Any levels that were marked as unacceptable for the respondent were given an arbitrary low partworth. Each respondent's partworths are estimated using only information from his own responses, so it provides strictly "individual-level" estimation. Table 5 provides hit rates for partworths estimated by monotone regression, compared to previously shown hit rates for HB estimates.

Table 5
ACBC Hit Rates, Including Monotone Regression Estimation

	HB Estimation		Monotone Regression
	CBC	ACBC	ACBC
First three holdouts	57.0%	55.7%	52.2%
Fourth holdout	50.0%	60.8%	57.0%

The first fact evident from Table 5 is that hit rates for monotone regression are inferior to those for HB. When even small samples are available, HB appears to be the preferred estimation method. The second fact regards the fourth holdout choice set, which was deliberately constructed so as to be difficult to predict from choice data in which respondents had paid attention to few attribute levels. Both methods for estimating partworths from ACBC seem to perform better than conventional CBC under HB estimation.

To examine ACBC's success at holdout share predictions when monotone regression is used for estimation, in Table 6 we repeat the overall results from Table 4, but with an additional column for monotone regression predictions.

Table 6
Mean Absolute Errors for Prediction of Holdout Shares,
Including Monotone Regression

	HB Estimation		Monotone Regression
	CBC	ACBC	ACBC
Mean Absolute Error	4.49	4.52	4.54

ACBC's share predictions from monotone regression are essentially as good as those from HB estimation. This is somewhat unexpected. We'd generally recommend that HB be used for estimation whenever possible (especially given an improved HB estimation approach detailed by Thomas Otter in his paper presented at this same conference), but that when strictly individual estimation is required, monotone regression can still provide useable results.

Because ACBC contains relatively more information than conventional CBC, it may provide additional benefits for segmentation research, whether via demographic variables or latent classes. There is less shrinkage to population parameters when using HB with ACBC data and correspondingly larger scale, which is beneficial when characterizing distinct preferences of segments. Additionally, monotone regression can produce partworths that are truly individual, uninfluenced by group averages.

SUMMARY AND CONCLUSIONS

Results from this study help in understanding the previously puzzling results of our earlier ACBC attempts.

- Our previous attempts assumed that respondents answered in a compensatory way consistent with the logit model, and created choice tasks so as to increase statistical efficiency as defined by that model (by increasing utility balance). But for respondents who behave non-compensatorily, utility balance is irrelevant, since such a model says that the respondent looks for presence or absence of specific attribute levels irrespective of other aspects of the products. In our second ACBC paper, where our attempt at ACBC seemed to have failed, we found that we *had* increased statistical efficiency, but without improved predictions. That could be expected for respondents behaving non-compensatorily.
- In our first three attempts, we measured success with holdout tasks that had minimal overlap. A non-compensatory respondent can easily make consistent holdout choices among alternatives with minimal overlap, where consistency may be even easier to achieve than for a compensatory respondent. Thus, prediction of holdout choices among alternatives that have minimal overlap is not necessarily a good test of success under the logit rule.

For this current research, we were motivated to investigate new ways of collecting choice data consistent with the idea that many CBC respondents simplify their decision processes, paying attention to only a few critical attributes. This is a convenient way of dealing with a task perceived as being confusing, repetitive and boring. We believed it might be possible to structure a more interesting and engaging interview which let respondents identify any “must have” or “must avoid” attribute levels, and which encouraged more thoughtful evaluation of products compatible with those requirements.

While some researchers have tried to accomplish a more thorough evaluation of attributes through partial-profile models (both ACA and partial-profile CBC), we have accomplished this while maintaining the more realistic full-profile context.

We believe that the Adaptive CBC (ACBC) method for collecting data provides several improvements over conventional CBC.

- Although the interview takes longer (11.6 minutes rather than 5.4 in our experiment), respondents appear to have found the ACBC interview more interesting and engaging than CBC, and a more faithful simulation of the buying experience.
- ACBC produces better predictions for a choice set that was custom-designed for each respondent from concepts preferred in previous choice sets. ACBC was superior to CBC when predicting choice shares from the group of holdout respondents who had taken longer to answer, and had therefore presumably been more thoughtful. In both of these cases methods bias significantly favored CBC, since the holdout tasks were CBC tasks.
- ACBC’s superiority over CBC is also particularly evident when used with small samples of respondents. ACBC also permits estimation of truly individual-level

partworths without the need to borrow information from other respondents (although they are probably not as successful as HB estimates).

Most choice researchers admit that task simplification at the individual level must exist, but many have believed that the aggregate effect of hundreds of respondents (each employing different simplification strategies) should counteract this problem and fairly accurately reflect the more careful processing of information of real-world decisions. Our results suggest that respondents who take more time to complete CBC questionnaires provide different aggregate shares, and that a data collection technique that encourages a greater depth of processing may produce more accurate share predictions. Of course, we cannot be certain that ACBC performs better at predicting real world choices than standard CBC until a more complete validation experiment involving actual purchases is available.

There are some types of CBC studies that wouldn't seem a good fit for the ACBC approach we've described here. Brand-Package-Price studies, for which conventional CBC has been very popular and quite successful, would not seem to us to benefit from this adaptive approach. However, for studies involving about five attributes or more, the adaptive procedure may offer compelling benefits.

Despite our success with this comparative study, there are ways that our approach to ACBC may be improved. For example:

- In a pilot test of the interview we created a pool of 50 concepts to be considered by the respondent in the Screening section, but in the experiment reported here we used 40. Further work is required to learn the optimal number, and how it may be related to the numbers of attributes and levels.
- Further work can be done to estimate the optimal amount to vary the attributes from the BYO concept when constructing the pool of products that each respondent evaluates.
- We showed those concepts to respondents in groups of five, which was an arbitrary decision based on screen size, legibility, and clutter. We don't know if this layout was optimal.
- In the Choice section we asked for identification of both "best" and "worst" alternatives. The information about worst alternatives was of almost no value in improving estimation, but the fact that we asked that question may have improved the quality of respondents' choices of "best."

Another significant potential source of improvement is in estimation rather than data collection. There is good reason to believe that respondents are more careful and provide better answers to the BYO section of the questionnaire than to the more repetitive and complex considerations of products profiled on many attributes simultaneously. Yet the HB algorithm we used for estimation assumes constant error levels in all parts of the questionnaire. At this same conference, Thomas Otter has presented a compelling way to deal with this problem, and his work shows that the predictive ability of ACBC can be further improved by using a specialized HB methodology that uses a better way to combine information from the three ACBC sections (Otter 2007).

REFERENCES

- Allenby, Greg, Thomas Otter, and Qing Liu (2007), "Endogeneity Bias: Fact or Fiction?" *Sawtooth Software Conference Proceedings*.
- Gilbride, Timothy and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23, 3, (Summer) 391-406.
- Hauser, John R., Ely Dahan, Michael Yee, and James Orlin (2006), "'Must Have' Aspects vs. Tradeoff Aspects in Models of Customer Decisions," *Sawtooth Software Conference Proceedings*, 169-181.
- Hoogerbrugge, Marco and Kees van der Wagt (2006), "How Many Choice Tasks Should We Ask?," *Sawtooth Software Conference Proceedings*, 97-109.
- Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33 (August) 307-317.
- Johnson, Richard M. (1975), "A Simple Method for Pairwise Monotone Regression," *Psychometrika*, 40, 163-168.
- Johnson, Richard M. and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" downloaded from www.sawtoothsoftware.com/techpap.shtml.
- Johnson, Richard M., Bryan Orme, and Jon Pinnell (2006), "Simulating Market Preference with 'Build Your Own' Data," *Sawtooth Software Conference Proceedings*, 239-253.
- Johnson, Richard M., Bryan Orme, Joel Huber, and Jon Pinnell (2005), "Testing Adaptive Choice-Based Conjoint Designs," *Design and Innovations Conference (Berlin)*, available at www.sawtoothsoftware.com/techpap.shtml.
- Johnson, Richard M, Joel Huber, and Bryan Orme (2004), "A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs)" *Sawtooth Software Conference Proceedings*, 217-234.
- Johnson, Richard M., Joel Huber, and Lynd Bacon (2003), "Adaptive Choice-Based Conjoint," *Sawtooth Software Conference Proceedings*, 333-343.
- Otter, Thomas (2007), "Hierarchical Bayesian Analysis for Multi-Format Adaptive CBC," *Sawtooth Software Conference Proceedings*.

APPENDIX A

CBC TASK LAYOUT

	If these were your only options, which laptop would you choose? Choose by clicking one of the buttons below:			
	14 inch screen, 5 pounds	17 inch screen, 8 pounds	15 inch screen, 6 pounds	17 inch screen, 8 pounds
Screen Size/Weight:				
Brand:	Toshiba	Acer	HP	Dell
Processor:	Intel Core 2 Duo T7600 (2.33GHz)	Intel Core 2 Duo T7200 (2.00GHz)	Intel Core 2 Duo T7400 (2.16GHz)	Intel Core 2 Duo T5600 (1.86GHz)
Operating System:	Vista Home Basic	Vista Ultimate	Vista Home Premium	Vista Ultimate
Memory:	2 GB	4 GB	512 MB	1 GB
Hard Drive:	100 GB	120 GB	160 GB	80 GB
Video Card:	128MB Video card, adequate for most use	Integrated video, shares computer memory	256MB Video card for high-speed gaming	Integrated video, shares computer memory
Battery:	6 hour	3 hour	4 hour	3 hour
Productivity Software:	Microsoft Works	Microsoft Office Basic (Word, Excel, Outlook)	Microsoft Office Professional (Small Bus + Access database)	Microsoft Office Small Business (Basic + PowerPoint, Publisher)
Price:	\$1,700	\$1,000	\$2,800	\$1,300
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
				NONE: I wouldn't choose any of these.

APPENDIX B

AVERAGE PARTWORTHS (NORMALIZED)

	ACBC n=282	CBC n=277
14 Inch, 5 pounds	-24.38	-16.47
15 Inch, 6 pounds	8.12	-0.50
17 Inch, 8 pounds	16.26	16.97
Acer	-25.81	-35.69
Dell	24.59	24.55
Toshiba	-3.44	-5.71
HP	4.65	16.85
1.86GHz Processor	-35.66	-11.93
2.00GHz Processor	2.36	-2.08
2.16GHz Processor	13.26	0.20
2.33GHz Processor	20.04	13.81
Vista Basic	-9.67	-4.30
Vista Premium	6.05	-2.14
Vista Ultimate	3.62	6.44
512MB RAM	-90.69	-89.30
1GB RAM	-11.21	-9.26
2GB RAM	40.94	35.69
4GB RAM	60.96	62.87
80GB Hard Drive	-48.07	-28.40
100GB Hard Drive	-3.18	0.39
120GB Hard Drive	18.22	4.20
160GB Hard Drive	33.02	23.81
Integrated Video	-37.88	-23.58
128MB Card	10.86	0.46
256MB Card	27.02	23.12
3 hour battery	-26.78	-19.32
4 hour battery	7.91	-2.02
6 hour battery	18.87	21.35
MS Works	-39.80	-34.31
MS Office Basic	12.58	7.72
MS Office Small Bus	19.44	13.33
MS Office Professional	7.78	13.25
Price	-101.29	-97.39

AVERAGE IMPORTANCES

	ACBC n=282	CBC n=277
Size/Weight	8.25	8.25
Brand	8.10	13.90
Processor	7.67	5.74
OS	4.84	4.46
RAM	16.68	17.14
Hard Drive	9.87	7.47
Video Card	8.96	7.57
Battery	6.49	5.92
Software	9.79	8.72
Price	19.35	20.83

APPENDIX C

A SECOND EXPERIMENT TO TEST ACBC

A few months after completing the first test of ACBC as reported in this paper, we used ACBC in a real client study of a mechanical product for recreational equipment. We are grateful to Joe Curry of Sawtooth Technologies for sponsoring this project. Because this was an actual client project, we were not able to design the experiment as rigorously as our first test of ACBC (e.g. the ACBC respondents were collected a few weeks after the CBC respondents, with minor deviations in the method of recruitment). For this second test the questionnaire did not include graphics representing an interviewer, which we believe would have made a positive contribution. Also, we weren't able to collect a separate sample of holdout respondents. Despite these differences, the findings are quite similar to those of the first test.

This second study involved the following characteristics for the ACBC interview:

- 8 attributes (29 total attribute levels) plus price
- 36 products used in the Screening Section, shown in triples
- The Choice Tasks section asked just first choice from triples
- No graphic representing an interviewer was shown

Approximately 500 respondents completed a standard CBC survey and 400 completed the ACBC survey. The CBC survey involved 14 choice tasks shown in pairs. Four CBC-looking holdout tasks were included for all respondents, with the final holdout composed of the three concepts chosen in the earlier three fixed holdout tasks.

Qualitative Findings:

The ACBC respondents spent about triple the time doing the conjoint section of their questionnaire as respondents who completed the more abbreviated standard CBC interview (about 15 minutes compared to 5 minutes). Approximately 6% of the ACBC respondents dropped out of their survey during the conjoint questions compared to 1% of CBC respondents. The ACBC respondents reported that their survey was more monotonous than CBC respondents did (this is opposite what we found with the first laptop study), but both groups reported equal interest in taking another survey like theirs in the future. In addition, the ACBC respondents reported that the products they were shown were more realistic (confirming findings of the laptop study).

Quantitative Findings:

The aggregate utilities were correlated 0.91 between ACBC and CBC respondents. Attribute importances were very similar, with one attribute (warranty) appearing significantly more important for CBC respondents. We did not note any enhanced "curvature" (loss avoidance) for the worst levels from ACBC utilities compared to the CBC utilities (in contrast to what we observed with the laptop study). The hit rates for the fixed CBC-like holdout tasks favored CBC slightly but not significantly. Hit rates for the customized holdout CBC choice task differed more strongly, and in favor of ACBC: 62.2 vs. 59.5 (difference not significant). Again, methods

bias was strongly in favor of CBC in terms of ability to predict this holdout, as the choice task was a CBC task.

In contrast to the first ACBC study where we found that the BYO section was the most valuable of the three sections (BYO, Screening Section, Choice Tasks), this time it was least valuable. The rank-order of contribution toward predicting holdouts was 1) Screening Section, 2) Choice Tasks, 3) BYO. We think we have an explanation for this discrepancy. The BYO section focuses on the tradeoff between each feature and price. In this second study, price overall was not very important relative to the other attributes. Therefore, BYO's focused effort on estimating price sensitivity didn't pay off as well here (in terms of predicting holdout choices) as with the laptop study where price carried much more importance.

DISCUSSION:

It is impressive that ACBC again beats CBC in predicting the customized (and difficult) CBC-looking task, despite the methods bias in favor of CBC and the fact that ACBC utilities are not equivalent to those of CBC. Of course, the best test of validity would involve actual purchases, for which we do not have data. Respondents described the ACBC interview as more monotonous compared to CBC (opposite our findings from the laptop study), and we wonder whether not showing a graphic of an engaging facilitator made a difference, or if it is principally explained by the greater relative difference in task length for ACBC relative to CBC in this study.

HB-ANALYSIS FOR MULTI-FORMAT ADAPTIVE CBC

*THOMAS OTTER
GOETHE UNIVERSITY*

1. INTRODUCTION

Johnson and Orme (this volume) propose a new interview technique, ACBC, to collect preference information (experimental choice data). ACBC combines different elicitation formats and the idea of adaptive designs. In their paper, Johnson and Orme compare ACBC to standard CBC. The purpose of this paper is to discuss and compare models for the analysis of data collected using ACBC. The paper is organized as follows: Section 2 briefly reviews the procedure and discusses implications for modeling. Section 3 summarizes the models compared and illustrates MCMC estimation. Section 4 describes the data and Section 5 presents the results of the comparison. Finally, I provide a discussion of the results and avenues for future research.

2. THE INTERVIEW PROCEDURE

The interview starts with a “Build Your Own” (BYO) questionnaire. In this part of the interview the respondent identifies, for each attribute, a preferred combination of a specific level and an associated price.¹ Thus, the respondent reveals a point in the space of all possible combinations that is attractive to him.

ACBC takes this point in the design space as a seed for creating a fixed number of product concepts ‘centered’ on the response from BYO. Product concepts are created such that all attribute levels are shown, but only varying subsets of attributes take levels different from the BYO response. Thus, each of the generated concepts shares at least some attribute levels with the response from BYO. Johnson and Orme have experimented with pools of 40 and 50 concepts.

In the SCREENER section of the interview, the respondent then is asked to identify for each concept whether it is a possibility or if it won’t work for him. Because of space constraints concepts have to be shown in subsets. The procedure keeps track of patterns in the responses that are consistent with attribute-based screening rules (Gilbride and Allenby 2004).

If such a pattern is observed, the respondent is offered a choice among screening rules that are consistent with the data so far, before the next subset is presented. The set of rules includes the option that no screening rule applies. From any rules question, the respondent can only select one rule. However, multiple rules questions may be asked throughout the interview. If a rule other than ‘no screening is taking place’ applies, the concepts left in the pool which are rejected by rules are marked as rejected, without ever being presented to the respondent.

Subsequently, the collection of concepts that were identified as possibilities enter a single-elimination play-off choice based conjoint (CBC). First, three randomly chosen concepts from this set are presented. The chosen concept is then put against two different randomly chosen

¹ Johnson and Orme mention the possibility to handle situations where it would be hard or impossible to a priori design pairs of attribute levels and prices. However, this issue is beyond the scope of this paper.

concepts from the remainder in the pool of accepted alternatives, until the overall winner is determined.

3. IMPLICATIONS FOR MODELING

The adaptive nature of ACBC is best illustrated in the form of a directed acyclic graph (Figure 1).

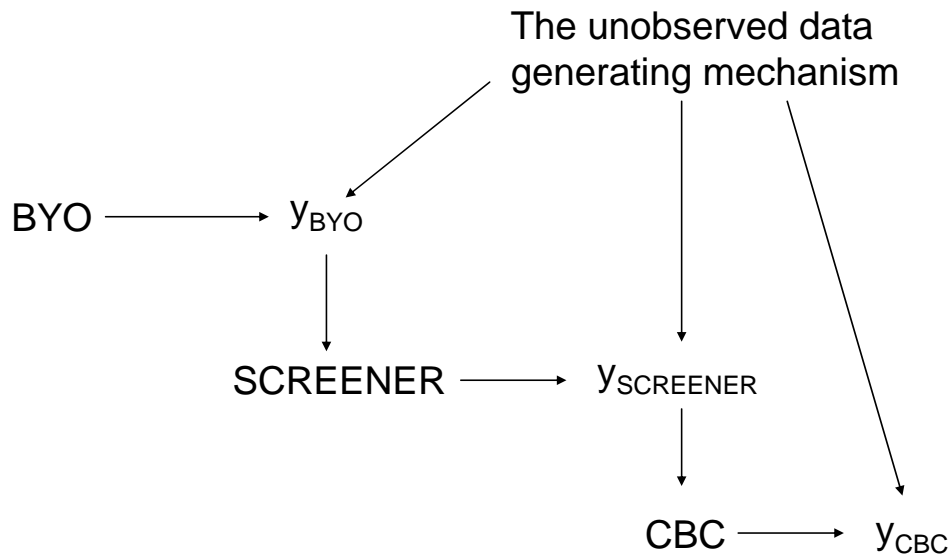


Figure 1
A Graph for ACBC

The data, y_{BYO} , $y_{SCREENER}$, and y_{CBC} obtained from one respondent are a priori dependent because they are assumed to jointly provide information about this respondent's unobserved preferences. Through the data, parts of the design, i.e. the concepts used in the SCREENER section and in the final CBC section are linked to respondents' unobserved preferences, too. However, Liu, Otter, and Allenby (2007 and this volume) show that, conditional on the data and the model, adaptive designs can and should be analyzed simply ignoring the fact that an adaptive design was used. This result has important implications for HB-analysis of ACBC data.

First, later parts of the interview, say CBC, cannot be analyzed in isolation using standard methods. Omitting earlier sections from the analysis removes, among other things, the data $y_{SCREENER}$ and thus directly connects what concepts are shown in the CBC part to a respondent's unobserved preferences. When the specific concepts shown can no longer be treated as a function of observed data, but have to be expressed as a function of unobserved preferences, standard hierarchical analysis is misspecified and may result in misleading inferences. Second, for the joint analysis of BYO, SCREENER, and CBC we need a sensible model to extract the joint information about respondents' unobserved preferences.

3.1 MODELS

I first discuss the formulation of full conditional likelihood functions for the different sections of the interview. The term full conditional likelihood refers to the likelihood specification at the level of the individual respondents, before introducing respondent heterogeneity.

BYO – full conditional likelihood

In ACBC, the BYO responses are modeled as, conditional on preferences, independent multinomial choices among attribute level-price pairs within attributes. With linear price, this specification is equivalent to the probability of choosing the concept combining all chosen attribute levels at the sum of their respective prices from a set of concepts. This set of concepts comprises all combinations of different attribute level-price pairs presented. The following algebra illustrates this relationship using a two-attribute example:

$$\frac{\exp(\alpha_i - p(\alpha_i)) \exp(\beta_j - p(\beta_j))}{\sum_{k=1}^{K(\alpha)} \exp(\alpha_k - p(\alpha_k)) \sum_{k=1}^{K(\beta)} \exp(\beta_k - p(\beta_k))} = \frac{\exp(\alpha_i + \beta_j - p(\alpha_i) - p(\beta_j))}{\exp(\alpha_1 + \beta_1 - p(\alpha_1) - p(\beta_1)) + \exp(\alpha_1 + \beta_2 - p(\alpha_1) - p(\beta_2)) + \dots + \exp(\alpha_{K(\alpha)} + \beta_{K(\beta)} - p(\alpha_{K(\alpha)}) - p(\beta_{K(\beta)}))} = \frac{V(\text{build})}{\sum_{j=1}^{K(\alpha)K(\beta)} V(j)}$$

Attribute one has $K(\alpha)$ and attribute two $K(\beta)$ levels with part-worths $\alpha_{.1}, \dots, \alpha_{.K(\alpha)}$ and $\beta_1, \dots, \beta_{K(\beta)}$, respectively. The price associated with level i of the first attribute, for example, is $p(\alpha_i)$. V is short for the deterministic utility².

SCREENER – full conditional likelihood

The likelihood specification for this section poses interesting challenges. If respondents essentially deterministically sort the concept pool into a set of accepted and a set of rejected alternatives, then one way to model the data is to enforce ordinal utility constraints between the two sets coupled with a deterministic decision rule, i.e. no error terms. If error terms are introduced, one has to think about how these are generated.

The assumption that each alternative is evaluated once leads to the *multichoice* model discussed by van Ophem *et al.* (1999). This model derives the likelihood for the two sets of concepts as the probability that the (unobserved) worst alternative in the accepted set is associated with a larger utility draw than the (unobserved) best alternative in the rejected set. For

² Economists refer to this more correctly as the indirect utility. To be consistent with most of marketing I will simply use ‘utility’ throughout the paper.

this model to make sense behaviorally, we have to assume that respondents somehow keep track of random utility draws across different computer screens.

The assumption that each alternative is evaluated once and compared to a fixed outside option, which itself is repeatedly evaluated each time, leads to a *binary logit model*. In this model respondents are only required to keep track of what they consider the standard of comparison.

Finally, the assumption that each accepted alternative is evaluated once but individually compared to repeated evaluations of all rejected alternatives leads to a *multinomial logit model*. This formulation is based on imaginary choice sets comprised of all rejected alternatives and one of the accepted alternatives in turn. It thus lacks a plausible behavioral motivation as it implies that e.g. even the first accepted alternative is compared to all rejected alternatives that may not even have been presented at this point. However, it is consistent with the application of rules that clearly differentiate between each accepted and all rejected alternatives.

The discussion so far has ignored the rejection of alternatives by rules. Consider an observed response to a rules question (see Section 2) causing multiple alternatives left in the pool to be jointly marked as rejected. In this case, associating these alternatives with independent error terms is at odds with how the data are generated. A complete characterization of the data generating process requires modeling the choice among screening rules offered by the rules question.

In effect, the response to the rules question reveals something about the certainty associated with the observed rejections as a function of the attribute levels. A respondent who rejects a particular alternative because of an unacceptable attribute or the equivalent lack of a must-have attribute is likely to behave almost deterministically given the rule. This reasoning suggests modeling some of the data in the SCREENER section as close to deterministic, given preferences for attribute levels and information about rules that apply (see Gilbride and Allenby 2004).

In this paper, I only offer a simplified attempt in this direction. I investigate task-specific differences in the amount of error associated with data from different parts of the interview. To the extent that decisions made in the SCREENER part are governed by rules, the distinction between accepted and rejected alternatives is close to deterministic, both before the rules question is asked and, of course, after the computer complies with the revealed rule.

CBC - full conditional likelihood

I use the standard multinomial logit model for the CBC part of the interview.

Heterogeneity

It is well known that a simple compensatory model can approximate many screening rules, e.g. unacceptable and must-have attribute levels, by setting the respective coefficients to extreme values. In the absence of a prior, the standard compensatory model is thus very hard to distinguish from models with screening rules. A hierarchical model, however, introduces the distribution of heterogeneity as a prior for parameters in the full conditional, i.e. individual level, likelihood functions. It is this (prior) distribution that effectively prevents a compensatory likelihood function from approximating decisions based on (heterogeneous) screening rules to an arbitrary degree.

In the case of ACBC, any information extracted from all three parts of the interview pooled across respondents implies the transfer of information from one person's SCREENER data to *another* person's CBC data. This is because what is unacceptable or a must-have is very likely to differ across respondents. Thus, a model that suitably connects all three parts of the interview is essential for pooling across respondents to make sense.

Models assuming prior independence between the screening of alternatives and the preferences revealed through compensatory processing (e.g. Gilbride and Allenby 2004) are special cases. These models imply that data known to be generated by screening rules can be analyzed independently from data known to be generated by compensatory processing, even in the context of adaptive designs.

3.2 ESTIMATION

I first show how to estimate an HB model with task-specific scale factors and then discuss Bayesian estimation of the multichoice model (van Ophem *et al.* 1999) using data augmentation.

Bayesian estimation of an HB model with task-specific scale factors

Consider three different parts of an interview that are expected to differ with respect to their scale only, with full conditional likelihood:

$$p\left(\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{y}_i^{(3)} \mid \boldsymbol{\beta}_i^*, \gamma^{(1)*}, \gamma^{(2)*}, \gamma^{(3)*}\right) = \prod_{t=1}^{T_1} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(1)})} \boldsymbol{\beta}_i^* \gamma^{(1)*}\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i^* \gamma^{(1)*}\right)} \prod_{t=1}^{T_2} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(2)})} \boldsymbol{\beta}_i^* \gamma^{(2)*}\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i^* \gamma^{(2)*}\right)} \prod_{t=1}^{T_3} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(3)})} \boldsymbol{\beta}_i^* \gamma^{(3)*}\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i^* \gamma^{(3)*}\right)} \quad (1)$$

Here, $\gamma^{(1)*}, \gamma^{(2)*}, \gamma^{(3)*}$ are task-specific scale factors and $\boldsymbol{\beta}_i^*$ collects respondent i 's part-worths. It is well known that all task-specific scale factors and $\boldsymbol{\beta}_i^*$ are not jointly identified in equation (1). A Bayesian solution to this problem is to specify proper priors for all parameters and to project down onto what is identified in equation (1), post-processing the MCMC draws. The following transformations identify the scales relative to, for example, the first task:

$$\boldsymbol{\beta}_i = \gamma^{(1)*} \boldsymbol{\beta}_i^*, \quad \gamma^{(2)} = \frac{\gamma^{(2)*}}{\gamma^{(1)*}}, \quad \gamma^{(3)} = \frac{\gamma^{(3)*}}{\gamma^{(1)*}} \quad (2)$$

Substituting in equation (1) we obtain the (likelihood) identified expression:

$$p\left(\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{y}_i^{(3)} \mid \boldsymbol{\beta}_i, \gamma^{(2)}, \gamma^{(3)}\right) = \prod_{t=1}^{T_1} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(1)})} \boldsymbol{\beta}_i\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i\right)} \prod_{t=1}^{T_2} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(2)})} \boldsymbol{\beta}_i \gamma^{(2)}\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i \gamma^{(2)}\right)} \prod_{t=1}^{T_3} \frac{\exp\left(\mathbf{x}'_{(y_{i,t}^{(3)})} \boldsymbol{\beta}_i \gamma^{(3)}\right)}{\sum_{j=1}^{J_t} \exp\left(\mathbf{x}'_j \boldsymbol{\beta}_i \gamma^{(3)}\right)} \quad (3)$$

For the purposes of MCMC based inference, equation (1) is preferable because it is likely to result in a more efficient MCMC sampler. A more efficient MCMC sampler delivers more information about the posterior than a less efficient MCMC sampler holding the number of draws from the posterior constant.

I use an MCMC sampler based on the following conditional distributions:

$$\begin{aligned}
 (1) \quad & p(\boldsymbol{\beta}_i^* | \bullet) \propto p(\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{y}_i^{(3)} | \boldsymbol{\beta}_i^*, \gamma^{(1)*}, \gamma^{(2)*}, \gamma^{(3)*}) N(\boldsymbol{\beta}_i^* | \boldsymbol{\alpha}, \mathbf{Q}) \\
 (3) \quad & p(\gamma^{(1)*}, \gamma^{(2)*}, \gamma^{(3)*} | \bullet) \propto \prod_{i=1}^N p(\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \mathbf{y}_i^{(3)} | \boldsymbol{\beta}_i^*, \gamma^{(1)*}, \gamma^{(2)*}, \gamma^{(3)*}) N(\ln \gamma^{(1)*}, \ln \gamma^{(2)*}, \ln \gamma^{(3)*} | \Gamma, \mathbf{V}) \\
 (4) \quad & p(\boldsymbol{\alpha} | \bullet) \propto \prod_{i=1}^N N(\boldsymbol{\beta}_i^* | \boldsymbol{\alpha}, \mathbf{Q}) N(\boldsymbol{\alpha} | \mathbf{0}, \mathbf{A}) \\
 (5) \quad & p(\mathbf{Q} | \bullet) \propto \prod_{i=1}^N N(\boldsymbol{\beta}_i^* | \boldsymbol{\alpha}, \mathbf{Q}) IW(\mathbf{Q} | \nu^0, \mathbf{S}^0)
 \end{aligned}$$

In blocks (1) and (2) I use a simple random walk Metropolis sampler. Steps (4) and (5) use standard conjugate results. I show the building blocks of steps (4) and (5) above to highlight the connection between the subjective prior for the scale factors and the subjective priors for the distribution of heterogeneity. If a priori very large scale factors are a possibility, this translates into a priori large amounts of heterogeneity, even if \mathbf{Q} , the variance-covariance of $\boldsymbol{\beta}_i^*$, is a priori small. Another way to see this is through the heterogeneity distribution of the (likelihood) identified quantity $\boldsymbol{\beta}_i$ in equation (2), $N(\boldsymbol{\beta}_i | \bar{\boldsymbol{\beta}}, \mathbf{V}_\beta)$, obtained by post-processing from

$$\begin{aligned}
 \bar{\boldsymbol{\beta}} &= \gamma^{(1)*} \boldsymbol{\alpha} \\
 \mathbf{V}_\beta &= (\gamma^{(1)*})^2 \mathbf{Q}
 \end{aligned} \tag{4}$$

Multichoice - Bayesian estimation

The multichoice model discussed by van Ophem *et al.* (1999) derives the likelihood for a set of rejected concepts and a set of accepted concepts as the probability that the (unobserved) worst alternative in the accepted set is associated with a larger utility draw than the (unobserved) best alternative in the rejected set.

With extreme value distributed errors the distribution of the maximum utility of the rejected alternatives is again extreme value. And in the special case of only *one* accepted alternative we are back to the simple multinomial logit model. With more than one accepted alternative, what is needed is the distribution of the minimum of draws from extreme value distributions. This is a hard problem. I use data augmentation to sidestep an explicit solution.

At each iteration of the MCMC, I generate extreme value distributed utilities for the accepted alternatives. These utilities imply a ranking among the accepted alternatives. Conditional on this augmented ranking, the likelihood is easily evaluated using the ‘exploded’ logit formulation introduced into marketing by Chapman and Staelin (1982). The full conditional distributions are

$$(1) p(\mathbf{z}_i^A | \boldsymbol{\beta}_i, \mathbf{y}_i^A) = EV(\mathbf{z}_i^A | \mathbf{x}_i^A \boldsymbol{\beta}_i, 1)$$

$$(2) p(\boldsymbol{\beta}_i | \bullet) \propto \text{ExplodedLogit}(\{\text{rank}(\mathbf{z}_i^A), \mathbf{y}_i^R\} | \boldsymbol{\beta}_i) N(\boldsymbol{\beta}_i | \bar{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$$

$$(3) p(\bar{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}} | \bullet) \propto \prod_{i=1}^N N(\boldsymbol{\beta}_i | \bar{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}) p(\bar{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$$

In block (1) a latent utility z for each accepted alternative is generated directly from an extreme value distribution with location $\mathbf{x}'\boldsymbol{\beta}_i$ and scale 1. Block (2) generates the part-worth vector $\boldsymbol{\beta}_i$ using a simple random walk Metropolis step. The likelihood in this step is exploded logit. The last term in this likelihood is equal to the probability of choosing the worst of the accepted alternatives from a set with all rejected alternatives, \mathbf{y}_i^R . Conditional on the augmented utilities for the accepted alternatives \mathbf{z}_i^A , we know which of the accepted alternatives is worst.

4. THE DATA

Johnson and Orme (this volume) give a complete description of the data. I repeat the essential details. They used ACBC and CBC to study preferences for laptop computers, described by 10 attributes with a total of 37 levels. The attributes and their levels are shown in Table 1.

Table 1
Attributes and Levels for Laptop Questionnaire

Screen Size/Weight:

14 inch screen, 5 pounds
15 inch screen, 6 pounds
17 inch screen, 8 pounds

Brand:

Acer
Dell
Toshiba
HP

Processor:

Intel Core 2 Duo T5600 (1.86GHz)
Intel Core 2 Duo T7200 (2.00GHz)
Intel Core 2 Duo T7400 (2.16GHz)
Intel Core 2 Duo T7600 (2.33GHz)

Operating System:

Vista Home Basic
Vista Home Premium
Vista Ultimate

Memory:

512 MB
1 GB
2 GB
4 GB

Hard Drive:

80 GB
100 GB
120 GB
160 GB

Video Card:

Integrated video, shares computer memory
128MB Video card, adequate for most use
256MB Video card for high-speed gaming

Battery:

3 hour
4 hour
6 hour

Productivity Software:

Microsoft Works
Microsoft Office Basic (Word, Excel, Outlook)
Microsoft Office Small Business (Basic + PowerPoint, Publisher)
Microsoft Office Professional (Small Business + Access database)

Price:

\$1,000
\$1,300
\$1,700
\$2,200
\$2,800

Data were obtained from the Opinion Outpost Internet Panel. Respondents first answered a brief screener to ensure that they had at least moderate familiarity with the product category. Respondents were randomly assigned to interview techniques resulting in 326 completed ACBC interviews. Before the actual interview, respondents received three holdout choice tasks, each consisting of four alternatives. A fourth holdout task was constructed for each respondent by combining the concepts preferred in the first three tasks.

Accordingly, Johnson and Orme used another group of 955 panelists who completed the same screener to assure familiarity with the product category, and who then answered 12 choice tasks (standard CBC format with 4 concepts per task, without a “None”) that were identical for all respondents. They arbitrarily deleted the fastest 28 and the slowest 27 respondents, leaving a total of 900.

5. RESULTS

In the following I compare the different models along their predictive performance. The data allow for predictive testing using holdout responses obtained from respondents in the estimation data set and 900 different respondents that were not used for estimation.

Predicting holdout responses

I report hit-rates, hit-probabilities, and predictive densities. Hit-rates summarize how often the alternative with the highest predicted probability is actually chosen. Posterior uncertainty in β translates into a posterior distribution of the probability of choosing a particular alternative. This probability is a non-linear transformation of β . Therefore, the probability computed at the posterior expectation of β is different from the posterior expectation of the probability. In practice it seems that reliance on posterior expectations of β is still common for data management reasons. I report both hit-rates for a comparison.

Hit-probabilities are the expected probability of making the observed choice. Both hit-rates and hit-probabilities are ad hoc summaries of predictive performance but are intuitively close enough to decision problems to be useful in practice.

Predictive densities provide a statistically more efficient summary of predictive performance. In our case, predictive densities can be interpreted as the probability of the entire holdout data set given the model. I report, on the log scale, a quasi joint predictive density. It is defined as the product of expected probabilities for all observed holdout choices. Ratios of predictive densities can be interpreted as Bayes factors, i.e. the posterior odds of one model over another, assuming equal prior model probabilities.

Each of the following tables summarizes one measure of predictive performance for all models investigated. Each model was run with three different subjective prior settings for sensitivity analysis. Specifically, I varied the prior expectation for V_{β} , i.e. the variance-covariance matrix of β , between 2, 6, and 10 holding the prior degrees of freedom constant at 5.³ As subjective prior parameters for β -bar I used a mean of zero and variance equal to 100.

Tables 2 and 3 summarize hit-rates computed from expected part-worths, i.e. the wrong way, and from expected probabilities. Much to my surprise, the results are quite comparable. However, the result is consistent with posterior uncertainty in β_i on a lower dimensional

³ Five degrees of freedom added to what is required by the dimensionality of the problem for the inverted Wishart prior to be proper.

subspace, such that different draws from the posterior are essentially consistent with respect to the maximal alternative. Across different subjective prior settings and models, the results suggest that allowing for more prior heterogeneity slightly improves hit-rates. Also, the MNL model with task specific scales (*scaled MNL*) seems to be doing slightly worse than the other models.

Table 2
Hit-rates computed from expected part-worths

	Prior V_{β}	1st	2nd	3rd	custom	overall
<i>Multichoice</i>	2	0.53	0.54	0.57	0.61	0.56
	6	0.54	0.55	0.57	0.60	0.56
	10	0.54	0.54	0.58	0.61	0.57
<i>Binary</i>	2	0.55	0.54	0.57	0.57	0.56
	6	0.57	0.55	0.56	0.60	0.57
	10	0.57	0.54	0.57	0.60	0.57
<i>MNL</i>	2	0.52	0.53	0.58	0.60	0.56
	6	0.52	0.52	0.60	0.60	0.56
	10	0.55	0.53	0.59	0.60	0.57
<i>Scaled MNL</i>	2	0.53	0.53	0.58	0.58	0.55
	6	0.52	0.52	0.58	0.58	0.55
	10	0.52	0.53	0.58	0.60	0.56

Table 3
Hit-rates computed from expected probabilities

	Prior V_{β}	1st	2nd	3rd	custom	overall
<i>Multichoice</i>	2	0.53	0.54	0.56	0.61	0.56
	6	0.54	0.55	0.57	0.60	0.56
	10	0.54	0.54	0.58	0.61	0.57
<i>Binary</i>	2	0.55	0.54	0.56	0.57	0.55
	6	0.57	0.55	0.56	0.59	0.57
	10	0.56	0.55	0.56	0.60	0.57
<i>MNL</i>	2	0.52	0.54	0.58	0.61	0.56
	6	0.52	0.53	0.60	0.59	0.56
	10	0.54	0.52	0.60	0.60	0.57
<i>Scaled MNL</i>	2	0.52	0.53	0.58	0.58	0.55
	6	0.52	0.52	0.58	0.58	0.55
	10	0.51	0.52	0.58	0.59	0.55

Hit-probabilities reported in Table 4 again do not really distinguish among the specifications investigated. Once more, the results suggest that more prior heterogeneity is better. In terms of

hit-probabilities the simple MNL model that puts accepted alternatives in the SCREENER against all rejected alternatives has a slight advantage.

Table 4
Hit-probabilities

	Prior V_{β}	1st	2nd	3rd	custom	overall
<i>Multichoice</i>	2	0.49	0.48	0.52	0.51	0.50
	6	0.50	0.49	0.53	0.52	0.51
	10	0.50	0.49	0.53	0.52	0.51
<i>Binary</i>	2	0.49	0.48	0.52	0.50	0.50
	6	0.50	0.49	0.52	0.51	0.51
	10	0.50	0.49	0.53	0.52	0.51
<i>MNL</i>	2	0.50	0.50	0.54	0.53	0.52
	6	0.51	0.50	0.55	0.54	0.52
	10	0.51	0.50	0.55	0.54	0.52
<i>Scaled MNL</i>	2	0.49	0.49	0.53	0.52	0.51
	6	0.49	0.49	0.54	0.53	0.51
	10	0.49	0.49	0.54	0.53	0.51

Table 5
Log - probability of holdout responses given the model

	Prior V_{β}	1st	2nd	3rd	custom	overall
<i>Multichoice</i>	2	-306	-320	-352	-279	-1,257
	6	-326	-335	-369	-290	-1,320
	10	-333	-346	-380	-298	-1,358
<i>Binary</i>	2	-285	-301	-333	-266	-1,185
	6	-298	-313	-351	-276	-1,238
	10	-309	-323	-361	-282	-1,275
<i>MNL</i>	2	-403	-413	-468	-343	-1,627
	6	-407	-424	-475	-351	-1,657
	10	-417	-441	-482	-360	-1,701
<i>Scaled MNL</i>	2	-380	-389	-433	-321	-1,524
	6	-390	-409	-444	-334	-1,577
	10	-396	-421	-446	-342	-1,606

The probabilities of the holdout responses given the models, reported in Table 5 (on the log scale) identify *Binary*, i.e. the specification that treats observations in the SCREENER section as independent comparisons to an outside alternative, as the overall winner. The second best specification is *Multichoice*. *MNL* and *scaled MNL* are clearly worse. In all cases, less prior heterogeneity results in better predictions than more prior heterogeneity. With equal prior model probabilities the posterior probability of *Binary* among the models investigated is one. In fact,

MNL and scaled MNL perform worse than the base case of chance predictions with a log-probability equal to $(3 \times \ln(1/4) + \ln(1/3)) \times 276 = -1,467$.⁴

Predicting holdout respondents

900 holdout respondents each evaluated the same 12 choice tasks with four alternatives. Thus I have 12 x 4 shares to predict. Assuming that the holdout respondents come from the same population as the calibration sample, the heterogeneity distribution connects the two samples.

Predicting shares from the heterogeneity distribution requires simulation because the multinomial logit likelihood cannot be analytically integrated by the multivariate normal distribution of part-worths. Moreover, I have a posterior distribution of β -bar and V_β and not just one multivariate normal. Taking full account of uncertainty, I therefore simulate 900 part-worths⁵ for each draw of β -bar and V_β , compute choice probabilities for each vector of part-worths, and then average to obtain predicted shares.

Table 6 summarizes the comparison to the observed holdout shares. I report the sum of squared prediction errors over 48 shares (SSE), the mean absolute errors (MAE*100—for direct comparison to the figures in Johnson and Orme’s paper) and the log-probability of all 900 x 12 holdout choices given the shares predicted from the model.

The log-probability of choices by holdout respondents assuming random choices is $12 \times \ln(1/4) \times 900 = -14,972$. All models predict considerably better than this naïve baseline. The results show that the *scaled MNL* model predicts shares best and that more prior heterogeneity results in better share predictions. However, the differences between models in terms of log-probabilities are less pronounced than those reported in Table 5.

The predictions in Table 6 alone suggest putting all weight on *scaled MNL*, but jointly predicting holdout responses (Table 5) and holdout respondents, *Binary* is still the clear winner. Based on joint predictions, *Binary* is chosen with a posterior model probability of 1 from all specifications investigated, assuming equal prior model probabilities (a probability of .0025 for prior $V_\beta = 6$ I and of .9975 for prior $V_\beta = 10$ I).

⁴ Three choices with four alternative each and one choice with three alternatives. Holdout responses were only available for 276 respondents out of the 326 that completed the ACBC interview. Bryan Orme indicated that the remaining holdout responses were discarded because of clearly outlying response times.

⁵ The goal is to integrate the multinomial logit by the heterogeneity distribution. The simulated part-worths are not meant to represent individual respondents but are used as ‘integration dummies’. Ideally, one would want to use a sample size of infinity here. However, I tried with much smaller samples of integration dummies too and found 900 to be considerably more than needed for stable solutions.

Table 6
Holdout shares predicted from posterior heterogeneity distribution

	Prior V_{β}	SSE	MAE	log-probability
<i>Multichoice</i>	2	0.21	4.71	-12,734
	6	0.19	4.60	-12,679
	10	0.18	4.44	-12,648
<i>Binary</i>	2	0.24	5.19	-12,817
	6	0.21	4.44	-12,719
	10	0.20	4.65	-12,676
<i>MNL</i>	2	0.18	4.33	-12,629
	6	0.18	4.29	-12,616
	10	0.17	4.19	-12,593
<i>Scaled MNL</i>	2	0.16	4.13	-12,587
	6	0.16	4.13	-12,579
	10	0.16	4.10	-12,572

It is of course disturbing to see models perform inconsistently across different predictive exercises. In general such a result indicates that all candidate specifications are only rough approximations and that a model reflecting the data generating mechanism has yet to be found. I will revisit this point in Section 6.

In practice it seems that share predictions are often made using part-worths obtained from the calibration sample directly. Table 7 reports share predictions using posterior expectations of individual part-worths in the calibration sample. Table 8 reports share predictions taking posterior uncertainty about individual part-worths into account. That is, I compute choice probabilities for each part-worth draw for each respondent in the calibration data and then average to get predicted shares.

A comparison of share predictions using posterior expectations of part-worths in Table 7 to predictions based on the distribution of heterogeneity in Table 6 indicates that the heterogeneity distribution generalizes considerably better than the collection of part-worths in the calibration sample. That is, the subjective hierarchical prior, updated using a collection of individual level likelihoods, successfully inter- and extrapolates the information in the individual level likelihoods.

Taking uncertainty in the posterior knowledge about individual part-worths into account improves share predictions considerably (Table 8). However, share predictions based on the heterogeneity distribution are still uniformly better.

Table 7
 Holdout shares predicted from posterior expectations of individual part-worths

	Prior V_{β}	SSE	MAE	log-probability
<i>Multichoice</i>	2	0.31	5.85	-13,075
	6	0.30	5.79	-13,021
	10	0.29	5.69	-12,978
<i>Binary</i>	2	0.40	6.71	-13,344
	6	0.37	6.42	-13,228
	10	0.35	6.22	-13,149
<i>MNL</i>	2	0.26	5.31	-12,843
	6	0.25	5.21	-12,840
	10	0.24	5.13	-12,804
<i>Scaled MNL</i>	2	0.23	4.98	-12,799
	6	0.23	4.98	-12,793
	10	0.22	4.94	-12,776

Table 8
 Holdout shares predicted from posterior distribution of individual part-worths

	Prior V_{β}	SSE	MAE	log-probability
<i>Multichoice</i>	2	0.21	4.81	-12,744
	6	0.21	4.75	-12,707
	10	0.20	4.65	-12,685
<i>Binary</i>	2	0.25	5.27	-12,828
	6	0.23	5.00	-12,757
	10	0.22	4.85	-12,727
<i>MNL</i>	2	0.19	4.48	-12,640
	6	0.18	4.44	-12,633
	10	0.18	4.42	-12,617
<i>Scaled MNL</i>	2	0.17	4.25	-12,600
	6	0.17	4.31	-12,599
	10	0.17	4.29	-12,593

6. DISCUSSION

I compared different likelihoods for the SCREENER part of the interview in combination with standard likelihood choices for BYO and the CBC part. Among all specifications investigated, treating the SCREENER observations as independent comparisons to some unobserved outside option (*Binary*), predicted best overall. However, *Binary* was not the preferred specification considering only share predictions to a holdout sample of respondents. This disparity strongly suggests that none of the specifications investigated should be settled on and further research is needed.

For the adaptive nature of the ACBC design to become ignorable in a hierarchical analysis, the likelihood has to extract at least the information contained in earlier parts of the interview entering design considerations for later parts of the interview. The critical junction in this respect appears to be the transition from the SCREENER to the CBC part of the interview. Any likelihood with shared coefficients between the SCREENER and CBC implies that information extracted from one respondent's SCREENER data ends up informing a prior for another respondent's CBC data and vice versa. To the extent that responses in the SCREENER part are a priori known to follow an entirely different behavioral protocol, shared coefficients do not make a lot of sense.

However, prior knowledge about behavioral differences between the SCREENER and the CBC responses is limited. Also, at the individual level, CBC only provides likelihood information about a subset of coefficients that is associated with variance in the levels among the accepted alternatives. Thus, a mechanism connecting the SCREENER and CBC through shared coefficients is inherently desirable.

I suggested the possibility that rejections in the SCREENER simply indicate that a respondent is certain about the inferiority of the rejected alternative. Differential certainty can be modeled using task specific scales. The results from *scaled MNL* indicate that the scale for the SCREENER is about 1.5 times the scale of the other parts of the interview, even with very informative priors suggesting that all tasks should be scaled the same.

In the limit, a very large scale for the SCREENER, everything else equal, implies ordinal constraints among the deterministic utilities in the SCREENER. Such ordinal constraints are more likely to translate into preference information that generalizes to the CBC part than individual SCREENER likelihoods that are maximized at disproportionately large values of individual parameters. This will be the case whenever a particular set of coefficients perfectly separates accepted and rejected alternatives.

In terms of future research answers to the following questions will further shed light on how to best model data collected using ACBC:

- Can the performance of *Binary* and *Multichoice* be improved by using task-specific scales? In this context it may also be worthwhile to distinguish between respondents with essentially deterministic SCREENER responses (large scale) and respondents whose SCREENER response are less internally consistent.
- How does a model separating attribute-based screening and compensatory decision making (Gilbride & Allenby 2004), calibrated on ACBC data, perform in predictions?

- Are responses to rules questions based on expected utility considerations? One way to obtain a direct likelihood for answers to the rules questions is to associate rules with the expected utility from sets with and without alternatives that have the screening attribute. An operational difficulty in this context is to define the set over which expected utility is formed.
- Can response times be used to directly inform a model about the increasing difficulty of choices throughout the interview?

In conclusion, I think that ACBC is an intuitively sensible, exciting format to collect preference information that presents some very interesting modeling challenges. The quest for tailor-made models to help bring the technique to its full potential has just begun.

REFERENCES

- Chapman, Randall G. and Richard Staelin (1982), Exploiting Rank Ordered Choice Set Data Within the Stochastic Utility Model, *Journal of Marketing Research*, 19, 288-301.
- Gilbride, Timothy J. and Greg M. Allenby (2004), A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules, *Marketing Science*, 23, 391-406.
- Johnson, Richard M. and Bryan K. Orme (2007), A New Approach to Adaptive CBC, *Sawtooth Software Conference Proceedings*, Sawtooth Software, Inc.
- van Ophem, Hans, Piet Stam and Bernard van Praag (1999), Multichoice Logit: Modeling Incomplete Preference Rankings of Classical Concerts, *Journal of Business & Economic Statistics*, 17, 117-128.

EM CBC: A NEW FRAMEWORK FOR DERIVING INDIVIDUAL CONJOINT UTILITIES BY ESTIMATING RESPONSES TO UNOBSERVED TASKS VIA EXPECTATION-MAXIMIZATION (EM)

*KEVIN LATTERY
MARITZ RESEARCH*

In the last decade, Hierarchical Bayes (HB) has advanced conjoint analysis significantly by enabling one to estimate individual level conjoint utilities. HB frames the problem by thinking of utilities as having a distribution that can be approximated. This paper outlines a completely different way of framing and solving the problem of individual utilities, one that has nothing to do with conceiving of utilities as distributions. As I will show later, this new framework brings certain advantages, such as incorporating individual level constraints into utilities, or even allowing different model structures across different respondents.

In the new framework, each respondent can be viewed as seeing all the tasks in a very large design plan, but completing a subset of them. If we knew how each respondent would have completed all the tasks then we can estimate a single respondent's utilities using just the data for the respondent.

For example, consider a case where we have 40 parameters to estimate, and each respondent completed 15 out of 60 tasks. If we knew the choices a specific respondent would have made for the remaining 45 tasks, then we could estimate the utilities for this respondent. We could then repeat this for the next respondent, doing each individual one at a time. The central assumption in each estimation is that the total set of tasks (in this case 60) will be larger than the degrees of freedom (in this case 40).

Adopting this new framework, the key tactical question becomes how to estimate the missing values – the responses to tasks that respondents did not complete. In this case, Expectation-Maximization (EM) seems like a natural (but not the only) choice to estimate these missing values. The following sections outline this method in some detail, and compare the results with HB. I also describe some advantages this new framework brings.

OUTLINE OF THE SOLUTION

As a simple example, consider an experimental design where each respondent completes 12 tasks out of 48 total tasks. Then consider a specific respondent who completed tasks 1-12, but not 13-48. This respondent's hypothetical data are shown below. The question marks show what we do not know: neither the individual utilities nor the individual's responses to tasks 13-48.

Experimental Design (48 Tasks with this Individual Doing 1-12)							Individual's Choices		
Task	var1	var2	var3	...	var19	var20	Option 1	Option 2	Option 3
1	1	1	0	...	0	1	1	0	0
2	1	1	1	...	1	1	0	0	1
3	1	1	0	...	0	1	1	0	0
4	1	0	0	...	1	1	0	0	1
...
12	0	1	1	...	0	1	0	0	1
13	0	0	0	...	1	0	?	?	?
14	1	0	0	...	0	1	?	?	?
15	1	1	0	...	1	1	?	?	?
...	?	?	?
48	1	0	0	...	0	0	?	?	?

Individual's Utilities	?	?	?	?	?	?
------------------------	---	---	---	---	---	---

If we knew the respondent's choices to tasks 13-48, we could estimate the utilities. Conversely, if we knew the utilities, we would know the respondent's choices to the incomplete tasks. This allows us to create an EM loop, alternating between estimating the hidden responses and computing the utilities. To better understand how this works, consider the following detailed steps.

1. The first step is to take a guess at the hidden data – what the respondent might have said if they had actually seen the tasks. The obvious candidate for the initial guess is the set of mean observed values. So for task 13, the initial guess would be the mean response of those who saw task 13. In this way we have an initial set of imputed missing values to tasks 13-48.
2. Using the complete data (all 48 tasks), we can now estimate utilities for the individual using typical conditional multinomial logistic regression (or another method if desired). Note that since we are solving for one individual at a time the estimation can be as customized as we want it. This can include individual constraints or even specific model structures.
3. Based on the individual's utilities (and the experimental design) we can compute the predicted values for the hidden responses. This gives us an updated estimate of how the respondent would have completed tasks 13-48. Note that the observed data in tasks 1-12 remains constant.

We repeat steps 2 and 3, several times. With each successive loop the imputed hidden responses will change and the observed log likelihood will continue to improve, with each new iteration improving the fit of the utilities to the individual's observed data. We repeat this loop until the observed loglikelihood converges. By doing this EM Loop for each respondent (one at a time) the final result will be an estimate of the utilities for each respondent. Moreover, the utilities will predict the observed data very well (in fact too well).

This EM Loop is the core of the method developed. But as we will see doing just this to derive utilities for each individual is overly simplistic.

FROM SIMPLE TO SOPHISTICATED EM

If one runs the simple EM loop described above, EM will consistently improve the log likelihood of observed responses with each iteration until the observed data is fit almost perfectly. This can result in a severe overfitting of the observed data which is manifest in the two following problems.

- i. The predicted responses for the observed data will likely have a completely different distribution than the hidden data. For any specific task, the observed and hidden predictions will have very different distributions (means and standard deviations).
- ii. Prediction of holdout tasks may be significantly worse than HB models. This again is because we have overfit each individual's data.

To solve these problems, I suggest the following steps.

- 1) The convergence criterion for EM is relaxed. EM will improve the observed log likelihood with each iteration. Rather than running EM until the observed data is fit almost perfectly, one should set a more relaxed criteria. This will give a set of predicted probabilities that do not overfit the individual's observed data. For the results here I have set the convergence criterion to be 90% of the loglikelihood between a random and perfectly fit model. I have also set a maximum of 10 EM loops, which is more than enough for all but a few cases.
- 2) Adjust the imputed hidden responses by comparing them with the observed responses. For any specific task, the distribution of predicted responses for the observed data should be similar to the hidden data. For instance, in scenario 1, 10% may have actually picked option 1. If the predicted hidden responses show in aggregate that option 1 = 40% and option 2 = 60%, then the predicted hidden responses are systematically overstating option 1 and need to be adjusted.
- 3) After adjusting hidden responses to match observed responses, loop back to step 1 and run EM for each individual using the adjusted hidden responses (and the observed responses).

Thus an EM loop is run for each respondent. But the convergence criteria is somewhat relaxed. Then after each respondent is estimated, we compare the predictions for the observed and imputed data. The imputed data is adjusted to match the observed data. And the EM loop is repeated for each individual. This repeats until the imputed data converges.

The net result is a set of utilities for each respondent that is consistent with the observed data, and where the imputed data matches the observed data in aggregate. As described in the appendix this aggregate matching includes both the mean and covariance matrix of the responses. Another way of viewing the algorithm is that it takes the observed means and covariance matrix of the responses and distributes the imputed responses across the entire sample in a way that is consistent with each individual's observed responses. EM is the method for determining which direction each individual's hidden responses should move in a manner consistent with their observed responses.

Steps 1 and 2 above introduce flexibility to the algorithm. What convergence criteria should be used and should this change with each iteration? Also, how should we adjust hidden

responses to match observed responses in aggregate? The parameters we have used are discussed in the appendix. But these are open for discussion.

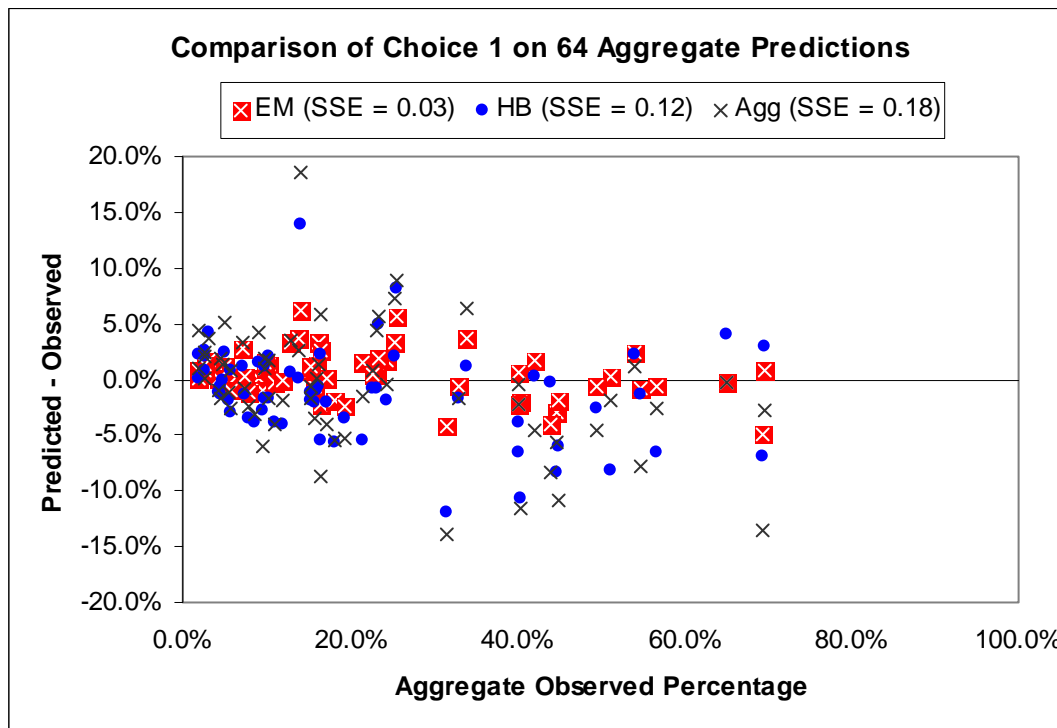
COMPARATIVE RESULTS (EM AND HB)

We ran 3 case studies, each with one holdout task.

	Study 1	Study 2	Study 3
Alternative Specific	Yes, 3 Brands and None Option	No (3 Choices)	No (4 choices)
Number of Attributes	8 per brand	4	4
Degrees of Freedom	46	17	15
Tasks Per Respondent	8	12	10
Total Tasks	64	36	40
Total N Size	951	935	602

Study 1 has the most degrees of freedom. Moreover, respondents only completed 8 of 64 tasks. So we expect study 1 to be the most difficult to model. The following shows how EM and HB compare on the in-sample predictions (later we will consider holdout tasks). Specifically, we look at the predicted versus observed aggregate share for each task.

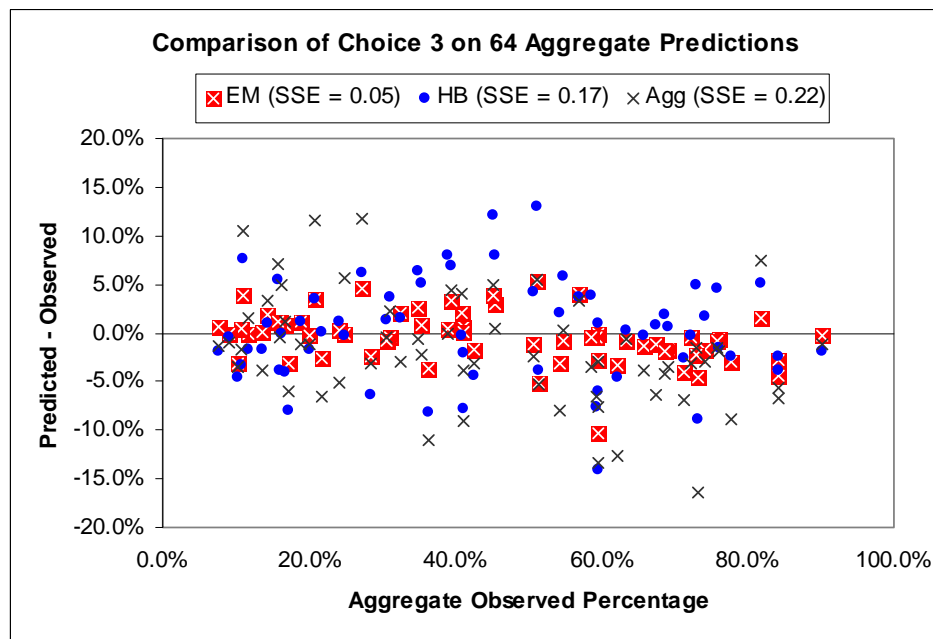
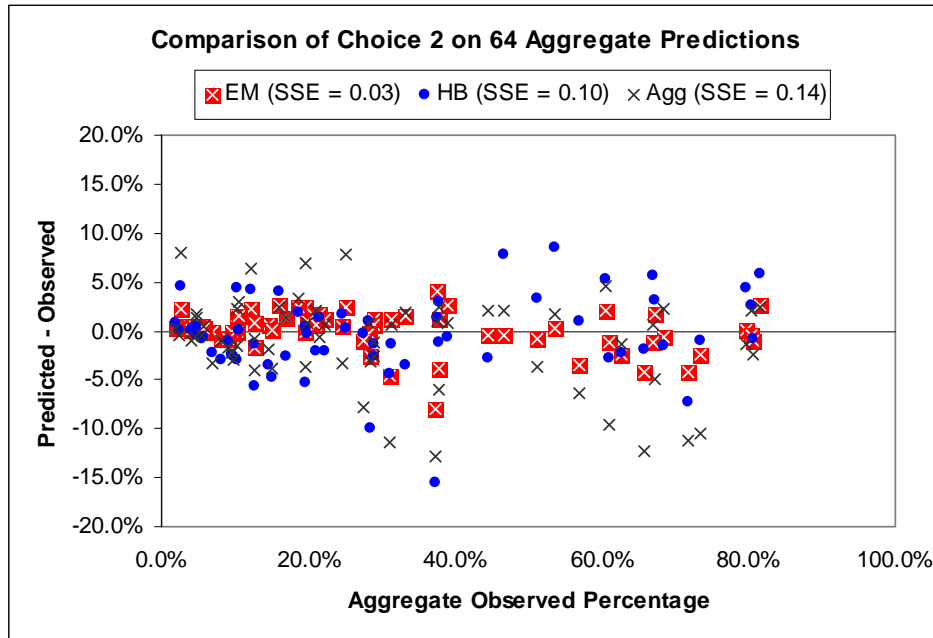
The chart below shows this for the first of four choices, for each of the 64 tasks in Study 1.

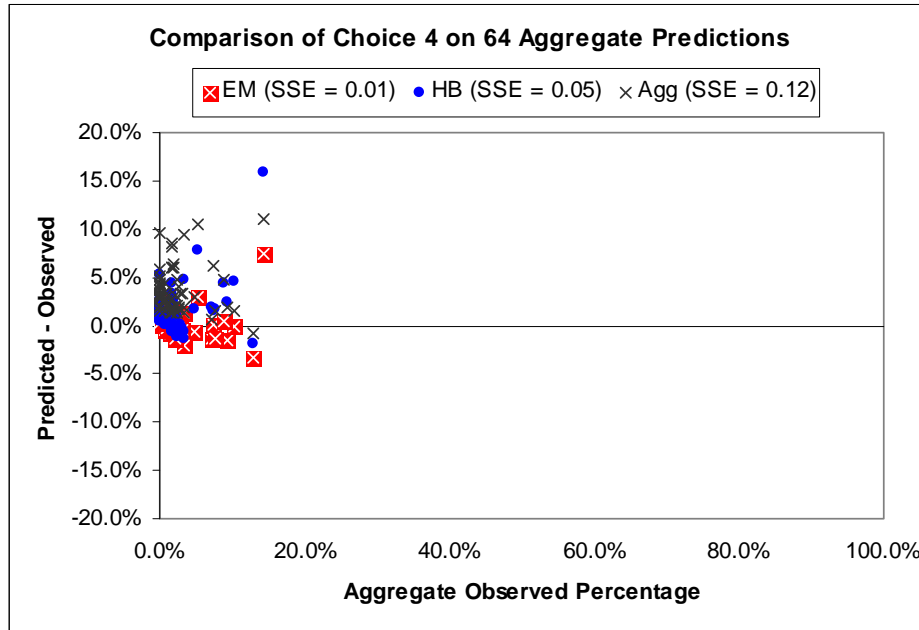


While HB does better than an aggregate model, EM clearly outperforms HB with respect to predicting in-sample share. Of course, this sophisticated version of the EM algorithm directly compares the observed and imputed in aggregate and adjusts them to match, so we should expect a very close fit from the EM model.

HB has no direct provision for checking aggregate observed data with predicted data. In many cases, the observed and predicted shares are excellent, but it is quite possible for them to vary significantly. In our experience, this problem can be exacerbated when weights are added. Considering the lack of direct checking against observed shares, HB does reasonably well, and is typically better than an aggregate model. Note that the HB global scale parameter in these studies was tuned to minimize $(\text{Observed mean} - \text{Predicted mean})^2$ of the in-sample tasks.

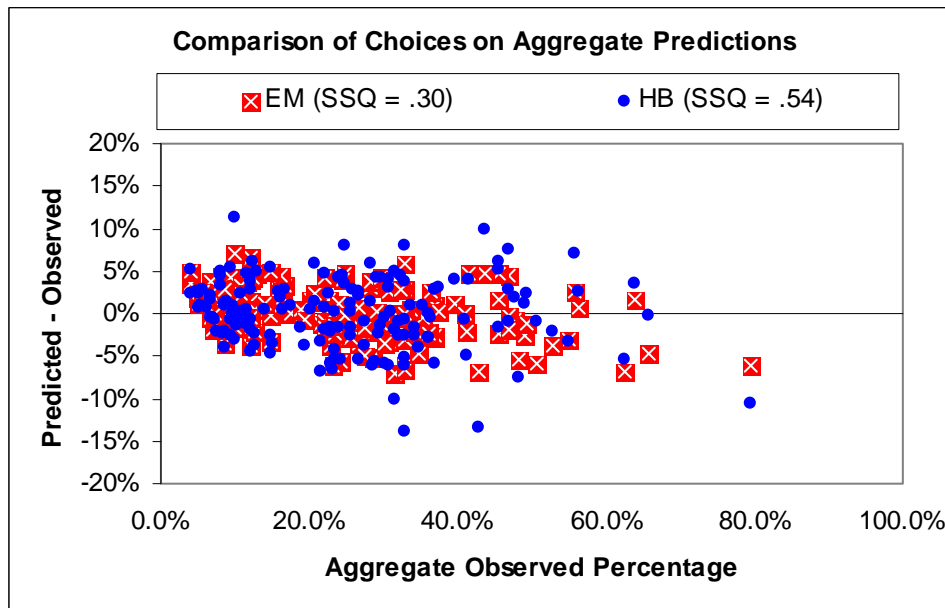
Studies 2 and 3 show similar results to Study 1, with EM doing a significantly better job than HB at predicting in-sample tasks.





Note that choice 4 was a None option. Hence the somewhat low share with small variance.

EM also does better in-sample than HB for the other two studies though to a lesser degree. This is most likely due to Studies 2 and 3 having more reasonable degrees of freedom for the number of tasks completed.



Given the EM algorithm as specified above with the comparison of predicted and observed shares, we expect it to always perform well predicting in-sample. HB may or may not do relatively well, and in those cases where HB does not do well, EM provides an excellent alternative.

PREDICTING HOLDOUT TASKS

One holdout task was completed for each of the three studies above. So for each choice, we have an observed share, the percentage of respondents who picked a specific choice. We can compare that aggregate percentage with the percentage predicted by the HB and EM models.

While EM had a clear advantage over HB predicting in-sample, EM has only a slight advantage over HB in predicting the holdout tasks in aggregate. As the tables below show, EM predicted the aggregate results better than HB, but not by as much as one might expect given the larger in-sample difference. In fact HB did better at predicting the holdout tasks than it did for many of the in-sample tasks.

Study 1	1	2	3	4	SSE
Observed	25.4%	19.6%	54.3%	0.7%	N/A
EM	28.6%	19.5%	51.2%	0.7%	0.20%
HB	27.5%	14.3%	56.4%	1.9%	0.37%
Agg	32.7%	16.0%	46.3%	5.0%	1.49%
Study 2	1	2	3		SSE
Observed	52.1%	15.9%	32.0%		N/A
EM	54.2%	14.7%	31.1%		0.07%
HB	46.6%	25.4%	28.0%		1.82%
Study 3	1	2	3	4	SSE
Observed	22.8%	37.6%	12.1%	27.5%	N/A
EM	19.3%	40.7%	13.8%	26.1%	0.27%
HB	16.9%	40.6%	16.0%	26.5%	0.60%

When we turn from aggregate prediction of holdout tasks to individual hit-rate, HB shows a slight edge in all 3 studies, with a hit rate of .6% to 2.7% better than EM.

	EM Hit Rate	HB Hit Rate
Study 1	50.9%	52.6%
Study 2	64.2%	64.8%
Study 3	65.1%	67.8%

One of the most remarkable findings is the level of agreement between HB and EM with respect to predicting individual hits. If we think of a Venn Diagram for HB and EM, most individuals are predicted correctly by both methods or incorrectly by both methods. In

particular, the overlap between the two methods is 65-78%. This remarkable similarity suggests EM and HB share something significant in how they treat individual respondents, despite their very different theoretical approach.

	Both EM and HB	Neither EM nor HB	EM Only	HB Only	Agreement of Methods (Both + Neither)
Study 1	34.8%	30.4%	16.1%	17.8%	65.2%
Study 2	53.0%	24.0%	11.2%	11.8%	77.0%
Study 3	55.7%	22.8%	9.4%	12.1%	78.5%

Study 3 asked respondents to state their preferred level for 3 of the 4 conjoint attributes. It is important to note that the preferred level of each attribute varies for each respondent. The attribute is not ordinal. We compared the conjoint utilities with the stated preference to determine the extent to which these match.

Attribute	Number of Levels	Predict Best Level	
		HB	EM
Attribute 1	5	41.2%	34.9%
Attribute 2	4	57.3%	54.2%
Attribute 3	5	61.5%	57.2%

Both HB and EM did much better than chance, with HB outperforming EM. However, neither HB nor EM did very well matching the stated preference levels. But again what is remarkable is that both methods showed the same pattern, with attribute 1 being hardest to predict the stated preference and attribute 3 the easiest. It is also worth noting the prediction rates are highest for the most important attributes (2 and 3), and lowest for attribute 1 which is least important. Again, this suggests some similarities between the two methods.

One of the advantages of EM is that regression utilities are estimated for each individual one at a time. This means constraints can be incorporated specific to each individual. Incorporating these individual level constraints significantly improved EM holdout task prediction from 65.1% to 72.1%. This gives the individual constrained EM model a higher hit rate than HB, and retains its other advantages – aggregate fit of holdout and non-holdout tasks. Obviously, it also improved the prediction of the best level to 100% since this was the constraint.

	Both EM and HB	Neither EM nor HB	EM Only	HB Only	Agreement of Methods (Both + Neither)	EM Hit Rate	HB Hit Rate
Study 1	34.8%	30.4%	16.1%	17.8%	65.2%	50.9%	52.6%
Study 2	53.0%	24.0%	11.2%	11.8%	77.0%	64.2%	64.8%
Study 3	55.7%	22.8%	9.4%	12.1%	78.5%	65.1%	67.8%
Study 3 with EM Individual Constraints	57.2%	17.8%	14.9%	10.6%	75.0%	72.1%	67.8%

This is only one study and we'd expect the improvement in predicting hit rate for the EM model to depend on the specific constraints and the importance of the attribute. In this study,

attributes 2 and 3 (the most important attributes) improved the hit-rate. However, incorporating attribute 1 constraint had no impact on hit rate. In the case of attribute 1, the respondent's preference appeared to have no impact on their decision. So the moral of the story is that it is not enough to incorporate individual constraints, they must be relevant individual constraints.

FINAL CONCLUSIONS

This paper outlined a completely different way of framing and solving the problem of estimating individual level utilities – one where we estimate hidden responses and model individual level utilities based on those estimates. Since we are solving for one individual at a time the estimation can be as customized as we want it. This can include individual constraints or even specific model structures. Having this option available may prove useful in many situations: when we want to customize the way we treat each individual via individual constraints, thresholds, models (compensatory or EBA), etc.

This paper only began to explore the benefits of this new framework. We saw that incorporating relevant individual level constraints improved hit-rate. The more information we can incorporate about each individual, the better each individual's utilities can be estimated.

To help this new framework along I have outlined one potential way to estimate hidden responses using the EM algorithm. It is quite possible that the specific method by which I applied EM could be improved. Moreover, EM itself may not be the best method to estimate hidden responses.

While the EM algorithm described uses a very different framework than a random coefficient model like HB, EM and HB tend to predict the same individuals correctly or incorrectly. This suggests both methods share something significant in how they treat individual respondents.

In studies where HB predicts aggregate holdout and non-holdout tasks very well, EM will not offer much or any improvement in these predictions. But in those cases where HB does not predict aggregate results well, EM will likely offer significant improvement due to its explicit checking of observed and imputed data. Also, at this point, we do not expect EM to yield a higher hit-rate of holdout tasks than HB, unless relevant individual level constraints are incorporated. Future changes to the EM algorithm may also help.

There is also a great deal of flexibility in setting convergence rates and how to adjust the hidden data after each individual is estimated. While I have not tested it, different starting points for the EM algorithm would definitely be interesting, and may significantly improve the algorithm at the expense of more computing time.

In conclusion, the new framework using sophisticated EM appears to work very well already. It also shows tremendous promise for broadening our approach toward conjoint analysis, especially in developing custom models for each respondent.

APPENDIX

DETAILS FOR ADJUSTING HIDDEN VALUES

The initial means are used for the first iteration of the EM Loop.

The first EM Loop yields predicted values which are adjusted based on a linear adjustment for the second iteration. The linear adjustment will have a mean that is within the margin of error of the observed mean for each task. Moreover, the adjusted percentages are a linear transformation of the unadjusted percentages for each task and choice. For example, choice 1 of task 1 has a linear adjustment, choice 2 of task 1 has different linear adjustment, etc.

For the remaining iterations, the NORTA algorithm is applied to match the mean and covariance of the predicted observed data.

The Linear and NORTA adjustment are discussed below.

LINEAR ADJUSTMENT

The hidden values for a specific task are adjusted by a linear function: $y = mx_i + b_i$, where each i indexes the n choices for a task. To derive b_i and m , first take the hidden data for one specific task and compute the Minimum and Mean for each of the n choices of the task.

Then for each of the n choices, we have one parameter to estimate b_i where b represents the intercept. The slope will be the same across all choices, equal to $(1 - \sum b_i)$.

Set the following constraints for b_i :

Minimum of -1 and max of +1

Linear transformation of min > 0

Linear transformation of mean \geq lower bound of observed mean

Linear transformation of mean \leq upper bound of observed mean

Minimize the sum of b_i which is equivalent to maximizing the slope $(1 - \sum b_i)$.

If the slope is 1 then minimize the sum of the squares of b_i . Note that in the worst case scenario, each b_i is the observed mean and the resulting slope is $1 - \sum b_i = 0$. This produces a constant mean value for each choice.

NORTA ADJUSTMENT

NORTA (Normal to Anything) is a general algorithm used for fitting multivariate normal data to a target covariance matrix and mean.

Let SVD of observed covariance matrix be: $u * d1 * t(u)$.

Let SVD of hidden covariance matrix be: $v * d2 * t(v)$.

Then a transformation of normally distributed data to a data set with the observed covariance matrix is given by the NORTA algorithm:

$$[u * \text{sqrt}(d1) * t(u)] * \text{Normal Data}$$

We can reverse this, so a transformation of the hidden covariance matrix to multivariate Normal is:

$$\text{Inv}([v * \text{sqrt}(d2) * t(v)]) * \text{Hidden Data}$$

So a transformation of Hidden data with its covariance matrix to the covariance matrix of observed data is:

$$[u * \text{sqrt}(d1) * t(u)] * \{\text{Inv}([v * \text{sqrt}(d2) * t(v)]) * \text{Hidden Data}\}$$

The mean can then be added to each column.

This will transform Hidden data to have the covariance matrix and mean of the observed data. All linear dependencies will be preserved (so the sum of choices will still 1 for each task). In some cases choices will be <0 and >1. Choices less than 0 are set to 0 and the data repercentaged.

In the NORTA algorithm I actually use the predicted observed data rather than the true observed data, with the idea being to balance the predicted results between observed and hidden data. Also note that NORTA is done for each listwise complete block of tasks. So if tasks 1-12 were a block of tasks, then we have one computation for that block of tasks, where each of the 12 tasks and n choices is a column of data.

REMOVING THE SCALE FACTOR CONFOUND IN MULTINOMIAL LOGIT CHOICE MODELS TO OBTAIN BETTER ESTIMATES OF PREFERENCE¹

JAY MAGIDSON
STATISTICAL INNOVATIONS INC.
JEROEN K. VERMUNT
TILBURG UNIVERSITY

INTRODUCTION AND GOAL

Multinomial logit choice models based on latent class (LC) or HB methods are utilized in marketing research today along with simulators which predict choices and market shares. However, a weakness in these models creates a potential interpretability and validity problem. The problem is that the part-worth preference (utility) parameters that are used to make such predictions are generally confounded with a *scale parameter* which reflects the amount of uncertainty by different respondents.

This scale parameter confound is assumed not to exist in current HB models under the common restrictive assumption that the level of uncertainty, inferred by the values of the scale parameter, is identical for each respondent. Similarly, no confound occurs in LC choice models under the assumption that the values of the scale factor are the same for all respondents regardless of the latent class to which they belong (Louviere *et al.*, 2000, page 206). If the equality assumption is violated, the predictions contain additional amounts of error as well as potential bias.

We begin this paper by introducing the scale factor issue in conjunction with some simple hypothetical LC choice models. We then propose some extended LC choice models that can be used to test the equal scale factor assumption, and when violated, to estimate and thus remove the effects of the different scale parameters that exist for different respondent subgroups. By not allowing differences in the scale factor to be a determinant in the formation of different latent classes, LC segments resulting from this new approach more clearly differentiate respondents based on their true preference differences by revealing their scale free part-worth preference utilities. Thus, isolation of the scale factors should result in more *pure* measures of part-worth preference utilities.

Applications of these new extended models² in 2 real world choice experiments find that in both cases the commonly made assumption of equal scale factors is unwarranted. Comparing results from the standard models to the extended models suggests that the former leads to some faulty conclusions regarding the actual part-worths as well as misclassification of a substantial number of respondents into the *wrong* LC segment. The results also suggest that discrete factor (DFactor) choice models, with and without a scale-factor adjustment, may well provide

¹ The authors wish to thank John Wurst for his helpful comments.

² These models, which include extensions of standard LC choice models as well as discrete factor choice models, have been implemented in the syntax version of the Latent GOLD Choice program (Vermunt and Magidson, 2008).

additional improvements in both data fit as well as interpretation of results over the more traditional LC choice models.

WHAT IS THE SCALE PARAMETER?

The scale parameter λ_i relates to the amount of *certainty* in respondent i 's *expected choices*. Swait and Louviere (1993) pointed out for a given respondent i , standard choice modeling estimates the product $\lambda_i \beta$ rather than β . Thus, differences in the scale parameters should be isolated in order to obtain meaningful comparisons of preference part-worth β -parameters across individuals or groups.

The scale parameter can also be explained in terms of the variance in observed responses. Response variance is inversely related to λ (variance = $\pi^2/6\lambda_i^2$). Thus response variance may be viewed as a measure of *uncertainty* (lack of certainty). For persons with λ approaching 0, the response variance approaches infinity, which reflects *complete* uncertainty. In this case, choice probabilities derived from the multinomial logit model are equal for all alternatives (Ben-Akiva and Lerman, 1985).

The magnitude and importance of such confounds are not well understood. Clearly, Louviere and Eagle (2006) believe the scale factor issue is very important. In their 2006 Sawtooth Conference presentation they state:

“All choice models confound scale and [preference part-worth] parameter estimates. The confound is particularly problematic in complex models like random coefficients [HB-like] models and latent class models if one cannot separate scale and [preference] parameters.”

“Thus, it is likely that random coefficient models are biased and misleading.”

“So, the bottom line is that one cannot estimate individual-level parameters from choice models unless one can separate scale and model parameter estimates.”

and

“The field needs research that leads to new models that can capture both scale and systematic component (mean) effects.”

As a simple illustration of the scale parameter, consider first the situation where there is complete homogeneity with regards to brand preference. Specifically, assume that all respondents prefer brand A over B, and B over C and have the following identical BRAND preference part-worth utilities: $\beta_A = 0.5$, $\beta_B = 0.1$, and $\beta_C = -0.6$.

Further suppose that 2 respondent subgroups exist, the first expressing less certainty than the other in several pair-wise brand comparisons. The scale factors for these two groups are denoted by $\lambda[1]$ and $\lambda[2]$ respectively. Without loss of generality, for purposes of parameter identification, we fix $\lambda[1] = 1$ for the first subgroup, and for concreteness assume that $\lambda[2] = 2$ for the second subgroup. This situation is reflected in Table 1.

Table 1:
Assumed values for the preference parameter β ,
and the product $\lambda[s]*\beta$ for each of the 2 subgroups.

BRAND	Preference Parameter (β)		$\lambda[s]*\beta$	
	Subgroup s=1	Subgroup s=2	$\lambda[1] = 1$ Subgroup s=1	$\lambda[2] = 2$ Subgroup s=2
A	+ 0.5	+ 0.5	+ 0.5	+ 1.0
B	+ 0.1	+ 0.1	+ 0.1	+ 0.2
C	- 0.6	- 0.6	- 0.6	- 1.2

These parameters yield choice probabilities (shown in Table 2) which are obtained from the multinomial (conditional) logit model equations:

$$\text{Prob}(\text{Brand} = A | \text{Subgroup } s) = \exp(\lambda[s]*\beta_A) / \text{SUM}(s),$$

$$\text{Prob}(\text{Brand} = B | \text{Subgroup } s) = \exp(\lambda[s]*\beta_B) / \text{SUM}(s),$$

$$\text{Prob}(\text{Brand} = C | \text{Subgroup } s) = \exp(\lambda[s]*\beta_C) / \text{SUM}(s),$$

where

$$\text{SUM}(s) = \exp(\lambda[s]*\beta_A) + \exp(\lambda[s]*\beta_B) + \exp(\lambda[s]*\beta_C).$$

From this example, it should be clear that a higher value for $\lambda[s]$ (as occurs in subgroup 2) translates into more extreme choice probabilities.

Table 2:
Choice probabilities for each subgroup obtained from the multinomial logit model equations under the parameter values given in Table 1.

BRAND	Choice Probabilities	
	Subgroup s=1	Subgroup s=2
A	.50	.64
B	.33	.29
C	.17	.07

SCALE PARAMETER AS A CONFOUND

LC choice analysis identifies $K \geq 1$ different latent classes, each of which is associated with a unique set of part-worth parameters. Since a *standard* LC choice analysis assumes $\lambda = 1$ for all respondents, if data were generated according to the probabilities in Table 2, this standard analysis would tend to mistakenly identify each of the 2 subgroups as different LC segments with *different* preference part-worths as shown in Table 3 below:

Table 3:
Expected results under a standard LC choice analysis which mistakenly assumes $\lambda[1] = \lambda[2] = 1$ (True values shown in parenthesis).

	Estimated (True) part-worth preference parameters (β) under assumption $\lambda[s] = 1$	
BRAND	Subgroup 1	Subgroup 2
A	+ 0.5 (+ 0.5)	+ 1.0 (+ 0.5)
B	+ 0.1 (+ 0.1)	+ 0.2 (+ 0.1)
C	- 0.6 (- 0.6)	- 1.2 (- 0.6)

Table 3 shows the incorrect results where the part-worth parameter is confounded by the scale parameter. It is incorrect to infer that subgroup 2 (labeled ‘Segment 2’ in Table 3) prefers Brand A *more* than subgroup 1, or that subgroup 2 prefers brand C *less* than subgroup 1 because the true brand preference part-worths are in fact identical for both subgroups 1 and 2. The only true difference between respondents in Segments 1 and 2 is that the responses obtained from the former group reflect greater amounts of uncertainty.

SCALE-EXTENDED LC CHOICE MODELS

As mentioned above, *standard* LC choice models contain a K-category latent variable X representing K homogeneous LC segments. Each segment k has its own unique preference part-worths which are estimated under the assumption $\lambda = 1$ for all respondents.

LC Choice Model extension #1 contains 2 categorical latent variables, X* and S: X* denotes LC segments differing in their preference part-worth utilities. S denotes latent subgroups that differ in their scale parameter, with $\lambda[1]$ fixed to the value 1 for purposes of identification. That is, the λ corresponding to the first subgroup, $\lambda[1]$, is set to 1. The values for the other $\lambda[s]$ parameters are assumed to be nonnegative.

In the example described above, the standard LC choice model results in 2 LC segments that appear to differ in their part-worth utilities for each of the attributes, but upon further inspection

it can be seen that these part-worth utilities are proportional to each other. This is an indication that the λ equality assumption does not hold true.

These 2 segments can be explained more simply as a single homogeneous segment having the same preference part-worths, but with 2 different latent subgroups which have different scale parameter values $\lambda[2] \neq \lambda[1]$. Each respondent in subgroup 1 shares the same scale parameter $\lambda[1]$, and each respondent in subgroup 2 shares the same scale parameter $\lambda[2]$. That is, in this example S consists of 2 subgroups differing only in their level of uncertainty, not their preference part-worth utilities. Thus, S is dichotomous, and X^* has only a single category (i.e., there is only 1 LC segment).

Next assume that X^* has 2 underlying LC segments that truly differ in brand preferences, segment 1 most preferring brand A, while segment 2 most prefers brand C. Further suppose that each of these segments consists of both more and less certain respondents. This situation can be represented in terms of the 4 cells in Table 4 below. LC segment 1 consists of the 2 cells (1,1) and (1,2), while LC segment 2 contains cells (2,1) and (2,2).

Table 4:
Example with 2 LC segments and 2 scale parameter subgroups

	Less certain subgroup (s = 1): $\lambda[s]=1$	More certain subgroup (s = 2): $\lambda[s]=2$
Segment 1 ($X^*=1$): $\beta_1 = (0.5, 0.0, -0.5)$	joint class (1,1)	joint class (1,2)
Segment 2 ($X^*=2$): $\beta_1 = (-0.2, -0.8, 1.0)$	joint class (2,1)	joint class (2,2)

Our scale-extended choice model generalizes the latent class (LC) multinomial logit choice model from its standard log-linear form to a log-bilinear form. The standard LC choice model expresses the expected utility of the j-th alternative as a linear function of the part-worth parameters. For example, with attribute A1 at level l_1 , and A2 at l_2 , the expected utility of class $X=k$ is $\beta_{l_1,k}^{A1.X^*} + \beta_{l_2,k}^{A2.X^*}$.

The scale-extended model incorporates the scale parameter $\lambda[s]$, and thus is bi-linear in the parameters. This yields a logit model with a linear term of the form $\lambda_s \beta_{l_1,k}^{A1.X^*} + \lambda_s \beta_{l_2,k}^{A2.X^*}$, where the β and $\lambda[s]$ parameters are estimated simultaneously. Note that each part-worth is multiplied by the same scale parameter $\lambda[s]$.

Each member belonging to latent class segment $k=1,2,\dots,K$ shares the same pure preference utilities, but some of these members differ in uncertainty (i.e., differ on their scale factor). For purposes of identification, we set $\lambda_1 = 1$ for the largest subscale group. For 2 or more latent subgroups, each of the K LC segments contains some respondents with scale parameter = 1, and some with a higher value (reflecting lower amounts of uncertainty) and/or a lower value (reflecting larger amounts of uncertainty). When there is only a single subgroup, the model

reduces to a standard LC choice model, where all respondents are assumed to have a common scale factor ($\lambda = 1$.)

How do the results compare between a standard LC choice analysis designed to identify different latent segments, and a scale-extended LC model where within each class, subgroups of respondents are allowed to have different scale parameters? Theoretically, if the true underlying population conforms to Table 4, the standard LC choice modeling approach might identify 4 LC segments that might be more or less confounded, while the scale-extended approach would identify 2 segments, each consisting of more or less certain subgroups.

Below, two real-world choice data sets are used to illustrate the various models; the first is a 5-attribute choice experiment for coffee makers (Skrondal and Rabe-Hesketh, 2004), the second is a 50% random sample provided by Sawtooth Software of TV choice data originally described by Huber *et al.* (1998). The latter data has an additional variable reflecting the length of time to complete the choice task, information that might be expected to be related to one's scale factor.

We also introduce a new class of LC choice models, called discrete factor (DFactor) choice models. In clustering applications, DFactor models were shown to outperform traditional LC models (Magidson and Vermunt, 2001), so they might be expected to outperform traditional LC choice models as well. The DFactor models are illustrated in the first example.

EXAMPLE 1: COFFEE MAKERS – STANDARD VS. SCALE-EXTENDED MODELS

Consider data from a discrete choice study with 5 attributes – BRAND, CAPACITY, PRICE, FILTER and THERMOS (Skrondal and Rabe-Hesketh, 2004). For these data, models containing between $K = 1 - 6$ LC segments were estimated, each with between 1 and 3 scale subgroups. For the standard models (i.e., those containing only a single scale factor), the 5-class solution fit best (lowest BIC). For scale-extended models with S^* dichotomous (i.e., 2 scale subgroups), again the 5-class model fit best. Table 5 provides the fit statistics for these models, which were estimated using the LG-Syntax module in Latent GOLD Choice. Appendix A1 provides the model specifications. For more technical details, see Vermunt and Magidson (2008).

Table 5:
Fit measures for the LC choice models estimated with Coffee Makers data

	Standard LC Choice Models			LC Choice Models + 2 Scale Factors		
	LL	BIC(LL)	Npar	LL	BIC(LL)	Npar
1-class	-1298.7	2639.2	8	-1180.1	2412.4	10
2-class	-1110.6	2310.0	17	-1089.8	2278.7	19
3-class	-1073.6	2283.0	26	-1051.2	2248.6	28
4-class	-1043.2	2269.1	35	-1021.9	2236.9	37
5-class	-1013.6	2257.0	44	-996.9	2233.9	46
6-class	-992.9	2262.5	53	-974.3	2235.7	55

Compared to the standard models with K segments, the corresponding K -class scale-extended model contains 2 additional distinct parameters to be estimated -- $\lambda[2]$, the scale factor

for the 2nd subgroup and P[2], the proportion of the population falling into this 2nd subgroup.³ Note that based on the BIC criteria, the scale-extended models are preferred over the standard models regardless of the number of segments.

For the scale-extended models, the 2 latent variables were assumed to be independent. This hypothesis was tested for the 5-class model by a likelihood ratio comparison with a model assuming dependence between the two latent variables and found to be supported. Also, models with different numbers of scale subgroups were estimated for the 5-class model and those containing 2 subgroups were found to fit best.

For the 5-class scale-extended model, each segment k consists of both lower (s=1) and higher (s=2) certainty subgroups. The larger S-class consisted of 52% of the respondents who expressed higher amounts of certainty ($\lambda[2] = 11.2$) than the other subgroup ($\lambda[1] = 1$). The scale-extended model segments differed considerably from those from the standard model. As shown in Table 6, only 62+16+20+9+9 = 118 of the 185 respondents remain grouped in the same class. For example, only 62 of the 83 respondents grouped into class X*[1] were grouped into the corresponding X[1]-class. Overall, 37% of the respondents were classified into an X-class that differed from the corresponding X* classification.

Table 6:
Cross-tabulation of classification with and without scale factor

	X*[1]	X*[2]	X*[3]	X*[4]	X*[5]	Total
X[1]	62	26	0	7	0	95
X[3]	4	16	0	5	0	25
X[4]	0	1	20	0	0	21
X[2]	17	2	6	9	1	35
X[5]	0	0	0	0	9	9
Total=	83	45	26	21	10	185

DFACTOR MODELS AS AN ALTERNATIVE TO LC CHOICE MODELS

Magidson and Vermunt (2001) suggested the use of Discrete Factor (DFactor) models as an alternative to traditional LC Cluster models and introduced a basic model containing dichotomous uncorrelated DFactors that generally provided a more parsimonious explanation of data. Thus, it might be expected that basic DFactor choice models might also outperform traditional LC choice models.

Discrete Factor (DFactor) Choice Models posit M latent DFactors denoted as X_1, X_2, \dots, X_M with K_1, K_2, \dots, K_M categories which yields a total of $K_1 \times K_2 \times \dots \times K_M$ segments. A *basic* DFactor model contains *dichotomous* DFactors which are mutually *independent* of each other. These models impose a factor structure on the parameters of a LC choice model. For example,

³ Given P[2], P[1], the proportion in the first subgroup can be computed as $1 - P[2]$, and thus is not counted as a distinct parameter to be estimated.

the part-worth utility parameters corresponding to segment $(X_1, X_2) = (k_1, k_2)$ of a *basic* 2-DFactor model⁴ are expressed as:

$$\beta_{l_1.k_1,k_2}^{A_1.X_1,X_2} + \beta_{l_2.k_1,k_2}^{A_2.X_1,X_2} + \dots$$

where:

$$\beta_{l_1.k_1,k_2}^{A_1.X_1,X_2} = a_{l_1} + b_{l_11}^{A_1} X_1 + b_{l_12}^{A_1} X_2$$

and

$$\beta_{l_2.k_1,k_2}^{A_2.X_1,X_2} = a_{l_2} + b_{l_21}^{A_2} X_1 + b_{l_22}^{A_2} X_2$$

Since this is a restricted (structured) version of a LC choice model with $K = K_1 \times K_2 \times \dots \times K_M$ classes, it is a parsimonious model, containing relatively few parameters. In fact, a *basic* DFactor choice model with M DFactors has the same number of parameters as a LC choice model with only $K=M+1$ latent class segments. Thus, for example, a *basic* 2 DFactor model is a restricted 4-class model that contains the same number of parameters as a 3-class model.

LC Choice Model Extension #2 utilizes a Scale-Extended (DFactor) Choice Model which posits M DFactors $X_1^*, X_2^*, \dots, X_M^*$ and S. By default, each X_m^* is dichotomous. As before, the categories of S represent subgroups with different scale factors, and for identification purposes, the restriction $\lambda[1] = 1$ is made for the 1st category of S, which is the category having the *lower* scale factor.

EXAMPLE 1: MODEL RESULTS FOR DFACTOR MODELS

For the coffee-maker data, models containing between 1 and 4 DFactors were estimated, with and without the scale parameter. As in the standard choice models, the scale-extended form of the model contains 2 additional parameters. (See Appendices A1 and A2 for the syntax equations corresponding to the standard LC Choice and DFactor models.) The fit statistics for these models are given in Table 7.

Table 7
Fit measures for the DFactor choice models estimated with Coffee Makers data

Model	D-Factor Choice Models			Scale Adjusted D-Factor Choice Models			
	LL	BIC(LL)	Npar		LL	BIC(LL)	Npar
1-dfac	-1110.6	2310.0	17	1-dfac w scale	-1089.8	2278.7	19
2-dfac	-1056.6	2249.0	26	2-dfac w scale	-1039.4	2224.9	28
3-dfac	-1021.7	2226.1	35	3-dfac w scale	-1001.6	2196.4	37
4-dfac	-988.1	2205.9	44	4-dfac w scale	-971.3	2182.7	46

⁴ For dichotomous DFactors, dummy coding is used ($X_1 = 0$ or 1 and $X_2 = 0$ or 1) so that the constant a_{l_2} reflects the part-worth associated with attribute A_1 for the first LC segment $(0, 0)$.

First, comparing Table 7 with Table 5 we see that each *standard* DFactor model fits these data better (lower BIC) than the corresponding *standard* LC choice model containing the same number of parameters (Npar). Also, the *scale-extended* DFactor models outperform (lower BIC) the *standard* DFactor models. Among the models estimated, the scale-adjusted 4 DFactor model fit best.

As in the Extension #1, 2 scale subgroups were found to be best for 4-DFactor models. The larger S-class consisted of 61% of the respondents who expressed higher amounts of certainty ($\lambda[2] = 9.0$) than the other subgroup ($\lambda[1] = 1$). Again, S was found to be independent of X*. That is, about 61% of the members of each X* segment were classified into this S-subgroup.

Compared to the 5-Class Cluster solution, the 4-DFactor Scale-Extended Model classifies more respondents into the higher certainty group, 77% of whom are classified into the corresponding group by the Cluster approach. Similarly, 91% of those classified into the less certain group by the DFactor approach are classified into the corresponding group by the Cluster approach.

Overall, the S-subgroups identified by the different types of scale-extended choice models show strong agreement -- both models classify the same 99 respondents into the *more* certain group and the same 52 respondents into the *less* certain group. Overall, a somewhat higher number of respondents were classified into the more certain subgroup, consistent with the fact the scale factor for this at this subgroup was somewhat lower than obtained under the traditional approach (9.0 vs. 11.2).

Table 8:
Fit measures for the DFactor choice models estimated with Coffee Makers data

Comparison of S-subgroups obtained from Cluster and Dfactor Models				
	5-Class Cluster Model	4-DFactor Model Classifications		
		$\lambda = 1$	$\lambda = 9.0$	Total
1	$\lambda = 1$	52	29	81
2	$\lambda = 11.2$	5	99	104
	Total	57	128	185

EXAMPLE 2: SAWTOOTH SOFTWARE TV CHOICE DATA

It is possible to include covariates in a LC model to better predict/explain/describe the latent variable. In example #2 we include the time to complete the choice tasks as a numeric covariate, specify linear and quadratic time effects, and examine whether completion time is related to latent variables S and/or X*. Overall, the time to complete the 18 choice tasks ranged from 1 minute to 22 minutes with a mean time of 6.4 minutes.

This choice experiment consists of N = 176 respondents, who selected from various choice sets TVs that differed on their levels across 6 attributes – Brand, Screen Size, Sound, Channel

Blocking, Picture-in-Picture availability and Price. The data were obtained as a random 50% sample of all respondents analyzed originally by Huber *et al.* (1998).

For these data we estimated a 4-class LC choice model where the 4th class has zero effects. That is, for segment #4, the choices are assumed to *not* be affected by the attributes (i.e., segment #4 is a random response segment). We compare models with and without a scale parameter. For the former, 2 scale subgroups were assumed. Both models included time as an active covariate. The model specifications are given in Appendix B.

For the standard LC choice model (no scale parameter), time was found to be a significant predictor of the classes (X), the 3rd and 4th classes having significantly lower mean completion time than classes 1 and 2. The scale-extended LC Choice model was preferred based on the BIC.

Results from the scale-extended model were as follows:

- The second subgroup of S consists of 69% of the cases. This subgroup is more certain ($\lambda[2] = 3.7$) than subgroup #1 ($\lambda[1] = 1$).
- Time was found to be a significant predictor of S, but not X*.

Table 9:
Cross-tabulation of classification based on original and scale adjusted 4-class models

Scale parameter = 1.0						
original classes	scale-adjusted classes				Total	Mean time to complete
	1*	2*	3*	4*		
1	7		1		8	4.6
2	0	1	1		2	8.0
3		2	21		23	4.4
4	13	9		0	22	4.2

Scale parameter = 3.7						
original classes	scale-adjusted classes				Total	Mean time to complete
	1*	2*	3*	4*		
1	61		0		62	7.5
2	2	48	0		49	7.5
3		1	5		6	4.2
4	0	0		4	4	8.0

Table 9 shows the correspondence between respondent classifications based on the two models. For clarity, the original classes are denoted as 1, 2, 3 and 4 while the classes from the scale-extended model are denoted using an * (1*, 2*, 3* and 4*). The comparison can be summarized as follows:

- S and X* were correlated – Classes 1 and 2 tend to be in the more certain subgroup while class 3 tends to be in the less certain subgroup.

- The random responder class (class 4) is much smaller when the scale factor is estimated, as many of those classified as random responders by the standard model become reclassified into the uncertain subgroup of classes 1 or 2.
- Relationship between individual classifications:
 - Class 1 mostly corresponds to class 1*, mean time = 7.5, more certain ($\lambda = 3.7$)
 - Class 2 mostly corresponds to class 2*, mean time = 7.5, more certain ($\lambda = 3.7$)
 - Class 3 mostly corresponds to class 3*, mean time = 4.4, less certain ($\lambda = 1$)
 - Class 4 mostly redistributed to 1* and 2*, mean time = 4.2, less certain ($\lambda = 1$)

Note that since classes 1 and 2 tend to have higher λ than class 3, its part-worths would be expected to be *overstated* relative to class 3. Table 10 below shows that all of the part-worths are in fact higher than the corresponding scale-adjusted values.

Especially noteworthy is the comparison of the part-worths that address price sensitivity. Under the standard LC choice model, class 3 appears to be *less* price sensitive: 0.38 vs. -0.35 differs by *less* than either 0.87 vs. -0.85 (class 1) or 0.61 vs. -0.93 (class 2). However, given the same scale value ($\lambda=1$ shown) the scale-adjusted differences suggest *greater* price sensitivity for class 3.

Table 10.
Comparison of Original vs. Scale-adjusted part-worths

	Traditional 4-class model				4-class model with scale adjustment			
	0.39	0.30	0.16	0.15	0.45	0.35	0.16	0.03
	Class1	Class2	Class3	Class4	Class1*	Class2*	Class3*	Class4*
brand								
JVC	-0.17	-0.39	-1.18	0	-0.05	-0.10	-1.16	0
RCA	-0.04	0.18	-0.38	0	-0.02	0.05	-0.24	0
Sony	0.21	0.21	1.56	0	0.06	0.05	1.40	0
size								
25" screen	-0.36	-0.28	-0.47	0	-0.10	-0.09	-0.29	0
26" screen	0.02	-0.09	0.10	0	0.00	-0.03	0.08	0
27" screen	0.34	0.37	0.37	0	0.10	0.11	0.21	0
sound								
Mono sound	-1.66	-0.83	-0.40	0	-0.50	-0.21	-0.43	0
Stereo sound	0.40	0.44	0.24	0	0.13	0.11	0.19	0
Surround sound	1.26	0.39	0.16	0	0.37	0.10	0.23	0
block								
No blockout	-0.15	-0.83	-0.25	0	-0.05	-0.24	-0.15	0
Channel blockout	0.15	0.83	0.25	0	0.05	0.24	0.15	0
pip								
No pip	-0.33	-1.05	-0.27	0	-0.11	-0.30	-0.18	0
Picture in picture	0.33	1.05	0.27	0	0.11	0.30	0.18	0
price								
\$300	0.87	0.61	0.38	0	0.26	0.17	0.39	0
\$350	0.40	0.50	0.13	0	0.12	0.14	0.13	0
\$400	-0.42	-0.17	-0.17	0	-0.11	-0.05	-0.21	0
\$450	-0.85	-0.93	-0.35	0	-0.27	-0.26	-0.32	0
mean time=	7.1	7.5	4.4	4.8	6.7	6.7	4.7	8
N =	69	52	29	26	83	61	28	4

SUMMARY AND CONCLUSIONS

Louviere and Eagle (2006) argued that in choice modeling it is not valid to compare part-worths across individuals, groups or latent classes, without removing the potential confound with the scale parameters. In this presentation we introduced new LC models that can estimate and thus remove the scale parameters, and applied these models in 2 CBC examples with real data.

In both examples the equal scale parameter assumption used to justify standard LC choice modeling was found to be violated, different scale parameters reflecting different amounts of uncertainty being present in different subgroups. Thus, in both cases the confound existed.

In example 1, 2 subgroups that differed in their scale parameters were found. While these subgroups were approximately equally distributed among the classes, ignoring these subgroups resulted in misclassifying 37% of the cases into a different class. Similar misclassification rates were found in DFactor Choice models, which provided improved fits over the more traditional models.

In example 2, the 2 scale subgroups found were correlated with the LC segments. While segments #1 and #2 appeared to be *more* price sensitive than segment #3, they tended to have higher scale factors than segment #3 which masked the actual relationship. Removal of this confound resulted in the conclusion that these segments were in fact *less* (not *more*) price-sensitive than segment #3.

Extensions of standard LC and discrete factor (DFactor) choice models are now available that can capture both scale and preference parameters, thus removing the scale confound from preference part-worth parameter estimates. We believe that application of these and related HB-like models based on CFactors (see e.g., Rabe Hesketh and Skron dal, 2004; Vermunt and Magidson, 2008) will result in improved estimates of preference part-worths and a better understanding of the effects associated with different respondent levels of uncertainty. This, in turn, can lead to improved targeting to relevant segments based on an improved understanding of segment preferences and levels of uncertainty.

REFERENCES

- Ben-Akiva and Lerman (1985). *Discrete Choice Analysis: Theory and application to travel demand*, Cambridge, MA. MIT Press.
- Huber, J., Arora, N and R. Johnson (1998). "Capturing Heterogeneity in Consumer Choices," *ART Forum*, American Marketing Association.
- Louviere, Jordan J., Thomas C. Eagle, (2006). "Confound it! That Pesky Little Scale Constant Messes Up," 2006 Sawtooth Software Conference Proceedings.
- Louviere, Hensher and Swait (2000). *Stated Choice Models*, Cambridge University Press.
- Magidson and Vermunt (2001). "Latent class factor and cluster models, bi-plots and related graphical displays" *Sociological Methodology*, 31, 223-264.
- Skrondal and Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. London: Chapman and Hall.
- Vermunt and Magidson (2008). *User's manual and Technical Guide for LG-Syntax™ – the syntax module for Latent GOLD and LG Choice 4.5*.

APPENDICES

Below we present the Equation section and the Latent Variables portion of the Variables section of the LG-syntax that define the various models described in this article. The remaining portions of the syntax contain technical and output options, as well as the definition of the variables which specify the attributes, the dependent variable, and various id variables.

APPENDIX A1: LG-SYNTAX SPECIFICATIONS FOR LC CHOICE MODELS ESTIMATED ON THE COFFEE MAKER CHOICE DATA WITH AND WITHOUT A SCALE ADJUSTMENT

A standard LC Choice Model is defined as follows:

```
variables
  latent Class nominal 5;
equations
  Class <- 1;
  choice <- brand | Class + capacity | Class + price | Class
           + filter | Class + thermos | Class;
```

Here, ‘Class’ is the name assigned to the categorical latent variable representing the 5 segments. One logit equation is for the class sizes, the other for the observed choices. Note that “| Class” means that the parameter concerned depends on Class.

A Scale-Adjusted version of the LC Choice Model may be defined as follows:

```
latent
  Class nominal 5, sClass nominal 2, scale continuous;
equations
  Class <- 1 ; sClass <- 1;
  scale <- (s) 1 | sClass;
  (0) scale;
  choice <- brand scale | Class + capacity scale | Class
           + price scale | Class + filter scale | Class
           + thermos scale | Class ;
  s[1] = 1; s[2] = +;
```

The scale factor named ‘scale’ is defined as a continuous latent variable with a mean (intercept) depending on sClasses and a residual variance equal to zero. The restrictions set the scale factor for the first sClass to 1, and insure that the scale factor for the second sClass is nonnegative.

APPENDIX A2: LG-SYNTAX SPECIFICATIONS FOR DFACTOR CHOICE MODELS ESTIMATED ON THE COFFEE MAKER CHOICE DATA WITH AND WITHOUT A SCALE ADJUSTMENT

A DFactor Choice Model is defined as follows:

```
latent
  dfac1 ordinal 2, dfac2 ordinal 2, dfac3 ordinal 2, dfac4 ordinal 2;
equations
  dfac1 <- 1; dfac2 <- 1; dfac3 <- 1; dfac4 <- 1; class <- 1;
  choice <- brand + capacity + price + filter + thermos
    + brand dfac1 + brand dfac2 + brand dfac3 + brand dfac4
    + capacity dfac1 + capacity dfac2 + capacity dfac3 + capacity dfac4
    + price dfac1 + price dfac2 + price dfac3 + price dfac4
    + filter dfac1 + filter dfac2 + filter dfac3 + filter dfac4
    + thermos dfac1 + thermos dfac2 + thermos dfac3 + thermos dfac4;
```

and a Scale-Adjusted DFactor Choice Model as

```
latent
  sClass nominal 2, dfac1 ordinal 2, dfac2 ordinal 2, dfac3 ordinal 2,
  dfac4 ordinal 2, scale continuous;
equations
  scale <- (s) 1 | sClass;
  (0) scale;
  dfac1 <- 1; dfac2 <- 1; dfac3 <- 1; dfac4 <- 1; sClass <- 1;
  choice <- brand scale + capacity scale + price scale + filter scale
    + thermos scale + brand dfac1 scale + brand dfac2 scale
    + brand dfac3 scale + brand dfac4 scale + capacity dfac1 scale
    + capacity dfac2 scale + capacity dfac3 scale
    + capacity dfac4 scale + price dfac1 scale + price dfac2 scale
    + price dfac3 scale + price dfac4 scale + filter dfac1 scale
    + filter dfac2 scale + filter dfac3 scale + filter dfac4 scale
    + thermos dfac1 scale + thermos dfac2 scale
    + thermos dfac3 scale + thermos dfac4 scale;
  s[1] = 1; s[2] = +;
```

APPENDIX B: LG-SYNTAX SPECIFICATIONS FOR 4-CLASS MODELS ESTIMATED ON THE TV CHOICE DATA

A standard 4-class model with the 4th class being restricted to be a random class and the covariate 'TimeR' affecting the classes and monotonic non-increasing price effect is obtained by:

```
latent
  Class nominal 4;
equations
  Class <- 1 + (b) TimeR;
  Choice <- (a1) brand | Class + (a2) size | Class
           + (a3) sound | Class + (a4) block | Class
           + (a5) pip | Class + (b2) price | Class;
  a1[4]=0; a2[4]=0; a3[4]=0; a4[4]=0; a5[4]=0; b2[4]=0;
  b2 = -;
```

A 4-class Scale-Adjusted model with the 4th class being restricted to be a random class and the covariate 'TimeR' affecting the scale factor and monotonic non-increasing price effect:

```
latent
  scale continuous, sClass nominal 2, Class nominal 4 ;
equations
  sClass <- 1 + (b) TimeR;
  Class <- 1;
  Class <-> sClass;
  scale <- (s) 1 | sClass;
  (0) scale;
  Choice <- (a1) brand scale | Class + (a2) size scale | Class
           + (a3) sound scale | Class + (a4) block scale | Class
           + (a5) pip scale | Class + (b2) price scale | Class;
  a1[4]=0; a2[4]=0; a3[4]=0; a4[4]=0; a5[4]=0; b2[4]=0;
  b2 = -; s[1] = 1; s[2] = +;
```

AN EMPIRICAL TEST OF ALTERNATIVE BRAND MEASUREMENT SYSTEMS

*KEITH CHRZAN
DOUG MALCOM
MARITZ RESEARCH*

BACKGROUND

Brand image research is a staple of applied marketing research, known by many names: brand equity research, brand positioning research and brand choice research to name a few.

When done well, a brand image questionnaire results in a matrix composed of ratings of multiple brands on multiple attributes. In addition, a brand study should collect a measure of the respondent's relative preference for the various brands (often a brand rating, but ideally a brand choice or brand share).

With this data matrix in hand, some of the key deliverables of a brand study include

- Perceptual maps of the brands' positions, relative to one another and relative to the attributes
- Reports of significant differences between brands on the ratings of the attributes
- Multinomial logit (MNL) choice modeling that quantifies the impact of each attribute on brand choice/share
- Quadrant analysis for each brand showing a joint plot of that brand's performance and of attribute importance

With such an analysis plan, the measures of brand position, the brand-attribute ratings, carry a heavy load: they are the raw material for all four analyses mentioned above.

Unfortunately, brand rating scales suffer from some limitations. Primary among these is the well-known brand "halo effect:" people who like a brand may give it higher ratings on all attributes, while those who dislike a brand tend to give it lower ratings on all attributes (Thorndike 1920). When respondents have this tendency to rate all of a brand's attributes similarly, the frequent result is that all the brand's attribute ratings are highly correlated. This high correlation can cause a variety of problems when running derived importance models, because it makes the effects of the individual attributes hard to separate from one another through statistical analysis. This can be (and frequently is) so severe that sign reversals occur, for example implying that perceptions of low quality or high price *increase* the likelihood of brand choice.

Add to this the fact that many respondents are "high raters" (they tend to rate all brands positively) and rating scales often are not very sensitive to differences between brands. This leads to the practical problem that the researcher may need to plan on relatively large sample sizes to detect significant between-brand differences, and to the practical problems of poorly distinguished brands if she does not.

So the use of rating scales to measure brand positions often leads to poor discrimination and a halo effect, which result in poorly identified brand positions and impoverished or incorrect assessments of the drivers of brand choice.

Probably the most commonly used attribute evaluation tasks for brand research are

- The Likert scale, a 5 point fully anchored scale ranging from strongly agree to strongly disagree
- A “pick any” task where respondents simply indicate which brands they associate with which attributes

PREVIOUS RESEARCH

Chrzan and Griffiths (2005) investigated three brand attribute measurement systems: Likert ratings, maximum difference scaling, and a comparative rating scale. They compared the measurement systems in terms of the face validity of their perceptual maps, the ability of the measurement systems to discriminate among brands and the ability of the systems to predict choices. The comparative rating scale and the maximum difference scaling both dramatically outperformed the Likert ratings, with the maximum difference scaling having an edge, in terms of brand discrimination and the comparative ratings having an edge in terms of predictive power (Chrzan and Griffiths 2005).

Whitlark and Smith (2004) test what they call “pick data,” or “pick K data.” In pick K data, the researcher has the respondent choose a fixed number K of attributes that the respondent most associates with a given brand. Whitlark and Smith recommend keeping K at about a third of the total number of attributes, so that for a study of 12 attributes, $K=3$, whereas for a study of 25 attributes K might be 8. Benefits of pick K data, Whitlark and Smith say, include that it taxes respondents less and that it produces similar brand positions in correspondence analysis of the resulting counts.

This seems like a promising idea, but a close read of the Whitlark and Smith paper reveals that their empirical comparisons do not actually use pick data at all: they transform data collected as standard attribute ratings into data that looks like pick data, but that lacks some of its crucial features. First, because of tied ratings, the transformation Whitlark and Smith perform results in no constant number K that constrains the number of attributes a given respondent “picks.” They supply no explanation of how ties are settled. More importantly, respondents never actually “pick” attributes at all, so that the cognitive process that goes into the ratings is not allowed to differ in Whitlark and Smith’s analysis from that which goes into the pseudo-pick data. Thus the pick K data results shown in the Whitlark and Smith article require several leaps of faith to accept, making the true test of pick K data below the only real evidence about the value of the method. In addition, this test of pick K data pays no attention whatsoever to the discriminating or predictive power of pick K data relative to ratings data.

Driesner and Romaniuk (2006) focus their comparison on ratings, rankings and pick any data, and they find that the pick any data works well. Unfortunately, Driesner and Romaniuk base their analyses on a small sample of 105 respondents who completed two or more waves of a three-wave research process. They conduct only univariate and bivariate analyses of their brand measures, though they note that multivariate analysis would be a good next step. Moreover, Driesner and Romaniuk devote their entire analysis to the brand positions and to weak tests like

whether top rankings correlate with top ratings: they pay no attention whatsoever to the important issues of the measures' ability to discriminate among brands or to predict brand choice.

POSSIBLE IMPROVEMENTS FOR MEASURING BRAND POSITION

Because of the heavy reliance a brand study has on measures of brand position, and because of the substantial problems that plague standard brand rating scales, an alternative could be valuable. We seek with a robust sample and a rigorously designed research plan to compare four alternative brand measurement systems.

- Comparative rating scales
- Semantic differential scales
- Pick any evaluations
- Yes/no scaling

Comparative Rating Scale

The first of the measures is a “comparative rating scale” which anecdotal evidence based on limited past experience suggests may make for a more discriminating rating scale. An example comparative rating scale might contain these five fully anchored points:

- Much better than other brands
- A little better than other brands
- About the same as other brands
- A little worse than other brands
- Much worse than other brands

Semantic Differential

Semantic differential scales anchor only the two endpoints of a scale with positive and negative descriptors and offers respondents the choice of any point along the scale to describe a brand. For an attribute like price, for instance, the endpoints might be labeled “high price” and “low price.”

Pick Any

In pick any data, a respondent simply checks all the attributes that he associates with a brand, or all brands which he thinks a given attribute describes. One can imagine several ways to collect pick any data, including

- Brand-wise pick any – for a given brand, the respondent checks all attributes he associates with the brand, leaving the other attributes blank
- Attribute-wise pick any - for a given attribute, the respondent checks all brands he associates with that attribute, leaving the other brands blank
- Matrix pick any – for a brand attributes matrix, respondent checks all cells that describe his beliefs about brand-attribute associations, leaving all other cells blank

- Yes/No pick any – because allowing respondents to leave blanks gives respondents an incentive to do so (to finish a survey more quickly) requiring either a yes or no response for each brand-attribute combination may produce better data.

In the empirical test below we include both attribute-wise pick any and Yes/No pick any.

EMPIRICAL STUDY

Planned Comparisons

Building on the more comprehensive analysis plan employed by Chrzan and Griffiths (2005), we plan to cover both different kinds of outputs from a brand study – (a) analyses concerning brand positioning and differentiation, and (b) analyses of brand choice.

Regarding analyses of brand positioning and differentiation, we compare

- Brand differentiation – the extent to which the four brand measurement systems allow us to detect differences among brands
- Face validity of perceptual maps – the credibility and consistency of the brand positions that result from each of the four measurement systems

Regarding analysis of brand choice, we have

- Goodness-of-fit for brand choice models – how well the four systems manage to predict brand choice
- Moreover, how much incremental fit does a measurement system provide over and above the halo effect?
- Face validity of the MNL model parameters – how credible are the coefficients that result from MNL choice models using the four kinds of measures of brand positioning as predictors

Research Design

During the spring of 2007, 800 members of a US-based internet research panel qualified as users of fast food burger restaurants, and they completed a web-based survey. The survey focused upon four brands of fast food restaurant (Wendy's, McDonald's, Burger King and Jack-in-the-Box) and on 12 attributes suggested by past experience to be significant drivers of fast food brand choice. All respondents also reported brand usage (both brand share and brand used most often) and various demographics. Respondents qualified to complete the survey if they reported dining experience with at least three of the four brands.

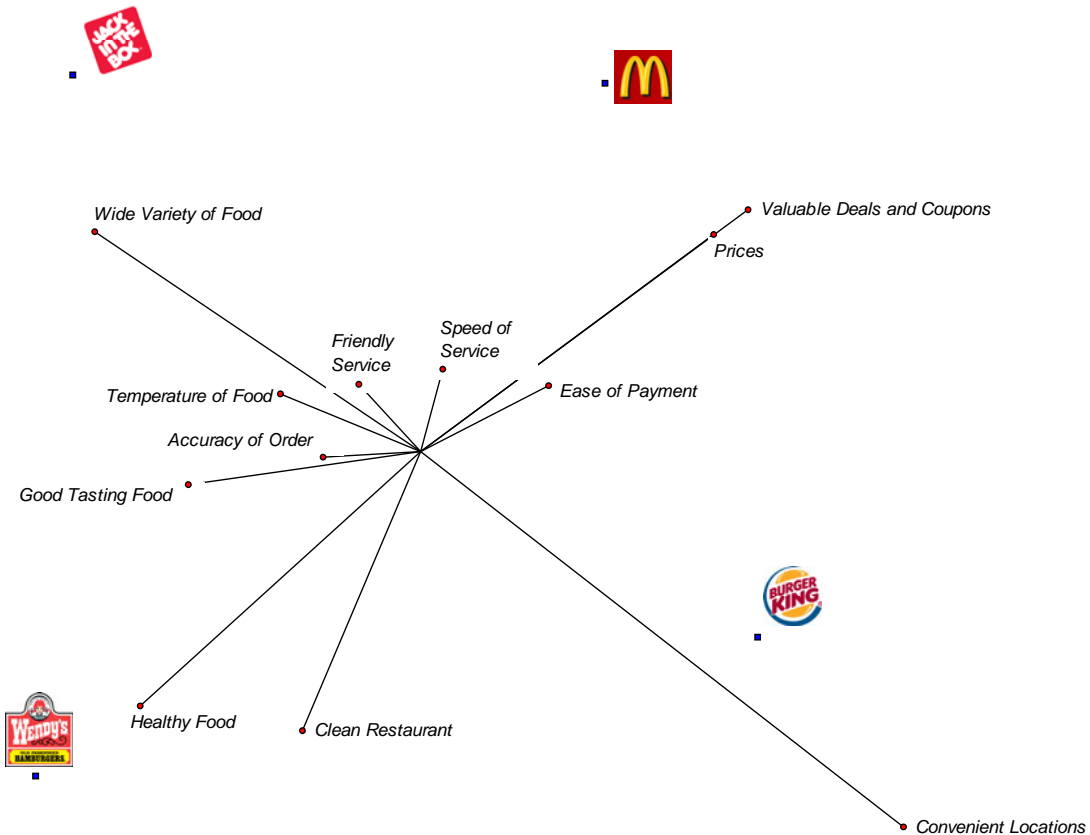
Respondents were randomly assigned so that they completed each of the four brand measurement systems.

Face Validity of Perceptual Maps

Assessing the face validity of the scales was done with perceptual maps. All maps were performed as bi-plots using doubled-centered data, since this technique is capable of producing maps with both metric and non-metric data. By examining the maps, we should be able to determine how well these scales perform at providing results important for positioning conclusions and recommendations.

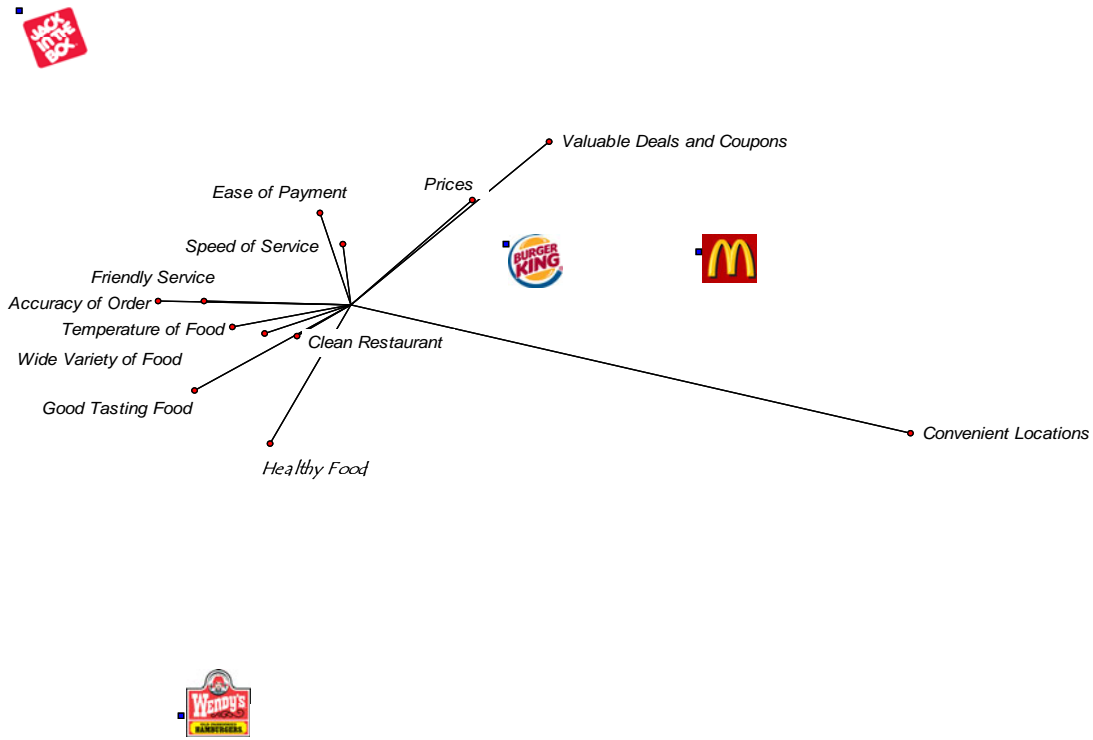
Comparative scale bi-plot.

The comparative scale map shows McDonald's associated with *valuable deals and coupons* and *price*, Jack-in-the-Box associated most with *wide variety of food*, Wendy's with *healthy food and clean*, and Burger King on *convenient locations*. All four brands have distinct positions.



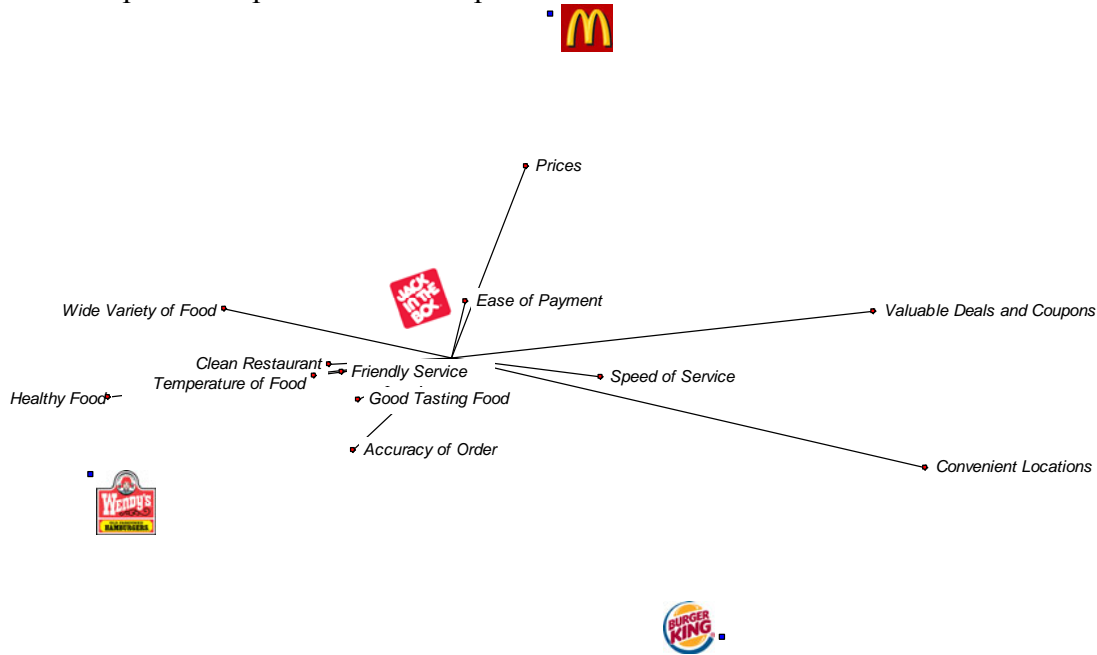
Semantic differential bi-plot.

The semantic differential scale map shows McDonald's associated with *valuable deals and coupons*, *prices* and *convenient locations*, Jack-in-the-Box associated most with *friendly service*, *ease of payment*, Wendy's with *healthy food* and *good tasting*, while Burger King occupies an interior position as a less excellent version of McDonalds.



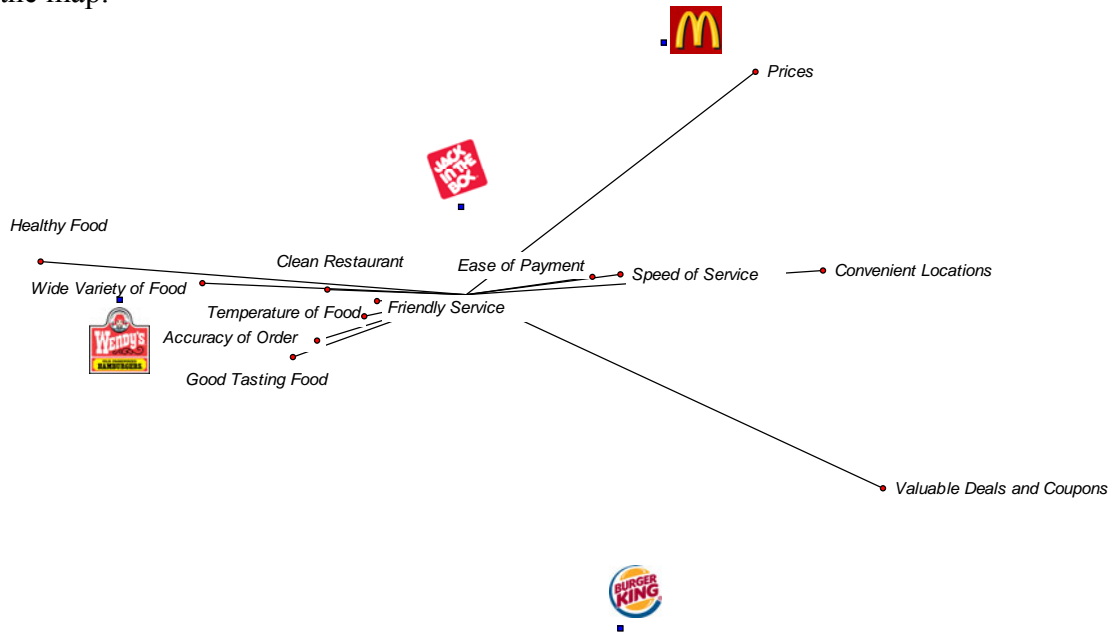
Yes/No bi-plot.

The Yes/No map shows McDonald's associated with *prices*, Wendy's with *healthy food* and *wide variety of food*, and Burger King on *convenient locations*. Jack-in-the-Box resides in the non-descript middle portion of the map.



Pick any bi-plot.

The pick any scale map shows McDonald's associated with *prices* and *convenient locations*, Wendy's with *healthy food*, *wide variety of food* and *good tasting*, and Burger King on *valuable deals and coupons* and *convenient locations*. Jack-in-the-Box again resides toward the center of the map.



Overall, there are many similarities in the results of the maps. All maps appear to do a reasonable job displaying the relationships between these attributes for these brands, with no obviously stronger solution apparent. As with any map, beauty is somewhat in the eye of the beholder on these maps, but, in the end, we think that all scales provide appropriate solutions for making brand positioning conclusions and recommendations. The maps do tell slightly different stories, however, especially regarding the position of Jack-in-the-Box and, to a lesser extent, Burger King. Because the four maps do tell slightly different stories, we do not conclude as previous authors did that the different scaling methods identify similar brand positions. We need to look at brand differentiation and prediction to make substantive judgments about the preference for some scales over others at this point.





Brand Differentiation

In order to assess how well these scales tease apart differences across brands, we performed appropriate statistical tests for the data, chi-square tests for the non-metric data and ANOVAs and T-tests for the metric scale data. The tests were run two ways, once on the raw data to understand what differences are detected across brands when the halo is not removed. These tests were chi-square tests for the non-metric data and ANOVA for the metric data performed across brands. For the second set of tests we wanted to examine the double-centered data for differences and performed cellwise tests, chi-squares for the non-metric data and t-tests for the metric data, since the data are essentially cell deviations from an expected value.

When viewing these results, we examine the findings on the raw data side-by-side with data transformed to be free of the halo effect. One way to remove the halo effect is to use a particular double centering method to eliminate it (Dillon, Mulani and Frederick 1984). The column on the right hand side of the following tables contains the relevant test statistic (F for ratings, ρ^2 for pick data) for each attribute, as well as average p-value for all tests for the brand for comparison. The majority of the table reports the cellwise test values for each attribute by brand. Non-significant results for these results have been replaced with a hyphen ('-').





Yes/No brand differentiation.

The statistical tests for the Yes/No scale show that about half the attributes are significant across these brands using the raw data with seven of the twelve attributes having a significant chi-square value (average p-value for all tests of .094). However, very few differences remain when the halo effect is removed.

					X ² With Halo Average p-value: .094
Speed of Service	-	-	-	-	19.6
Prices	-	-	-	-	25.5
Convenient Locations	2.2	-	-	-2.8	97.0
Healthy Food	-	-2.1	2.9	-	36.9
Good tasting	-	-	-	-	-
Wide Variety	-	-	-	-	19.8
Clean	-	-	-	-	-
Friendly Service	-	-	-	-	-
Deals & Coupons	2.2	-	-1.8	-	37.5
Accuracy of Order	-	-	-	-	-
Temperature of Food	-	-	-	-	-
Ease of Payment	-	-	-	-	9.1





Pick any brand differentiation.

The Pick Any statistical tests show massive levels of significance on the raw data with all tests significant and an average p-value of .0003. When the tests are run on the double-centered data, the majority of the significant differences disappear leaving only 12 of the 48 cellwise tests as significant. It is clear that the Pick Any scale discrimination comes largely from the halo effect.

					X ² With Halo Average p-value: .0003
Speed of Service	-	-	-	-	40.5
Prices	3.2	-	-1.9	-	76.6
Convenient Locations	2.6	-	-2.3	-	108.2
Healthy Food	-	-2.5	3.8	-	67.0
Good tasting	-2.0	-	-	-	21.0
Wide Variety	-	-	2.0	-	44.7
Clean	-	-	-	-	34.3
Friendly Service	-	-	-	-	14.9
Deals & Coupons	1.7	2.1	-2.7	-1.9	53.7
Accuracy of Order	-	-	-	-	17.2
Temperature of Food	-	-	-	-	16.3
Ease of Payment	-	-	-	-	38.7

Semantic differential brand differentiation.

The statistical tests for the semantic differential scale show nine of twelve tests on the raw data as significant, and nearly half of the cellwise tests on the double-centered data are significant as well (19 of the 48). Semantic differential is doing a better job of identifying differences than the nominal scales.

					ANOVA F With Halo Average p-value: .048
Speed of Service	-	-	-	-	-
Prices	.14	-	-.16	-	5.4
Convenient Locations	.35	.12	-.17	-.62	43.4
Healthy Food	-	-.11	.18	-	8.3
Good tasting	-.17	-	.16	-	5.6
Wide Variety	-	-.09	-	-	3.7
Clean	-	-	-	-	3.1
Friendly Service	-.09	-	-	.14	-
Deals & Coupons	.13	.15	-.27	-	9.0
Accuracy of Order	-.19	-	-	.16	3.0
Temperature of Food	-.14	-	-	-	2.7
Ease of Payment	-	-	-	-	-

Comparative scale brand differentiation.

The Comparative Scale statistical tests show ten of the twelve attributes with significant F-values (average p-value of .03) on the raw data. When the halo is removed from the data, there are 25 of the 48 cellwise tests showing significance.

					ANOVA F With Halo Average p-value: .030
Speed of Service	.13	-.08	-	-	6.6
Prices	.26	-	-.25	-	10.4
Convenient Locations	.45	-	-.32	-.34	31.6
Healthy Food	-	-.15	.23	-	14.4
Good tasting	-.30	-	.19	-	10.2
Wide Variety	-.12	-.15	.21	.15	12.5
Clean	-	-	.12	-	6.6
Friendly Service	-.10	-	-	-	-
Deals & Coupons	.20	-	-.27	-	7.5
Accuracy of Order	-.18	-	-	-	3.6
Temperature of Food	-.23	.08	.11	.12	5.2
Ease of Payment	-	-	-.10	-	-

Overall, the discriminating power of the four scales varies quite a bit. Yes/No really doesn't discriminate brands well on the raw data or on double-centered data. Pick Any appears to be very powerful in discriminating brands, when the data are not standardized at all, but almost all of its capacity to discriminate is lost when the halo effect is removed. Semantic Differential and Comparative ratings scales both detect significant differences well on the raw data, but more importantly do not lose their discriminating power when halo is removed. Comparative Scale appears to be particularly robust in this regard.

RESULTS FROM BRAND CHOICE MODELS

The remaining comparisons shift attention away from brand position and onto the ability of the four measurement methods to support multinomial logit (MNL) brand choice modeling. Using the four systems in turn, we can predict respondent-reported share of burger dining and brand used most often. Since both measures (share and brand used most often) lead to the same conclusions, the results below are for the brand used most modeling.

The four measurement systems produced these four vectors of MNL coefficients:

<u>Attribute</u>	<u>Semantic Differential</u>	<u>Comparative Rating</u>	<u>Yes/No</u>	<u>Pick Any</u>
Fast	.43	-	1.05	-
Price	.48	.39	.95	-
Location	.80	1.04	2.03	1.74
Healthy	-	-	-	.72
Taste	.79	.76	1.80	.77
Variety	-	-	.88	.79
Clean	.63	-	-	-
Service	-	-	-	.85
Deals	-	.43	.61	.73
Accuracy	-	-	-	-
Temperature	-	-	-	-
Payment	-	-	-	-

All four methods identify location as the most important attribute. All but pick any identify taste as the second most important (pick any has it as fourth, behind variety and service). Pick any fails to identify price as a significant driver of share of wallet, while semantic differential fails to identify deals as a significant driver. Only yes/no and semantic differential identify fast service as a significant driver and only the two binary methods show a significant coefficient for variety. Based on the face validity of the resulting coefficients, it is difficult to pick a winner among these methods.

Like ordinary least squares multiple regression, MNL produces a goodness-of-fit measure, called in this case rho-squared (ρ^2). Like R^2 from regression, a higher ρ^2 means a better fitting model, a model that explains more about brand choice than a model with lower ρ^2 .

<u>Method</u>	ρ^2
Semantic differential	.33
Comparative ratings	.34
Yes/no	.29
Pick any	.35

Of course a powerful halo effect, as seen in the brand differentiation analysis above, could be the source of some of the predictive power of the models. If we use a simple mean of a brand's attributes as a proxy for the brand's halo, we see that the majority of prediction for all four methods comes from the halo, particularly for semantic differential and pick any:

<u>Method</u>	<u>ρ^2 raw</u>	<u>ρ^2 mean</u>
Semantic differential	.33	.27
Comparative ratings	.34	.22
Yes/no	.29	.19
Pick any	.35	.29

The third column below shows the ρ^2 for each of the four measurement methods after the double centering transformation described above:

<u>Method</u>	<u>ρ^2 raw</u>	<u>ρ^2 mean</u>	<u>ρ^2 double centered</u>
Semantic differential	.33	.27	.15
Comparative ratings	.34	.22	.14
Yes/no	.29	.19	.11
Pick any	.35	.29	.05

Again, pick any appears to contribute little to prediction except the halo effect, with semantic differential and comparative ratings contributing the most non-halo information to the prediction.

While these analyses of predictive validity in light of the halo effect do not identify a definitive winner, they do point to a distinct loser: pick any seems to provide little predictive attribute information apart from the halo effect, a finding that reinforces that of the discrimination tests above. All-in-all, pick any appears to be a particularly feeble way of measuring brand attribute positions.

DISCUSSION

The four brand measurement systems produce similar, but not identical perceptual maps and MNL choice models without distinct winners and losers in terms of face validity. The quantitative tests of discriminating power and predictive power are more telling, however.

While pick any has the greatest ability to discriminate among brands, almost all of that ability comes from the halo effect. When the halo is removed, the ability of pick any to discriminate among brands, or to predict brand choice, mostly goes away.

Yes/No scaling also discriminates poorly after the halo is removed, and it also predicts brand choice poorly after a double centering transformation. We find little evidence to recommend use of yes/no scaling for measurement of functional brand attributes.

Semantic differential retains much of its discriminating power even after the removal of the halo effect, and more of its predictive power than either of the binary measurement methods.

Comparative ratings retain the greatest power to discriminate among brands after double centering. Under both methods for accounting for the impact of halo on prediction, comparative ratings seem to retain a relatively substantial amount of their predictive power.

Comparative ratings or semantic differential are probably the most robust methods for measuring brand position, of the four methods tested. All of the methods owe much of their predictive power to the brands' halo effect, and a method that solved this problem would be welcome.

REFERENCES

- Chrzan, Keith and Jeremy Griffiths (2005), "An Empirical Test of Brand-Anchored Maximum Difference Scaling," paper presented at the Design and Innovations Conference, Berlin, May 2005.
- Dillon, William R., Narendra Mulani and Donald Frederick (1984), "Removing Perceptual Bias in Product Space Analysis," *Journal of Marketing Research*, 21, 184-193.
- Driesener, Carl and Jenni Romaniuk (2006), "Comparing Methods of Brand Image Measurement," *International Journal of Market Research*, 48, 681-698.
- Thorndike, E. L. (1920), "A Constant Error in Psychological Rating," *Journal of Applied Psychology*, 4, 25-29.
- Whitlark, David B. and Scott M. Smith (2004), "Pick and Choose," *Marketing Research*, 16, 9-14.

ALTERNATIVE APPROACHES TO MAXDIFF WITH LARGE SETS OF DISPARATE ITEMS – AUGMENTED AND TAILORED MAXDIFF

PHIL HENDRIX
IMMR
STUART DRUCKER
DRUCKER ANALYTICS

INTRODUCTION

Since Maximum Difference scaling was developed by Louviere in the late 1980’s, marketing and opinion researchers have adopted and applied the methodology in a wide range of applications. In addition to its desirable scaling properties and outputs, one of the primary benefits of MaxDiff is the ease with which respondents are able to complete the fundamental task, e.g., designating the “best” and “worst” within sets of three to six items, according to some stated criterion. The “ease of use” for respondents has encouraged researchers to apply MaxDiff to an ever expanding array of issues, particularly where the items are seemingly disparate and numerous.

As an example, Moriarty (2007) recently presented results from a large study in which MaxDiff was used to examine the degree to which consumers identify with a wide range of membership and aspirational groups – items presented reflected demographic characteristics (gender, ethnicity, region, age, etc.), activities/lifestyles, products, brands, and celebrities. In the study described in this paper, MaxDiff was used to examine the importance of various benefits offered by a wide range of wireless services that are beginning to appear on new mobile devices – for instance, the ability to check e-mail, even when away from home or office; the ability to view the locations of individuals (friends, family, etc.); and many other benefits shown in the Appendix.

In addition to scaling seemingly disparate items or objects, these two studies and others like them have in common an important feature that complicates matters – namely, large sets of items. The first study contained more than 50 items, while the second contain some 40 items. As the number of items grows, the number of tasks and the amount of time required by respondents both increase, as shown below. These issues, in turn, raise concerns about respondent fatigue, data quality, and measurement error.

Tasks and Time Increase with Number of Items		
Impact	With 40 items...	With 60 items...
Design expands	24 Choice Sets	36 Choice Sets
Task gets burdensome	> 10 mins.	> 15 mins.
Assumption: MaxDiff consists of five items per task, average of three exposures per item		

This paper examines several alternative approaches to applying MaxDiff with large sets of seemingly disparate items, provides a framework and criteria for comparing the approaches, and presents empirical evidence regarding the relative performance of the approaches. Results are

presented from a study with a nationally representative sample of consumers in the U.S. in which we employed three variations on MaxDiff to scale the 40 items analyzed. More information on the items, study, and methodologies examined is provided in the next section.

MOTIVATION FOR USING MAXDIFF TO CALIBRATE CONSUMER NEEDS

Conjoint and related methods are often used to gauge consumer demand for new products and services. However, results of such studies do not necessarily reflect the significance of the underlying need or consumers' interest in a solution – individuals may desire a solution, but regard the products and services presented as inadequate or even unacceptable. For instance, consider “maps and directions” on a mobile device – individuals have significant interest in the benefits or solution, but until recently consumers have shown limited interest in such services on their mobile phones, due to concerns about ease of viewing, usability, cost, and other criteria.

Extending perspectives from early research on Means-End Chain Analysis (MECA) (Olson and Reynolds 1983), Christensen and his colleagues (2007) urge researchers and companies to focus more attention on consumers' underlying needs – in their view, products and services are hired by consumers to satisfy needs. Other researchers (Johnson 1984) have observed that consumers make tradeoffs across product categories, choosing not just within a product category (say, televisions), but across categories (television vs. movies) and indeed, across spheres (entertainment vs. health and other categories). Better understanding the relative importance of underlying needs provides a more fundamental barometer of market potential than demand for a particular solution or instantiation.

In the current study, we use MaxDiff to calibrate consumers' interest in the benefits offered by a wide range of new wireless services. In order to minimize the potentially confounding effects of availability, reliability, and cost of the services, respondents were instructed to assume that the services are all 1) reliable; 2) easy to use; and 3) free. Respondents were shown sets of the items, using the MaxDiff approaches described in the Methodology section, and asked in each to indicate the “most” and “least” desirable to have. Again, the assumptions and task were structured to measure consumers' intrinsic need for and interest in the solutions, not their willingness to buy. An illustrative MaxDiff task is shown in the Appendix, along with the Glossary provided to respondents.

IMPACT OF NUMBER OF ITEMS ON MAXDIFF

While MaxDiff was initially employed for traditional multiattribute conjoint problems, following Cohen's breakthrough paper (2003), the methodology has been extended to many other applications, including attitudinal descriptors of product benefits, brand imagery, flavors/varieties within a product line, lifestyle/psychographic measures, new product/service offerings, and others.

A cursory review of the literature on empirical use of MaxDiff in marketing research consistently presents analysis of 30 or fewer items at a time. For example, using latent class analysis, Cohen examined models with 13 and 20 product benefit attributes, while Sa Lucas (2004) studied 25 statements of personal vision. In a line optimization application, Buros (2006) applied a TURF-like procedure to MaxDiff scores for 30 coffee varieties based upon Hierarchical Bayesian estimation.

Orme's simulation study (Orme 2005) of the accuracy of HB estimation for MaxDiff suggests that to recover reasonably accurate hit rates on holdout tasks, each item should be exposed to respondents an average of three times (particularly not less than two times), and that four to six items per task (ideally, five) are recommended. These findings were further confirmed empirically by Chrzan and Patterson (2006).

Using the formula developed by Orme, a study with 30 items, taken five at a time, would result in 18 MaxDiff tasks per respondent $((30 \text{ items}/5) * 3)$. The study mentioned in the introduction with close to 60 items, taken five at a time, would result in 36 tasks. In addition to the number of tasks, descriptions provided to respondents can also add to the complexity of the task. For instance, in the current study, respondents were shown a list with a 2-3 sentence description for each of the 40 services. In the MaxDiff exercise, the choices were presented with the Service Name and a brief, 6 to 8 word description. This level of complexity is not unusual in studies of new product/service concepts.

As a result, in applications with more than 30 items, the number of tasks and complexity of the stimuli presented make it difficult to design and implement a MaxDiff study that takes advantage of the benefits of HB estimation, e.g., accuracy in capturing respondent heterogeneity and recovering predicted preferences on validation tasks.

VARIATIONS ON MAXDIFF EXAMINED

In addition to "ordinary" MaxDiff, we developed and tested two new variations intended to alleviate the burden on respondents resulting from large sets of items. In both of the new approaches, which we call Augmented MaxDiff and Tailored MaxDiff, respondents completed a Q-Sort task prior to the MaxDiff that partially revealed their preferences – the Q-Sort, explained more fully in the next section, yields a partial rank ordering of the services.

In the survey, conducted in Sept.-Oct. 2007, respondents were randomly assigned to one of three experimental cells, completing either the Ordinary, Augmented, or Tailored MaxDiff. We subsequently examined the performance of each of the MaxDiff approaches on the basis of holdout prediction, interview length, and respondents' ratings of the tasks.

New Variations on MaxDiff Examined	
In both approaches, respondents first complete a Q-Sort task. For each respondent, the Q-Sort yields a partial rank ordering of the items. Respondents then complete an Ordinary or Tailored MaxDiff.	
New Approach	Description
Augmented MaxDiff	<ul style="list-style-type: none"> ▪ Data from respondent’s Q-Sort used to augment MaxDiff data ▪ From respondent’s Q-Sort, comparisons of items across categories in the Q-Sort (e.g., items in Category 1 preferred over items in Categories 2, 3, and 4; etc.) used to construct sets of respondent-level judgments ▪ Data from Q-Sort and MaxDiff concatenated, used to estimate model
Tailored MaxDiff	<ul style="list-style-type: none"> ▪ Data from Q-sort used to customize MaxDiff ▪ For each respondent, MaxDiff choice sets constructed from a subset of items using disproportionate sampling – items preferred by respondent (as revealed in Q-Sort) shown more often, and vice versa ▪ Data from Q-Sort also used to augment data from MaxDiff exercise, as in Augmented MaxDiff ▪ Data from Q-Sort and MaxDiff concatenated, used to estimate model

We also considered testing Adaptive MaxDiff (Orme, 2006) as an alternative methodology. While attractive conceptually, the software for implementing Adaptive MaxDiff is not yet commercially available. While including Adaptive MaxDiff in the comparison is worthwhile, we limited our study to methods that could be programmed by current Web data collection suppliers, and analyzed by marketing science practitioners with “hands-on” ability to manipulate datasets in broad-application packages such as SAS or SPSS.

The Q-Sort and each of the MaxDiff approaches are described more fully below.

Q-SORT

Q-Methodology was originally developed in the 1930’s by Stephenson (McKeown and Thomas, 1988) to study individuals’ cognitive and preference structures, and in the following decades popularized in numerous fields, including political science, sociology, and marketing. While the tool has been implemented in many different ways, for different purposes, we implemented Q-Sort in the online survey with an innovative “drag-and-drop” user interface to obtain from respondents a partial rank order of the 40 items. Prior to the MaxDiff, respondents viewed a randomized list of the 40 items, and were asked to select the 10 apps that would be “most desirable to have”; then, from the remaining 30 items, the 10 least desirable; and then the next 10 most desirable. The third quartile of items consists of the remaining items. Using a drag-and-drop interface programmed by Kinesis Survey Technologies (<http://www.kinesisurvey.com/>), respondents were able to click on a particular service in the list, drag it into a box on the right side of the screen, and repeat the process until all 40 items were categorized (see illustration in the Appendix).

As a respondent completes the Q-Sort, we obtain for that respondent a partial rank ordering of the services grouped on desirability into the top quartile, second quartile, third quartile, and bottom quartile. The structure of the Q-Sort could be varied, of course – we could have

instructed respondents to identify the “top 5,” followed by the next 10, and so on, producing more of a normal distribution of partial rank orders. For our purposes, a uniform distribution was sufficient and somewhat simpler to implement than other distributions.

The information that respondents provided in the Q-Sort task was used in two different ways:

- For the Augmented MaxDiff, after the interview was completed, the Q-Sort data were concatenated to the MaxDiff data, providing additional paired comparisons that were used to estimate the parameters of the MaxDiff model, as explained below.
- For the Tailored MaxDiff, within the interview information from the respondent’s Q-Sort was also used to determine the subset of items shown to the respondent. In the spirit of tailored testing, items revealed in the Q-Sort as “more desirable” by a respondent were presented more frequently in the MaxDiff tasks for that respondent, and vice versa for less desirable items. Critical Mix (www.criticalmix.com) programmed the MaxDiff portion of the study, including the Tailored MaxDiff.

To recap, the Q-Sort was used to elicit partial rank orders, which were then used 1) to supplement the MaxDiff data post hoc (for both Tailored and Augmented MaxDiff) and 2) for the Tailored MaxDiff, to customize the MaxDiff for each respondent, given the preference structure revealed in the Q-Sort. Based on anecdotal evidence from qualitative pretests, a Q-Sort also appears to encourage respondents to acquaint themselves with the full list of items, which one would expect to lead to more reliable scaling estimates. However, a Q-Sort does add time (2-3 minutes) to overall interview length, so the question is whether the incremental benefit of incorporating outweighs the cost.

AUGMENTED MAXDIFF

The premise behind Augmented MaxDiff is to “borrow” and incorporate information from the Q-Sort into the MaxDiff in order to potentially 1) shorten and make the interview more engaging for the respondent, reducing monotony and improving reliability, and 2) improve estimation.

To extract and concatenate data from the Q-Sort into the MaxDiff, the partial rank orders provided by a respondent can be converted into a set of paired comparisons – assuming that items in the top quartile are numbered 1-10; 11-20 in the second quartile; and so on, we know that item 1 is preferred over items 11-40; similarly for items 2-10. Likewise, we know that each of the items in the second quartile is preferred over all of the items in the third and fourth quartile, and that each of the items in the third quartile is preferred over all items in the fourth quartile. The coding of these “supplemental” paired comparisons from the Q-Sort task for analysis with the MaxDiff data is explained more fully in the next section.

In the Augmented MaxDiff, respondents were shown only 16 tasks (each containing 5 items), compared to 24 tasks in the Ordinary MaxDiff. This allows us to assess whether the additional paired comparisons from the Q-Sort can make up for the loss of information resulting from fewer MaxDiff tasks (16 vs. 24).

TAILORED MAXDIFF

In the experimental design for an Ordinary MaxDiff, the task sets are constructed so that each item appears, across the sample, an equal number of times. In the absence of any additional information, the assumption is that each item is of equal interest. In Tailored MaxDiff we take a different approach – from the Q-Sort, we know that a particular respondent prefers certain items over others. Therefore, rather than allowing each of the items to appear with the same frequency, we present more preferred items more frequently, and less preferred items less frequently. In effect, for a fixed number of judgments provided by a respondent, we focus the comparisons disproportionately on preferred items, and less on less preferred items. This philosophy is consistent with the notion of Adaptive Testing, which in turn is based upon Item Response Theory (van der Linden and Hambleton 1997).

In the implementation of Tailored MaxDiff, out of the full set of 40 items, for a particular respondent 24 items were selected for that respondent using the disproportionate sampling decision rules shown below. The decision rules were selected based on judgment and are presented as a “proof of concept” rather than definitive guidelines to be applied across the board.

	Number of Items in Quartile¹	Sampling Proportion	Number of Items Selected	Total number of items in Subset
Top Quartile (Most Desired)	10	100%	10	} 24
2 nd Quartile	10	80%	8	
3 rd Quartile	10	40%	4	
Bottom Quartile (Least Desired)	10	20%	2	

¹Based on the individual's partial rank ordering in the Q-Sort

The decision to retain 24 out of the 40 items, while somewhat arbitrary, was driven by several considerations:

- Our goal of presenting fewer tasks in the Tailored MaxDiff, compared to the Ordinary MaxDiff.
- Our expectation that consumer preferences for services drop off significantly after the first 10-15 – the decision rules we adopted provide very good “coverage” of the first and second quartiles, and much less in the third and fourth. In other contexts, this assumption may or may not be warranted.
- To insure that each included item was shown on average three times (as per the recommendation in Orme's 2005 paper), which with 24 items required 18 tasks. Since respondents had already spent several minutes on the Q-Sort, we felt that 18 tasks was near the limit of what they should be asked to complete in the MaxDiff.

STUDY METHODOLOGY

A nationally representative sample of n=619 U.S. cell phone owners aged 18-64 were the subject of our study, examining 40 new wireless services (see the Appendix for a detailed description). Respondents were randomly assigned to one of the three cells, corresponding to the MaxDiff approaches examined:

- Ordinary MaxDiff (3x/average exposures per item, as per the Orme 2005 paper)
- Augmented MaxDiff: Q-Sort, followed by an Ordinary MaxDiff with the full set of 40 items (2x/average exposures per item)
- Tailored MaxDiff: Q-Sort, followed by a MaxDiff on the 24 items selected through disproportionate sampling (with 3x/average exposures for the subset)

We used the paradigm of five items per task for the Ordinary and Augmented MaxDiff Cells, but to preserve whole number divisibility in calculating number of tasks per respondent, used four items per task for the Tailored MaxDiff. While not necessarily as optimal as five per task, our belief was that the increased number of tasks would more than compensate for the “simpler” effort within each task for that cell.

We had planned to obtain samples of n=250 per analytic cell – however, due to timing and resource constraints we ended up with n=150 for the Ordinary MaxDiff cell. Due to technical issues we also had to use two separate vendors for the Q-Sort and the MaxDiff – as a result, we ended up with n = 225 in the Tailored MaxDiff cell.

Experimental Designs for the Three MaxDiff Models Examined				
MaxDiff Models	# of Items (Total/Subset)	MaxDiff Tasks per Respondent	# of Items per Set	n
Ordinary MaxDiff	40	24	5	152
Augmented MaxDiff	40	16	5	245
Tailored MaxDiff	40/24	18	4	222

Within each experimental cell, the sample was balanced across eight strata, defined by gender and four age breaks. We verified that given the final sample compositions, there were no statistical differences at a 95% level of confidence in the assignment of gender/age subgroups within each cell (i.e. the 18-29 male subgroup accounted for a similar proportion of the Ordinary MaxDiff Cell as it did in the other two cells).

EXPERIMENTAL DESIGN AND MODEL ESTIMATION

The experimental designs for the Ordinary and Augmented MaxDiffs were developed using commercially available software, controlling for level and pairwise balance of the items with optimal D-efficiency. Within interviews, items were further randomized at a task and respondent level to minimize presentation bias.

For the Tailored MaxDiff, we first generated an experimental design for 24 items. Critical Mix (the firm who programmed and hosted each of the MaxDiffs) then selected the items from

each of the quartiles, using the disproportionate sampling rules described earlier. More details are available from the authors.

In the Augmented MaxDiff cell, the Q-Sort and MaxDiff could have been presented in any order. However, for consistency with the Tailored MaxDiff cell (where the Q-Sort had to be completed prior to the MaxDiff), respondents completed the Q-Sort task first, followed by the MaxDiff.

All three cells also received three holdout tasks, each with five items, randomly drawn without replacement from the full set of 40 benefits. In each holdout task, we had respondents rank the items from most to least desirable. The holdout tasks were presented before both the Q-Sort and MaxDiff exercises, depending on the appropriate cell. Before any of the tasks, respondents were also shown and encouraged to review a glossary that described each service more fully. In the subsequent Q-Sort and MaxDiff exercises, only the names of the services were shown.

Having respondents rank items in the holdout tasks – rather than simply indicate best/worst – allowed us to accomplish two goals. First, rankings provide more data than the MaxDiff “most desirable/least desirable” comparison, allowing us to examine performance of the various methods on such measures as the average number of items predicted correctly and Spearman’s rho (correlations of actual and predicted rank). Secondly, in the psychometric spirit, we can assess how well the MaxDiff utilities predict a somewhat different, but clearly related, set of outcome measures – since rankings contain more information, they provide a stronger measure against which to evaluate model performance.

CODING THE Q-SORT/MAXDIFF DATA FUSION

One of the biggest challenges in analyzing the Augmented MaxDiff and Tailored MaxDiff cells was deciding how best to incorporate the data from the Q-Sort. Q-Sort has traditionally been viewed as a scaling technique, ultimately deriving a value, or scale point, for each “category” in the sort (i.e. the four quartiles of the sort). Although a “full” Q-Sort, with a normal distribution of items on a scale, or ranking within the extremes (1st quartile and 4th quartile, respectively) was considered, we did not implement these steps due to the existing questionnaire length.

The first approach we considered was to follow the Thurstonian tradition of paired comparisons that underlies MaxDiff and “unfold” the Q-Sort into its component pairings. Specifically, if item A is in the most desirable (1st) category, it is preferred over the 30 items in the remaining three quartiles. Likewise, if item B is in the 2nd quartile, it is preferred to the 20 items in the remaining two quartiles, and if item C is in the 3rd quartile, it is preferred to the last 10 items in the final quartile.

This would yield 60 binary judgments (30+20+10) that could be concatenated to the MaxDiff data. Taking into account both the winners and losers in the pairwise relationships, we would have 120 tasks to model from the Q-Sort section alone, to go with the 32 (16 tasks * 2) from the Augmented MaxDiff exercise. After concatenating these data to the MaxDiff data, we discovered with a test data set that the HB run times could take as long as 3-4 days.

As a result, we employed an approach suggested by Bryan Orme at Sawtooth Software. We visualized the different quartiles as “thresholds” to be either met or not met and coded the data as follows:

- In Step 1, we created 40 binary tasks, representing the 1st quartile (10 most desirable items). In each task, we had a dummy code for each item (1 for the item in question, 0 otherwise), followed by a dummy coded constant representing that quartile being tested (1=Yes, 0=No). The dependent variable was simply a binary choice: item k was above that threshold (that is, being in the 1st quartile) vs. not – put differently, the item was either one of the 10 “winners” or 30 “losers”.
- In Step 2, we created 30 binary tasks, representing the 2nd quartile (10 next desirable). In each task, we had a dummy code for each item (1 for the item in question, 0 otherwise), followed by a dummy coded constant representing that quartile being tested (1=Yes, 0=No). The dependent variable was again a binary choice: item k above the threshold (that is, being in the 2nd quartile) vs. not – meaning that it was either one of the 10 “winners” or 20 “losers” from that step.
- In Step 3, we created 20 more binary tasks, representing the 3rd quartile (10 next desirable). In each task, we had a dummy code for each item (1 for the item in question, 0 otherwise), followed by a dummy coded constant representing that quartile being tested (1=Yes, 0=No). The dependent variable was again a binary choice: item k beating that threshold (that is, being in the 3rd quartile) vs. not – meaning that it was either one of the 10 “winners” or 10 “losers” at that level.

For the remaining 10 items, no further tasks were needed because the last quartile (10 least preferred) was mathematically redundant. It should be noted that within each of these steps, the tasks were formatted as two choices per task: row 1 representing the item in question, and row 2 representing the threshold, with appropriate parameters set to zero where not applicable.

The MaxDiff portion of the data was simple – as is the custom, each task was broken into two multinomial logit tasks: one for the item being chosen as “best”, and another for the item being chosen as “worst”. For each of the k items, the coding for the “best” tasks was 1=appearing in that task, 0=not appearing, with the complement (-1=appearing in the task, 0=otherwise) for the “worst” tasks. The dependent variables were the standard multinomial choice – the item in question for that task being picked as best and worst respectively among the services included in that task.

In this way, the Q-Sort data were transformed into 90 tasks (40+30+20), coded appropriately for a Sawtooth data file, and added to the MaxDiff data through custom programming in SPSS. We also appended the custom Sawtooth headers needed for the analysis (dropping the redundant last item’s parameter from the design matrix for model identification purposes as well).

HB ESTIMATION

All models were generated using CBC/HB, using the custom covariance matrix setting recommended for convergence in Appendix K of the CBC/HB manual (Sawtooth Software, 2007). We ran 30,000 burn-in iterations, followed by 10,000 draws that were retained to produce the point estimates of utility at the respondent level, as a conservative measure of convergence – which appeared to occur well before the burn-in iterations were complete. In addition, the prior

variance parameter of the covariance matrix was tuned against the holdout tasks to minimize potential over-fitting issues. We found, though, that the standard assumption of prior variance=1.0 held up across all models.

FINDINGS

The utilities for the three models are shown below, after rescaling them to a 0 to 100 probability scale at the respondent level and averaging. The utilities were rescaled by first imputing the score of 0 for the redundant benefit, zero-centering the scores, and then converting them through the adjusted logit formula presented in the CBC/HB manual to more accurately represent the probability of choice relative to the average four other items shown in the Ordinary MaxDiff exercise (the cell used as a reference point for this paper):

Summary of Utilities Estimated by Each MaxDiff Method							
		Estimated Utility			Item Rank, based on Utilities		
Item	Statement	Ordinary MaxDiff	Augmented MaxDiff	Tailored MaxDiff	Ordinary MaxDiff	Augmented MaxDiff	Tailored MaxDiff
20	Make free domestic calls	84.4	77.6	77.3	1	1	1
13	Directions	70.0	69.8	71.1	2	2	2
33	Text Messaging	65.0	68.6	64.1	3	3	6
14	E-mail	64.7	68.0	66.5	4	4	4
6	Cameraphone	62.7	64.1	65.0	5	6	5
1	Access the Internet	61.4	65.9	67.5	6	5	3
29	Routing	59.5	49.6	54.3	7	9	8
18	Get landline reception at home/public places	59.2	54.4	47.2	8	7	9
3	Alerts (Weather)	52.1	52.5	54.8	9	8	7
2	Alerts (Information)	43.3	41.4	44.5	10	11	10
	...						
23	Mobile Tickets	12.1	11.9	13.2	36	35	33
9	Connect to Social Networking sites	9.4	9.2	7.2	37	37	38
27	Receive lessons, instructions	7.6	6.1	6.7	38	39	39
32	Store, share personal profile	7.4	7.8	8.0	39	38	37
7	Compete against others in multiplayer games	4.1	5.7	6.5	40	40	40

Upon inspection, there were few differences between cells in the relative desirability of the individual services (although the range between average utilities in the Ordinary MaxDiff model appears somewhat greater than in the Augmented or Tailored MaxDiff cells).

The top benefit, “Make free domestic calls”, was clearly the same for all three models, and the next most desirable benefits were very similar (“Directions”, “Text Messaging”, “E-mail”, “Cameraphone”) in ordering for the Ordinary and Augmented MaxDiff models, and somewhat less so between the Ordinary and Tailored cells.

Overall, nine of the top ten benefits were consistently identified by all three models. Results were most consistent between the Ordinary and Augmented approaches – results varied somewhat between the Ordinary and Tailored approaches, particularly with respect to the top and bottom benefits, and the Augmented and Tailored MaxDiff results. Our hypothesis is that this is partially due to the difference in the way the MaxDiff exercise was conducted between the cells: a reduced set of tasks based on all 40 benefits for the Augmented cell, and the disproportionate sampling of 24 of the 40 benefits for Tailored.

We also correlated the aggregate-level rescaled utilities, and found Pearson correlations of 0.984 between the Ordinary and Augmented MaxDiff utilities, and 0.976 between the Ordinary and Tailored MaxDiff results. In summary, all three models yield very similar conclusions at the aggregate level.

ABILITY OF THE MAXDIFF METHODS TO PREDICT HOLDOUT TASKS

The three holdout ranking tasks were evaluated on four measures, all computed at the individual respondent level and averaged:

- Best % - Correctly predicting the top ranked benefit
- Worst % - Correctly predicting the lowest ranked benefit
- Average number of items predicted correctly – a bottom line accounting of how the ranks were replicated
- Spearman’s rho – an overall measure of association in actual vs. predicted rank

The first two statistics captured the essence of the “MaxDiff” exercise, while the latter take into account the replication of the ranking process by the models. In all cases, higher results are preferred to lower.

Model Fit					
Cell	MaxDiff Models	Best %	Worst %	Avg # of Items Correct	Spearman’s rho
1	Ordinary MaxDiff	59.4%	55.7%	2.21	.635
2	Augmented MaxDiff	62.4%	62.2% ^{††}	2.50 ^{††}	.707
3	Tailored MaxDiff	64.7% [†]	56.6%	2.40 ^{††}	.676
^{††} Significantly different from Model 1 at p<0.05 [†] Significantly different from Model 1 at p<0.10					

Both the Augmented MaxDiff and the Tailored MaxDiff had better prediction on all four fit measures than the Ordinary MaxDiff, and all four were far ahead of the predicted best percent and worst percent one would expect due to chance (1/5, or 20%).

Significance testing between the cells, using t-tests for the average items predicted correctly and Pearson χ^2 for the percent best (most desirable benefit) and percent worst (least desirable) revealed:

- The Augmented MaxDiff (t=3.30, p<0.01) and Tailored MaxDiff (t=2.15, p<0.05) approaches outperformed Ordinary MaxDiff in recovering the average number of correctly predicted benefits, by margins of 0.29 and 0.19 respectively
- Augmented MaxDiff was better at predicting the worst percent ($\chi^2=4.90$, p<0.05) relative to Ordinary MaxDiff, with a margin of 6.5%, but the best percent was not statistically different for those cells at p<0.10 (although the actual best percent rate exceeded the Ordinary MaxDiff rate by approximately 3%)
- Tailored MaxDiff was directionally different in capturing the best percent rate than Ordinary MaxDiff ($\chi^2=3.23$, p<0.10), ahead of that cell by 5.3%. Conversely, the Worst percent rate wasn't different from what we found for the Ordinary cell
- The Augmented and Tailored MaxDiff showed similar results compared to one another, except that the Augmented Cell had better capturing of the worst percent ($\chi^2=4.50$, p<0.05)
- The Spearman's rho was higher for both of the Q-Sort/MaxDiff fusion techniques, but in line with the average number of items predicted correctly, was higher for Augmented than Tailored MaxDiff

Based on these results, it is apparent that at least for this study, both of the new variations on MaxDiff match the performance of Ordinary MaxDiff on predicting all key measures; and certainly not any worse. Both are better than MaxDiff on at least the overall ability to recover the average number of items predicted correctly, and depending on the MaxDiff-type criterion and level of confidence accepted, both methods offer at least one point of differentiation compared to the more traditional Ordinary MaxDiff.

EVALUATING THE EXERCISE: RESPONDENT FEEDBACK ABOUT THE EXERCISE

Although testing variations on MaxDiff may be of intrinsic interest to researchers, understanding how consumers view the various approaches is equally, if not more important, to gauge the future potential value of the various methods.

To determine how respondents felt about their experience with each of MaxDiff approaches, we followed the tradition of Cohen, Orme and others by using the ratings suggested by Huber et al (1991). Following completion of the MaxDiff and Q-Sort tasks, respondents rated their experience on the dimensions shown below, using a seven-point Likert scale, with 1=strongly disagree and 7=strongly agree:

Cell	Method	Attributes on which MaxDiff Exercise Rated				
		Was Enjoyable	Was Confusing	Was Easy	Made me feel like clicking answers just to get done	Allowed me to express my opinion
1	Ordinary MaxDiff	4.88	2.62	5.47	3.13	5.35
2	Augmented MaxDiff	5.20 ^{††}	2.32 [†]	5.65	2.70 ^{††}	5.72 ^{††}
3	Tailored MaxDiff	5.19 ^{††}	2.57	5.61	2.77 [†]	5.68 ^{††}
^{††} Significantly different from Model 1 at p=0.05 [†] Significantly different from Model 1 at p=0.10						

Based on these respondent attitudes, we can conclude:

- Both Augmented and Tailored MaxDiff were perceived as more enjoyable than Ordinary MaxDiff (t=2.20, p<0.05 and t=2.16, p<0.05)
- Augmented MaxDiff was least likely to make respondents feel they were clicking just to get the exercise done (t=2.30, p<0.05); but only directionally so for the Tailored MaxDiff (t=1.92, p<0.10)
- Ordinary and Tailored MaxDiff were perceived similarly as being confusing, which given the added Q-Sort for the latter cell, was a pleasant surprise. We are a bit more pleased that the Augmented MaxDiff was directionally less likely on that score (t=1.83, p<0.10)
- All methods were viewed as similarly easy
- On a more definitive measure, “allowing me to express my opinion”, respondents viewed Augmented MaxDiff to be significantly different than Ordinary MaxDiff with a high degree of confidence (t=2.67, p<0.01), and slightly less so, but still comparatively robustly, for the Tailored method (t=2.35, p=0.05)

We also conducted significance testing between the Augmented and Tailored MaxDiff cells on these measures. There were no differences outside of random variation between the new methods on the ratings.

A THOUGHT EXPERIMENT: HOW MUCH Q-SORT INFORMATION IS “ENOUGH”?

One question that came to mind in testing the new variations on MaxDiff was to get some sense of how much information from the Q-Sort is necessary as part of the data fusion. As stated previously, we used an approach with four “categories” from a uniform distribution (10/10/10/10). On one hand, given the exercise, is it really necessary to have a respondent go through the step of dividing the benefit items three times (the last category/quartile is defined by default)? Conversely, do we really need 16 MaxDiff tasks after a Q-Sort with four quartiles being defined?

Thus, we took the Augmented MaxDiff data, with the reduced set of 16 tasks (vs. the 24 for Cell 1's Ordinary MaxDiff) that were based on the full set of benefits, and looked at three additional model variants:

- Three categories: 10 most desirable/next 20 desirable/10 least desirable, mimicking the data we would get from only having respondents pick the 10 best and 10 worst benefit items by collapsing the Q-Sort data into those strata
- Two categories: 10 most desirable/30 less desirable, mimicking the data we would have from a single exercise of picking the 10 best benefits from the list
- Retaining all the Q-Sort data, but using only 8 of the 16 MaxDiff tasks

We followed the same methodology for HB model estimation that we did for the main test of the methodologies. Analyzing the performance on the three holdout tasks shows the following:

Model Fit					
Cell	Augmented MaxDiff Models	Best %	Worst %	Avg # Items Correct	Spearman's Rho
2	All four quartiles	62.4%	62.2%	2.50	.707
2a	Three Categories: 10/20/10	62.0%	59.2%††	2.45	.693
2b	Two Categories: 10/30	62.6%	56.3%††	2.35††	.682
2c	Four quartiles/8 tasks	61.1%	61.0%	2.43	.696
†† Significantly different from Model 2 at p<0.05					
† Significantly different from Model 2 at p<0.10					

Conducting another round of significance tests, this time using a McNemar test for within-sample comparisons, we found that neither collapsing the number of Q-Sort categories, nor randomly removing eight of the 16 MaxDiff tasks but using the entire Q-Sort information collected, had an adverse impact in recovering the best percent (most desirable benefit).

Conversely, there appears to be more of a penalty for reducing the amount of Q-Sort information in recovering the worst percent (least desirable benefit), with both the variant adding only three categories to the MaxDiff data and the variant adding only the simplest 10 best vs. 30 worst Q-Sort buckets to the MaxDiff showing differences at p<0.05 ($\chi^2=4.16$ and $\chi^2=12.34$). When it came to recovering the number of items ranked, the reduced amount of information of the Q-Sort did not make any difference until we only worked with two categories (10/30), where a dependent t-test found differences at p<0.01 (t=3.30).

Notably, we captured substantially the same results by augmenting only 50% of the MaxDiff portion of the data with the full Q-Sort, which shows the strength of the technique as a source of information to a relatively simple use of a Q-Sort as an analytic tool compared to a more standard full distributional sort.

DISCUSSION

On every measure studied, both the Augmented and Tailored MaxDiff approaches displayed, if not always statistically significant differences on key measures, at least no worse (and always better) performance than the standard Ordinary MaxDiff method, as applied in this study.

Augmented MaxDiff predicted the average number of items correctly more accurately than the Ordinary method at $p < 0.01$ (2.50 vs. 2.21). While the Augmented approach did not recover the best percent rate better than Ordinary MaxDiff, it did have a higher prediction rate numerically (62.4% vs. 59.4%). We found that the Augmented method appeared to predict the other end of the MaxDiff task, the worst percent rate, more accurately at $p < 0.05$ (62.2% vs. 55.7%).

Regarding the Tailored MaxDiff, it directionally offered an improvement in the best percent rate, and no statistical improvement in predicting the worst percent rate, but overall an enhanced ability to capture the ranking structure correctly compared to Ordinary MaxDiff (2.40 vs. 2.21, $p < 0.05$).

Attitudinally, both of the new MaxDiff variations offered a more enjoyable task and enhanced perception that respondents were registering opinions, and were at a minimum no more confusing than the Ordinary approach. Augmented MaxDiff had the additional benefit (at $p < 0.05$) of making respondents feel less like clicking to get through the exercise compared to Ordinary MaxDiff, but no differently than perceived among respondents in the Tailored cell.

Given the effort required in both programming and requiring respondents to go through two exercises (Q-Sort and MaxDiff) rather than one, and the relative prediction rates of Augmented and Tailored MaxDiff, there is at least preliminary evidence that both approaches offer promise in allowing researchers to add more items to a MaxDiff exercise with proper augmentation from other data sources.

If forced to choose only one approach, the Augmented MaxDiff method requires less programming effort, assuming the Q-Sort can be programmed by the research supplier. On the other hand, the directionally superior prediction of best item in the Tailored MaxDiff (64.7% vs. 59.4%, $p < 0.10$) suggests that this method should be pursued as an encouraging “adaptive” alternative to the Ordinary approach.

We conclude that fusing data from a smaller MaxDiff exercise than an Ordinary approach with other sources of information (from either an Augmented or Tailored variation on MaxDiff) may yield no worse an ability to recover respondent preferences than the standard MaxDiff presented. The only question is how this information should be constructed, and if using the Q-Sort augmentation to MaxDiff, how far one can go in trading off the number of Q-Sort categories used in the modeling compared to the number of MaxDiff tasks asked.

CAVEATS OF STUDY

We recommend that this study be considered a “proof of concept” for the alternative approaches examined. Furthermore, given the limitations noted earlier, we caution against generalizing these findings to other product categories until the following issues have been examined more fully:

- Very large item sets – while the number of items examined here (40) is moderately large compared to published studies of MaxDiff, the results may or may not hold in very large sets of items (e.g., as examined by Moriarty 2007).
- The disproportionate sampling weights that we employed in the Tailored MaxDiff were selected based on judgment. Other sampling weights, and the basis for defining those weights in the context of Q-Sort or other sorting methods, should be considered and tested for comparable results.
- Our decision to use equally sized (uniform) Q-Sort categories was guided by judgment. Other distributions, such as a normal distribution, triangle distribution, or logistic distribution, could be considered. In addition, in the case of the Augmented MaxDiff, the order in which the two exercises are completed should be examined.
- Q-Sort as a standalone method compared to MaxDiff should be tested more thoroughly. Results obtained by Chrzan and Golovashkina (2007) suggest that Q-Sort compares favorably to MaxDiff in capturing stated importance.
- Given the relatively small sample sizes in our study (all less than $n=250$), we did not investigate differences between the methods on any potential subgroups (i.e. heavy users, early adoptors, etc.).
- In terms of implementing the approaches, programming the Tailored MaxDiff, let alone creating the Q-Sort/MaxDiff fusion for HB or Latent Class estimation, is not trivial, but can be done with existing off the shelf software.
- MaxDiff assumes that respondents have a common origin, e.g., the “average desirability” of new wireless services is constant across individuals. As Lenk and Bacon point out (2007), this assumption may or may not be valid. Their rescaling methods should be investigated and used to assess the approaches outlined here.

FUTURE DIRECTIONS FOR RESEARCH

As applied researchers, we investigated data fusion as a solution to a potential limitation of MaxDiff in requiring two to three exposures per item for optimal individual-level results. This emerged from a requirement to model a potentially “large” set of new wireless services for new mobile devices beyond what we had thought respondents could handle in a single online survey. Based on the results obtained from our initial investigation, we conclude that Augmented MaxDiff and Tailored MaxDiff approaches could be very useful in other applications where the size of the item set may tax respondents’ ability to complete an Ordinary MaxDiff.

We intend to employ these new approaches in future studies and encourage marketing science practitioners to investigate their value in other applications, such as attitudinally-based segmentation, line optimization, portfolio management, and multinational research, as well as the new product development application that inspired this project.

APPENDIX

Q-Sort Task, using Drag and Drop UI

Part 1 of 3:

From the list below, select the **10** services for your cell phone that are **MOST** desirable to you by clicking on a service and, while holding down the mouse button, "dragging and dropping" it from the box on the left to the box on the right.

Note: If you wish to remove a service from your "top 10," just drag and drop the service back into the box on the left. You may also view the full description by clicking on the Glossary button.

[Click here to view Glossary](#)

Access the Internet
Alerts (Information)* - Receive critical information alerts
Alerts (Weather)* - Receive notification that severe weather is approaching
BusinessFinder* - Electronic Yellow Pages on your Phone
Camcorderphone - Make movies with your phone
Cameraphone - Take photos with your phone
Compete against others in multiplayer games
Concierge - Make reservations for restaurants, hotels, transportation
Connect to Social Networking sites (Facebook, MySpace, etc.)
Watch TV on your cell phone with TIVO - pause, record, rewind programs
Your Availability - others can see your availability/best way to reach you



MaxDiff Task, Item Explanations

Which one of the following services for your cell phone would be most desirable, and which one would be least desirable?

(Assume that each service is available, reliable and easy to use, and free of charge)

Most Desirable	Table 2 of 24	Least Desirable
<input checked="" type="radio"/>	BusinessFinder* - Electronic Yellow Pages on your Phone	<input type="radio"/>
<input type="radio"/>	Watch TV with TIVO - pause, record, rewind TV programs	<input type="radio"/>
<input type="radio"/>	Text Messaging - send/receive text messages	<input checked="" type="radio"/>
<input type="radio"/>	Camcorderphone - Make movies with your phone	<input type="radio"/>
<input type="radio"/>	Store, share personal profile*	<input type="radio"/>

*Automatically detects your location and provides information specific to your location.

Glossary

1	Access the Internet	Access the internet, visit websites, upload/download files
2	Alerts (Information)* - Receive critical information	Receive alerts containing time-sensitive information (e.g., change in stock price,
9	Connect to Social Networking sites (Facebook, MySpace, etc.)	One-click access to Facebook, Myspace, Twango, other sites where you can share photos, stay connected with friends and family
10	Control appliances, lighting in your home	Control lights, appliances, temperature in your home.
11	Coupons/offers* - receive coupons and offers	Receive special offers and discount coupons from selected businesses for products and services that you use. Digital coupons/offers transmitted to your phone to redeem, simply scan digital image at checkout. Offers can be "location-based," e.g., made available when you enter a store.
13	Determine others' availability, best way to reach	Indicates if the person you wish to contact is available, and best way to reach them.

TECHNICAL APPENDIX: EXAMPLE OF COMBINING MAXDIFF AND Q-SORT DATA

To present a typical respondent's coded data for HB estimation, imagine that the respondent saw some set of items 1, 2, 3, and 4 in a particular MaxDiff task, and picked item 1 as most desirable and item 3 as least desirable. Previously, in the Q-Sort, assume that for those four items, item 1 was assigned to quartile 1, item 2 to quartile 2, item 4 to quartile 3, and the remaining item 3 fell into quartile 4 (ignoring where the other 36 items were assigned for this illustration). For these particular items, the data would look something like this:

Section	Item 1	Item 2	Item 3	Item 4	Threshold 1	Threshold 2	Threshold 3	Answer
Quartile 1 (10 most desirable)	1	0	0	0	0	0	0	1
	0	0	0	0	1	0	0	0
	0	1	0	0	0	0	0	0
	0	0	0	0	1	0	0	1
	0	0	1	0	0	0	0	0
	0	0	0	0	1	0	0	1
	0	0	0	1	0	0	0	0
	0	0	0	0	1	0	0	1
Quartile 2 (10 most desirable after Quartile 1)	0	1	0	0	0	0	0	1
	0	0	0	0	0	1	0	0
	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	1
	0	0	0	1	0	0	0	0
Quartile 3 (10 most desirable after Quartiles 1 and 2)	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	1	1
	0	0	0	1	0	0	0	1
MaxDiff Task: Most Desirable	1	0	0	0	0	0	0	1
	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	0	0	0	1	0	0	0	0
MaxDiff Task: Least Desirable	-1	0	0	0	0	0	0	0
	0	-1	0	0	0	0	0	0
	0	0	-1	0	0	0	0	1
	0	0	0	-1	0	0	0	0

In this case, this respondent was perfectly consistent in her best and worst choices in the Q-Sort and a MaxDiff task that fortuitously happened to include just those items. Note that once an item has “won”, e.g. item 1 in quartile 1, it no longer needs to have more information coded from the Q-Sort. Also, we see that item 3 “lost” in all three quartiles by falling into Quartile 4 (which is thus redundant, and not coded). The item numbers, and the non-random ordering, are purely a device for showing the paradigm. A more representative illustration of actual MaxDiff/Q-Sort data sets can be obtained from the authors.

REFERENCES

- Bacon, Lynd, Lenk, Peter, *et al.* (2007), "Making MaxDiff More Informative: Statistical Data Fusion by way of Latent Variable Modeling," 2007 Sawtooth Software Conference Proceedings, Sequim, WA, in press.
- Buros, Karen (2006), "Product Line Optimization Through Maximum Difference Scaling," 2006 Sawtooth Software Conference Proceedings, Sequim, WA.
- Cohen, Steve (2003), "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," 2003 Sawtooth Software Conference Proceedings, Sequim, WA.
- Christensen, Clayton, Scott D. Anthony, Gerald Berstell and Denise Nitterhouse (2007), "Finding the Right Job For Your Product," MIT Sloan Management Review Spring 2007 Vol. 48, No. 3.
- Chrzan, Keith and Patterson, Michael (2006), "Testing for the Optimal Number of Attributes in MaxDiff Questions," 2006 Sawtooth Software Conference Proceedings, Sequim, WA.
- Chrzan, Keith, & Golovashkina, Natalia (2006), "An empirical test of six stated importance measures." *International Journal of Market Research*. 48:6, pp. 717-740.
- Huber, Joel C., Dick Wittink, John Fiedler, and Richard Miller (1991), "An Empirical Comparison of ACA and Full Profile Judgments," 1991 Sawtooth Software Conference Proceedings, Sequim, WA.
- Johnson, Michael D. (1984), "Consumer Choice Strategies for Comparing Noncomparable Alternatives," *Journal of Consumer Research*, 1984, 11 (December), 741-753.
- McKeown, Bruce and Thomas, Dan (1988), "Q Methodology," Sage Series: Quantitative Applications in the Social Sciences, Sage Publications, Inc., Newbury Park, CA.
- Moriarty, Patrick (2007), "Mapping Consumer Identities and Relationships with Popular Culture," presented at the MSI Conference, Accelerating Market Acceptance in a Networked World, March 16, Los Angeles.
- Olson, Jerry C. and Thomas J. Reynolds (1983), "Understanding Consumers' Cognitive Structures: Implications for Advertising Strategy," in *Advertising and Consumer Psychology*, eds. Larry Percy and Arch Woodside, Lexington, MA: Lexington Books.
- Orme, Bryan K (2005), "Accuracy of HB Estimation in MaxDiff Experiments," Technical Paper available at www.sawtoothsoftware.com.
- Orme, Bryan K (2006), "Adaptive Maximum Difference Scaling," Technical Paper available at www.sawtoothsoftware.com.
- Sa Lucas, Luiz (2004), "Scale Development with MaxDiffs: A Case Study," 2004 Sawtooth Software Conference Proceedings, Sequim, WA.
- Sawtooth Software (2007), CBC/HB Manual, Sequim, WA.

PRODUCT OPTIMIZATION AS A BASIS FOR SEGMENTATION

CHRIS DIENER
LIEBERMAN RESEARCH WORLDWIDE

INTRODUCTION

A market researcher has about as many ways to segment a market as there are ways to get from a point on one side of a ball to a point on the opposite side. The key to determining the best way to segment is to understand more about the underlying business issues driving the need for segmentation. Some methods or ways align much more productively with certain business objectives than others. Segmentation by way of product optimization (“SO”), discussed in this paper, aligns best with the purpose of product development or product line management.

The SO process is fairly straightforward, as shown in Figure 1. In short, SO first involves estimating a choice or ratings-based conjoint model. That model is then built into a simulator with an optimization engine. The simulator is then used to find an optimal set of product configurations. Using a simple rule, respondents are then grouped into segments according to which of the optimal products they most prefer (e.g., all those who most prefer product “A” are then assigned to segment “A,” those who most prefer product “B” are assigned to segment “B,” and so forth).

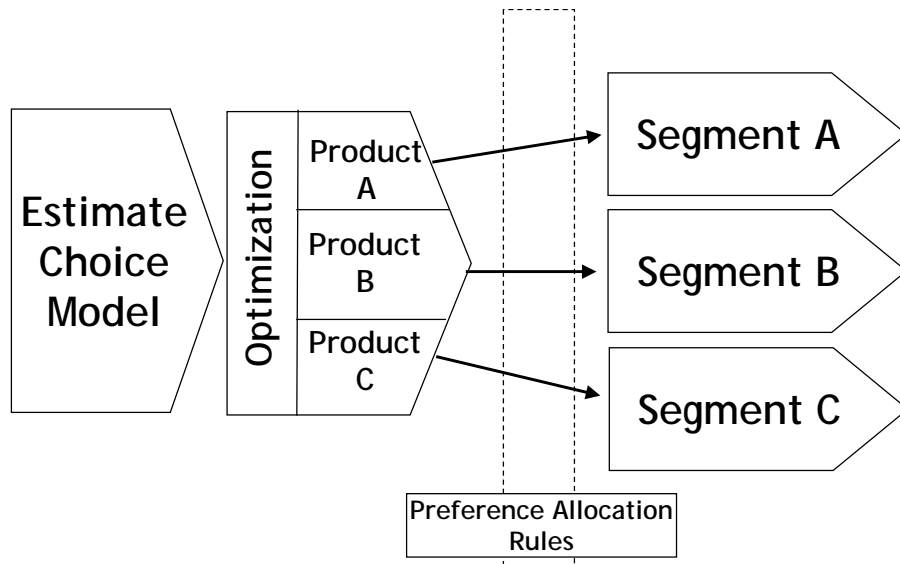


Fig. 1:
Overview of SO process

This paper will first show why SO would be used. Next, it will explain the procedure in more depth. Then, the paper will illustrate the approach using results from an empirical application. Finally, the paper will end with a discussion of specific practical issues involved with implementing the approach.

MAKING THE CASE FOR SO

Based on the company situation, segmentation may be completed for a number of different reasons. One reason for a segmentation study is to identify the best new opportunities for product development or for managing an existing product line. In this case the company has more of a product development/management objective. Resulting segments should give clear direction on new product attributes and levels. Another reason for segmentation may be to figure out how to divide the market to allow for more effective communication of existing products or services. This communication targeting has two potential applications: subjective targeting (getting the message right) and objective targeting (getting the message to the right people). Figure 2 shows how these issues can be arranged into a grid with axes of “Subjective Focus” and “Targeting Focus.”

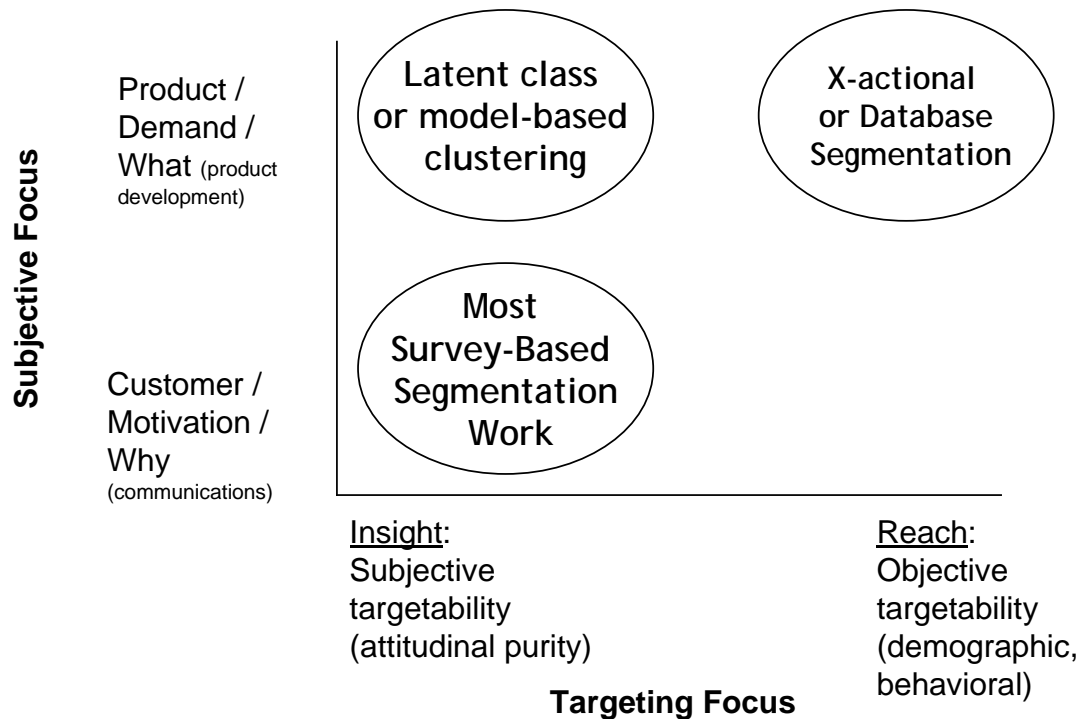


Fig. 2:
Segmentation priorities

Based on the priority in the Targeting Focus and the company situation in the “Subjective Focus” a researcher can find an appropriate methodological approach. The interior of the grid in Figure 2 contains several possible approaches. This paper concerns the Insight portion of the Targeting Focus shown in Figure 2. The paper directly addresses the spectrum between the two ends of the Subjective Focus. The two ends of the line in Figure 3 illustrate the Subjective Focus axis in Figure 2. Figure 3 diagrams the relationship between goals for resulting insight and the methods used in pursuing those goals.

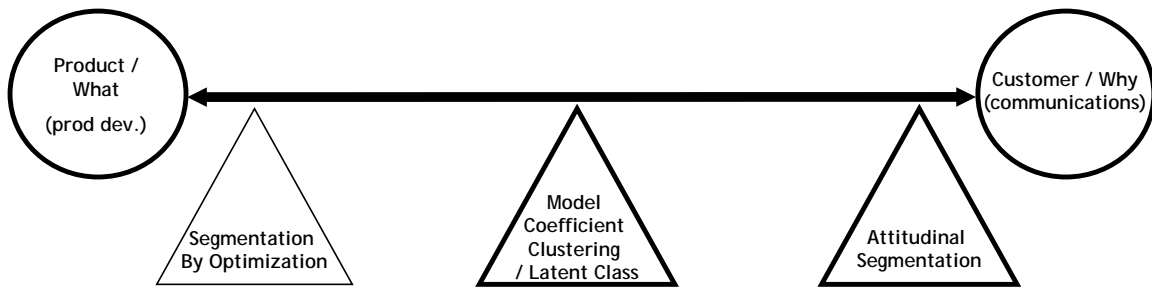


Fig. 3:
The subjective focus trade-off

In focusing on the two triangles to the right in Figure 3, the triangle closest to the right end is labeled: “Attitudinal Segmentation.” Clustering on attitude statements is a very good approach for generating segments that provide guidance on different communications strategies. However, this approach does not provide as good of insight into the specific products that each segment will be most likely to purchase. The middle triangle represents “Model Coefficient Clustering / Latent Class” (“CS”) methods for obtaining segments. These CS approaches use conjoint-type data which is much more specific to product attributes and features. The CS approach will generate segments which do much better in defining desired products. However, these segments will not discriminate as well in terms of attitudinal measures.

The triangle to the far left in Figure 3 represents the method that is the subject of this paper: “Segmentation by Optimization.” It is situated to the far left to show that this approach can be seen as a further extension of methodology toward creating better product-oriented segments and to support the notion that it is a further extension of conjoint-based segmentation approaches. The SO (“segmentation by optimization”) approach fits this position because it reduces risk or uncertainty further in terms of the goal of a product-oriented segmentation.

This uncertainty or risk differs at the opposite ends of the spectrum. For a person interested in a communications segmentation, for example, a researcher employing an SO or a CS approach (left two triangles) would be wondering whether the segments actually differ on attitudinal measures for clear messaging direction for the different segments. On the other hand, a researcher interested in segmentation to drive product development would question whether an attitudinal segmentation would produce segments that differ in terms of preferred products. This is perhaps the main concern or risk that led researchers to employ CS approaches (the middle triangle). CS approaches virtually guarantee that the segments will differ in terms of their product and feature preferences. CS has been the gold standard for product oriented segmentation.

However, a risk still remains with CS approaches. The question still remains after employing a CS approach as to whether the products which the CS segments prefer will actually work harmoniously to minimize cannibalization and maximize reach and depth of sales or profits across the market. It is this issue that SO directly addresses, taking the CS method another step by employing optimization in model simulations to define segments. The reason SO is the triangle to the far left is because the SO approach is likely to better accomplish the goals of product development or product line management segmentation than the CS methods. It does so because the optimization of the model results ensures this to be the case.

With SO, not only does the analyst know that each segment will prefer different products, but the analyst also knows that these products, when combined in the marketplace, will minimize cannibalization and maximize sales, revenues or profits. Also researchers can set up SO or CS methods that the products will be ones that the company can actually produce.

These different concerns along the spectrum are illustrated in Figure 4 which is Figure 3 overlaid with the concerns that a researcher may express as they employ a method aimed at a specific objective.

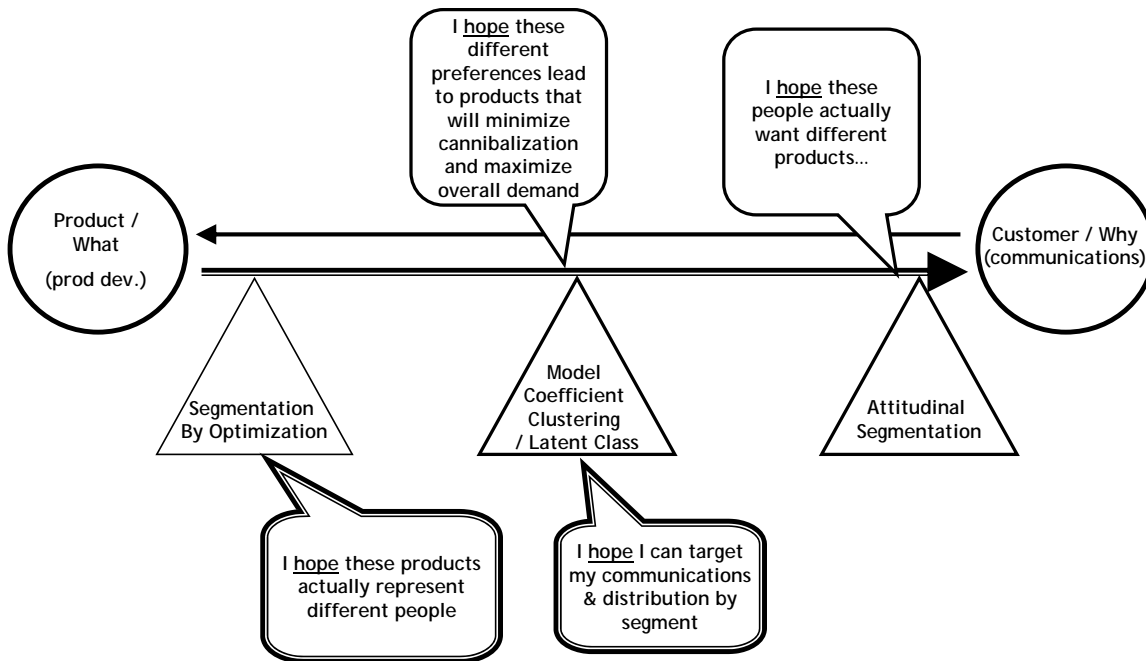


Fig. 4:

The concerns possibly expressed by researchers employing particular methods

Figures 3 and 4 represent a spectrum on which risks can be defined. Any point along the spectrum has its unique mix of risks or its unique *risk profile*. On the right side of the spectrum the risk profile is one where risk is minimized for obtaining clear insightful attitudinal discrimination between segments. But this comes at the cost of increased risk that the segments will want different products, that the products will actually be ones that the company can produce, and that these products will actually maximize the market. On the other side, the risk profile is one where risk is minimized for obtaining a segmentation which clearly guides product development. But this comes at the cost of increased risk that there will be meaningful and significant differences between the segments on attitudes.

The big question is: which risk profile is the best one? To a certain extent this depends on the situation and the objectives of or reasons for the research. The company may already have set products and are just looking for the best way to communicate about them. Or, the company may be segmenting to find “white space” for new products or to add to or rationalize their product line.

One way of assessing the value of one or the other risk profile is to ask which profile is less risky overall. Is it less risky to find an optimal set of products that the company can produce and

then be able to find a compelling attitudinal story? Or, is it less risky to find a clear attitudinal story and then hope the segments will differ on the products most preferred and that these products will be producible by the company and that they will actually work harmoniously in the marketplace?

Typically, because of the subjective nature of attitudinal segmentations, it appears more likely that an attitudinal “story” or that attitudinal meaning can be found for an SO or CS method than it is that an attitudinal segmentation will produce actionable product management information.

To summarize, SO makes a positive contribution to the market research practice for those situations in which product or product line configuration is a main objective of the segmentation.

THE APPROACH IN DETAIL

Several tools and processes are necessary for SO:

- A trade-off model
- A simulator
- An optimization routine.

First, a researcher must employ a trade-off model. This trade-off model must be able to generate individual-level utility estimates. These approaches would include choice-based conjoint estimated with Hierarchical Bayes (“HB”) methods, ratings-based conjoint or adaptive conjoint approaches.

Additionally, the modeling results must be integrated into a simulation tool. The simulation tool allows the user to input potential product configurations into the model and then generate predicted take rates. These take rates can be transformed into revenue, volume or profit forecasts.

Finally, SO requires the application of an optimization algorithm to the simulator. The optimization algorithm automates the process of inputting product configurations into the model and evaluating resulting predictions. The optimization algorithm typically has a goal that it seeks and then constraints or rules that it must use when seeking the goal. A goal would be to maximize share of preference. Or it may be to maximize revenue. A constraint would be something like only being able to put certain combinations of attribute levels together in the same concept configuration or (if you have cost information for each of the attributes and levels) making sure cost does not go over a certain level.

The optimization algorithm/simulator combination must be capable of jointly optimizing across product configurations. This means that the analyst must be able to set a goal such as maximizing the combined take rates of more than one product simultaneously in the simulator.

Using these tools, the process is rather simple:

1. Estimate a trade-off model.
2. In the simulator, optimize multiple products.

3. Assign respondents to a segment based on the product for which the respondent has the greatest demand, revenue or profitability (depending on the nature of the model and the goal of the optimization).

In the first step of the process, the analyst estimates the trade-off model. As previously stated, the model must be one that provides individual respondent-level utility estimates. Next, insert the model estimates into a simulator. The simulator can be based in a spreadsheet or be a custom program. Either way, the researcher must make sure that the simulator can accommodate an optimization algorithm or application.

After setting up the simulator the analyst will then run the optimization in which the analyst maximizes a goal subject to certain constraints. The analyst will be maximizing the goal across multiple products, arriving at a simultaneously optimized set of products. There are many optimization algorithms. The author prefers algorithms based on genetic algorithms for a number of reasons including that they work well with non-linear and discontinuous inputs and have proven to work well for the author across many projects.

Once the analyst finds the optimal products, the analyst would next assign respondents to segments. The easiest way to do this is to create segments based on products most preferred. Using this rule, respondents in a segment most prefer one of the newly optimized products. Respondents in different segments most prefer different products. The result is a set of segments, the number of which is the same as the products optimized in the simulator.

EMPIRICAL EXAMPLE

The approach will be illustrated using a dataset that has been masked to protect client confidentiality. As such, data comes from a study regarding hotels, which has 1200 respondents.

The purpose of the hotel study was to find the best product line of hotel configurations and to understand which features to emphasize to maximize communications effectiveness. The analytic process followed the pattern described above:

1. Develop and estimate a choice model
2. Gather cost information from client so that the simulator can output profit estimates
3. Use an optimization algorithm to find the best set of hotel configurations products (maximizing profit)
4. Assign individual respondents to segments based on likelihood to buy
5. Magnify attitudinal differences by “fusing” results with attitudinal data.

To illustrate the relative value of SO, several segmentation solutions were developed using different analytic approaches. First, segments were developed using the SO approach, the “SO” solution. Next, a solution was developed by clustering individual-level utilities (after utilities have been normalized to remove scale factor differences) – labeled the “CS” approach. Then, a pure attitudinal solution was created by clustering on attitudinal statements – labeled the “AT” approach. Finally, a solution was developed by combining both SO and AT information into a single set of clusters – labeled the “FS” approach. The FS approach is a hybrid between SO and AT solutions arrived at by including the SO segment membership as one of the inputs to the AT

solution – this is a simple approach used to basically illustrate the FS option. More complex and perhaps more effective approaches are discussed later in the paper. The discussion will compare these solutions (SO, CS, AT, FS) on cluster membership, the resulting product configurations, the resulting forecasts and the attitudinal differences.

Figure 5 summarizes the cluster membership overlap between the different solutions. It shows that the SO solution produces a significantly different pattern of clusters than the CS or AT solutions. It is only when trying to bring solutions together do we get more overlap, by design, in the FS approach. This shows that, somewhat surprisingly, even the CS approach does not have more overlap. This suggests that at least for this case the CS segments are not optimized for best coverage. In other datasets the author has found results similar to this, but still does not have enough experience to state this as a general finding or characteristic. However, the overall finding of low correspondence with AT solutions and dissimilar results from CS does show that the SO approach brings a new perspective that adds to what is currently done.

Transition Matrices		Hotel			
		1	2	3	4
		CS Solution			
SO Solution	1		X		
	2				
	3				
	4	X	X		X
		AT Solution			
SO Solution	1	X			
	2				
	3				
	4	X	X	X	X
		FS Solution			
SO Solution	1				X
	2	X			
	3				
	4	X	X	X	
		AT Solution			
FS Solution	1			X	
	2		X		
	3				X
	4	X			
X = Substantial Overlap					

Fig. 5:
Crosstab between segment solutions – showing overlap in segment membership

To further explore whether SO provides more useful results, products were optimized using the model within each segment of each type of solution. The question addressed with this examination is how well the solutions perform in producing segments that want different products. For product development, ideally, the different segments would want different products. This is especially the expectation with CS solutions. The differences between products in the different segments within each solution were tallied and the tallies are shown in Figure 6.

# of Differences in Attribute Levels Per Between Segments		
	Bank	Hotel
SO Solution	13	14
CS Solution	11	2
AT Solution	3	1
FS Solution	6	4

Fig. 6:

Number of differences between products optimized to specific segments

To show some variety, results of this analysis are shown also with an additional dataset from a banking study. Figure 6 shows that with the Hotel data, there were very few product differences in the optimized products for any of the segments other than the SO solution. While this same pattern held up in the Bank dataset (also shown in Figure 6), the CS solution showed a similar level of product differentiation. More generally, the SO approach will produce more differences between segments in optimally preferred products.

Finally, the solutions were compared on how much variance they explained in the utilities, attitudes, and the SO segment memberships. In general, solutions which are based on the choice model explain choice model data (utilities) well, while solutions based on attitudinal data explain that data well. This is probably no surprise to the reader. This suggests that a weakness of the SO approach is that it does not produce solutions that have a high level of attitudinal discrimination.

IMPLEMENTATION ISSUES FOR SUCCESSFUL SO APPLICATIONS

In applying SO, the practitioner should be aware of several specific weaknesses or issues with the approach:

- *Disconnect between the model estimates and attitudinal survey data.* Segments based on attitudinal data will have little discrimination on the model output and segments based on model data will have little discrimination on the attitudinal data. This means that if the researcher pursues an SO approach, the researcher should not expect to see strong attitudinal differences between the segments. However, the researcher can take the learning from SO and then apply various techniques to increase the attitudinal discrimination (such as re-clustering, discriminant analysis, or other approaches as described later in this paper). The disconnect will likely be smaller for attitudes that are more closely aligned with product features or likelihood of purchase.
- *Too little data per respondent.* If the model has a large number of attributes and levels then each respondent will see a small proportion of the required variation between those attributes and levels. This means that each respondent will “borrow” more information from the sample as a whole and this borrowing will reduce differences between respondents. This is referred to as “shrinkage” and occurs in the HB estimation process. Reduced differences between respondents will reduce the ability of the optimization algorithm to find products that are highly differentiated. Across the different studies with which the models were evaluated for this paper, each individual respondent saw between 50% and 25% of the required variation. The author would not recommend the approach for anything less than 25%, and ideally

not less than 33%. However, this recommendation is based on a limited set of experiences with the approach.

- *Value-ordered attribute levels.* If the levels of the attributes in the model are clearly ordered, such as price or a quality continuum, then the optimization algorithm will have a more difficult time finding differentiated products. This happens because on these kinds of continuum more is always better to everyone so it is unlikely that the optimization will find meaningful product differences on these attributes – such as a segment that wants a five-star friendliness of staff in the hotel room while another segment only wants three-star friendliness. If they like friendliness then they will like it all the way to five stars unless there is a cost incorporated into the optimization. So, it will almost be mandatory to include costs or other constraints like not allowing co-occurring high levels on different attributes at the same time – thus forcing the algorithm to trade them off.
- *Segment assignment complexities.* Issues may arise when developing a rule for assigning respondents to segments based on preferences of optimized products. The analyst may have products which appeal to a small number of people. If so, then the analyst may want to eliminate that product or combine it with another. Also, many respondents may prefer none of the products. If so, then either make them a separate segment or distribute them to the given segments based on the next most preferred alternative. The model may include competitive products. If so, then the analyst may want to generate some sort of threshold rule for share that allows the reassignment of respondents from the competitive product segments to the focus product segments even if the focus products were not the most preferred.
- *Business or project process integration.* In terms of project process considerations, the process of picking an optimal product set quickly in order to proceed with the segmentation may be difficult for a client (internal or external). Many clients will need to deliberate and do further internal analyses before feeling they can commit to a specific set of products. To guard against this mindset derailing a project in midstream, set the expectation from the beginning and reinforce it as the project continues that the “optimal” products that go into creating the segments do not need to be final products, but simply represent the best way of splitting the market on actionable attributes. However, if agreement can be reached on the optimal products, then that is the best situation even though it is not necessary.

OPPORTUNITIES TO IMPROVE THE PROCESS

Market researchers have only just started to harness the power of optimization algorithms as applied to this domain of modeling. This paper has earlier mentioned the value in applying some sort of “data fusion” process to increase the attitudinal discrimination between the SO segments. The author has used the NLM procedure described by Diener *et al.* (2002). Also the author suggests several other methods (e.g., Jones (2006)) for increasing attitudinal discrimination:

- Include the optimal segment membership information along with attitudinal measures in a new clustering process.

- Model individual-level share of preference, revenue or profit information from the simulator output for each optimal product as dependent variables in driver models using the attitudes as independent variables. Use these models to find the attitudes most closely associated with preference for the optimal product(s). Then use these or a subset of these attitudes to find a new set of segments.
- Run a discriminant analysis on the SO segments using attitudes and then use the predicted segment membership to define the final segments.

Another way to increase the attitudinal discrimination is to modify the optimization algorithm that you use to find the optimal products in the first place. There is great flexibility in defining the objective function of an optimization. Constraints as well can be creatively devised to allow convergence to specific types of outcomes. It may be that not only would the objective function include a goal of maximizing overall demand/revenue/profit, but would also include a benefit for increasing attitudinal discrimination. Thus the algorithm may trade off some optimality in one area to increase optimality in another. The sensitivity to one or the other aspect of the goal can be set by the researcher.

CONCLUSION

Overall the news is good from this paper that SO works as an additional approach that addresses product development needs in segmentation work. The main reasons that motivate the application of a CS approach also motivate to further pursue an SO solution to reduce the risk of producing a sub-optimal set of final products. SO ensures not only the creation of a segmentation solution that differentiates on product feature preferences, but one that differentiates in a way that leads to finding products that are maximally different to avoid cannibalization and increase market penetration or success across a set of products. In this way, SO adds a new dimension of value to segmentation. As this is a new process, a significant amount of work can be done to further refine it and leverage the approach across different types of research.

REFERENCES

- Diener *et al.* (2002). Fusing Data from Multiple Sources: Segmenting for Effective Linkages. Presented at American Marketing Association Advanced Research Techniques Forum. Vail, Colorado.
- Jones, U., Frazier, C., Murphy, C., and Wurst, J. (2006). Reverse Segmentation: An Alternative Approach. Proceedings of the 2006 Sawtooth Software Conference.

JOINT SEGMENTING CONSUMERS USING BOTH BEHAVIORAL AND ATTITUDINAL DATA

LUIZ SÁ LUCAS
IDS-INTERACTIVE DATA SYSTEMS

ABSTRACT

This paper presents a novel way to segment consumers from behavioral and attitudinal data. A fusion of Machine Learning and Marketing Research techniques, the idea is to develop a couple of segmentation and classification devices, based essentially on a distance matrix that is a combination of attitudinal, behavioral etc. distance matrices. The resulting classifier must be as simple and easy to implement as possible, and will be the resulting segmenter in the process.

INTRODUCTION

Joint Segmentation (the use of more than one dimension of basis variables for identification of segments) is usually an application of a Latent Markov approach, Means-End Chain or Positioning techniques. This paper takes another approach, very close to Reverse Segmentation, and presents a technique that segments consumers from behavioral and attitudinal data:

- the attitudinal data are obtained using traditional Marketing Research techniques, from a sample of consumers selected from the Client's database (a Telecom or an Energy Company, for example)
- the behavioral data are associated with the same respondents of the sample and come from the Client's database
- so in this case we have a fusion of Marketing Research and Data Mining / Machine Learning techniques, as we will see later.

In the course of the paper we will comment briefly on the other techniques quoted above.

In our case, the segmentation is performed in such a way that it is possible to efficiently classify the remaining records of the database (that, for sure, contain only behavioral data). Those records can be in the order of millions, for example.

The resulting classifier should be simple enough to be easily implemented by the Client, using, say, SPSS syntax, and used whenever this Client needs a classification of new records, or even update a previous classification.

BACKGROUND

There has long been interest in segmenting consumers with usual Marketing Research Segmentation tools (cluster analysis, for example), followed by the use of an independent classifier (based on discriminant analysis, decision trees, neural networks etc.) to classify the remaining records of the database (those out of the sample).

The problem with an *a posteriori* use of classifiers is that hit rates are usually low (ranging from 45% to 60%). Experts disagree about the reason for this, but one possible approach to the problem follows this line: suppose we have a very good cluster analysis solution where an element has a 75% probability of belonging to a group. Suppose also that the classifier assigns a probability of 80% for the same element to belong to this group. Assuming independence between the two devices, the probability that the two methods assign the element to the same group is $.75 * .8 = 60\%$.

Anyway we should never expect the usual approach to work perfectly. A better technique would be to use this *in tandem* approach (cluster and classify), measure the quality of the solution, and use the result of the classifier (whose quality must be also measured) as the final solution to the segmentation problem. In this sense, the classifier **is** the segmenter.

Reverse Segmentation (Jones *et al.* (2006)) addresses the same problem in a different way. The method *reverses* the process:

- first we define a classifier based on the database variables: here we define objects (sets of sample units) based on demographics, firmographics, behavioral etc. data
- then we cluster these objects

So in Reverse Segmentation, misclassification is nil. In our method we follow the more traditional sequence (clustering / classifying) and work with the single sampling units and not the objects. The misclassification is also nil.

Besides, this approach handles another usual criticism on segmentation: the lack of stability of the resulting segments. If we have a classifier, the solution to the problem is always going to be the same: that same data will always produce the same groups.

That's not the case with the usual cluster analysis techniques. They usually converge to local minima, and those will in general be different, depending on the initial solution.

Cohen (2003) even stresses the fact that sorting the same data base in different ways would induce k-means to produce different solutions...

THE PROBLEM

So the problem we posed is:

- Given two data sets:
 - Behavioral, obtained from the client's database
 - Attitudinal, obtained from a conventional marketing research survey applied to a sample of the original client's database
- Combine these two sample data sets, with different weights, so that we can define a classifier that can be applied to the whole behavioral dataset.
- This weighting scheme suggests for the method the name Weighted Distance Matrices Method - WDM
- This classifier should be simple enough to be programmed by the Client's IT team and applied efficiently to, say, a three-million records database.

- Most probably this classifier should be in the format of IF ... THEN...ELSE... rules
- So a decision tree would be very appropriate for that
- The whole process should be able to produce a stable set of clusters
- Finally the method should be able to be applied to any set of variables (nominal, ordinal, interval etc.)

JOINT SEGMENTATION

Segmenting using data from distinct data bases has been the object of several interesting works based on distinct points-of-view. We could insert WDM in this set with the following scheme:

- Simply include the (behavioral) data base variables in the standard clustering routine
- Reverse Segmentation (Jones *et al.* (2006))
 - Here we first **classify** the elements / observations into classes / grids / objects based on the values of data base variables. Then we **cluster** the objects we found in the previous step
- WDM, the method described here
 - For several weights of ‘data base’ and ‘attitudinal’ variables we first **cluster** the combined data set based on a distance matrix. Then we fit a **classifier** to the previous clustering solution.
- Concomitant Variables (Wedel and Kamakura (2000)):
 - This is a kind of a Mixture Model approach that allows for simultaneous profiling of the derived segments.
- Latent Markov Models / Latent Transition Models:
 - Ramaswamy *et al.* (1996) and others like Collins *et al.* (1997) present related models. They are also described in Wedel and Kamakura (2000)
 - In the first of these works, “the joint latent segmentation model explicitly considers potential interdependence between the bases... while extracting segments on each distinct basis”.
- Positioning Problems:
 - Buchta (1999) presents a model that works on three-way data
 - Here “consumers rate a set of brands on a set of dimensions, compare their perceptual brand profiles to their preferential profile, and make a choice”
 - See also the STUNMIX model in Wedel and Kamakura (2000).
- Consumer means-end chains (MEC):
 - Hofstede *et al.* (1999) and Perkins *et al.* (2007)

- The first of these papers states that in MEC theory, “three concepts are linked hierarchically in a cognitive structure in that product attributes yield particular benefits upon consumption, which contribute to value satisfaction”
- The model is also commented on in Wedel and Kamakura (2000).

So we see that our WDM is very close to the two other initial methods. The difference between WDM and the traditional approach is that:

- We use weights for the two sets
- We use the classifier as the segmenter:

The traditional approach needs the posterior creation of a classifier with the misclassification problem we have already commented upon

WDM has no problems of misclassification.

On the other hand, Reverse Segmentation needs an *a priori* criteria to create the classifier (the objects), but, as commented before, has also the advantage of no misclassification problems.

Properly speaking, our paper has a misleading title, since the name Joint Segmentation is usually associated in the literature to the model described by Ramaswamy *et al.* (1996).

THE WDM META-ALGORITHM

Our approach (that we could call *Weighted Distance Matrices Method – WDM*) could be described in the following way:

- Take two sets of data (the variables can be of any kind – nominal, ordinal, interval, ratio)
- For each of these two data sets, obtain a distance matrix (say $D_{\text{Behavioral}}$ and $D_{\text{Attitudinal}}$ or, more shortly, D_B and D_A)
 - This generalizes the kind of variables used in the process (nominal, ordinal etc.) the distance matrix can be based on a Gower distance (Wedel and Kamakura (2000))
- Create a summary distance matrix based on the weighted average of the two partial matrices;
 - $D = \alpha D_B + (1 - \alpha) D_A, 0 \leq \alpha \leq 1$
- Based on D for different values of α (0.1, 0.2, ..., 0.9) solve a cluster analysis problem for different values of k (number of groups), evaluating the quality of the solution (we will be back on this issue shortly).
- We have then n_α values for the weights.
- Based on the solution for each k and each α , create a classifier (also evaluating the quality of classifier, as we will comment soon)
- The $k * n_\alpha$ classifiers are the possible solutions for our problem

We can summarize the meta-algorithm with the following pseudo-code:

```
calculate the distance matrices  $D_B$  and  $D_A$ 

for k in ( $k_{\min}$  :  $k_{\max}$ )
  for  $\alpha$  in ( $\alpha_{\min}$  :  $\alpha_{\max}$ ) step  $\alpha_{\text{step}}$ 
    calculate  $D = \alpha D_B + (1 - \alpha) D_A$ 
    solve the cluster analysis problem
    fit a classifier to the solution
    evaluate the Hit Rate, Kappa and NPI for the classifier (see below)
  end of loop in  $\alpha$ 
end of loop in k
```

SOME COMMENTS ON CLUSTER ANALYSIS/SEGMENTATION

Cluster analysis/Segmentation has been the subject of some criticisms:

- “Segmentation is like slicing a single watermelon” (Rich Johnson – comment in the 2000 Sawtooth Conference)
- “Segmentation is not a stable process” (Tim Renken – comment in the 2007 ART Forum)
- “K-means will give you different solutions for different sortings of the database” (Steve Cohen (2003), quoted above)

We will illustrate Rich’s comment with our good old Fisher’s Iris data example. But first let’s examine a simple case. Let’s imagine we are a canned tea producer and we have segmented our consumers into three categories:

- Daily Drinkers – those who drink every day
- Weekly Drinkers – those who drink at least once a week
- Less than Weekly Drinkers – the rest of consumers

Now let’s imagine two consumers: one drinks every day and another drinks every day but Sundays. The first one drinks seven times a week. The other drinks six times a week. This 6-time-a-week drinker is much “closer” to the “Daily Drinker” segment than he/she is to the core of the “Weekly Drinkers”: we’re really slicing a single watermelon. But if this helps the company to develop its Marketing strategies, this is really not a problem...

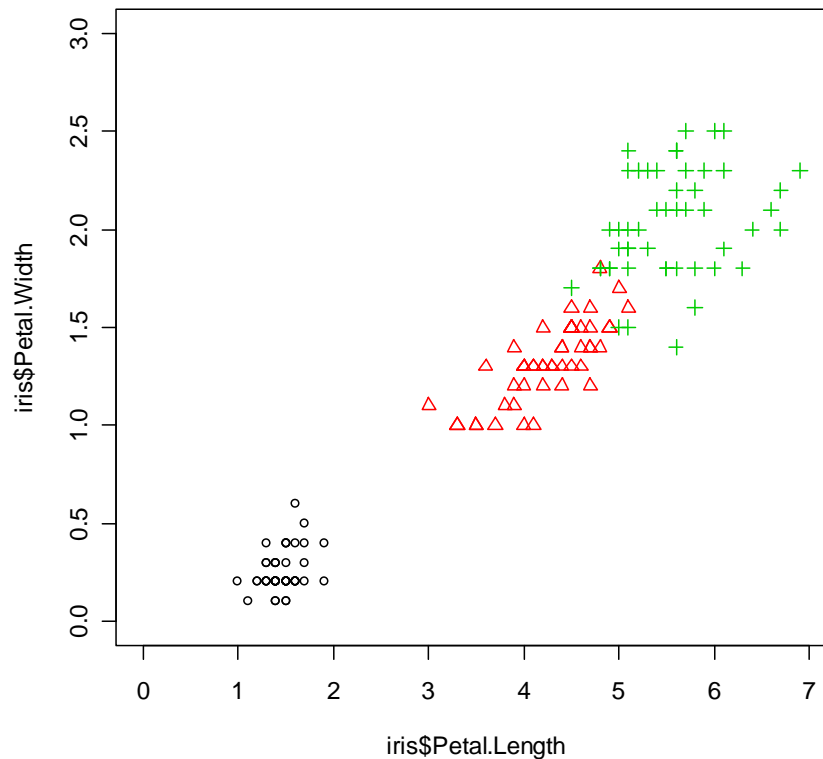
About stability: again if we solve with local minima methods the clustering problem, with the same sample we can come out with a lot of different solutions. If we want a stable solution, one possible way to achieve that is to define, as we did, one classifier that will always give the same solution to the same sample: we will always “slice the watermelon” in the same way.

Local minima solutions: several authors suggest solving the problem several times (say 10 or 20 times), keeping the best solution among the tries. We will be back to this at the conclusion of the paper.

Other comments could be made about the homogeneity of the resulting clusters and the number of them. We will illustrate this with Fisher’s Iris data.

These data have 150 observations on petal (length and width) and sepal (length and width) features, with 50 observations for three kind of Iris: *setosa*, *versicolor* and *virginica*. It's well known that the most discriminating variables are the petal ones (mainly petal length) so we present the data in Figure 1 below:

Figure 1:
Fisher's Iris data

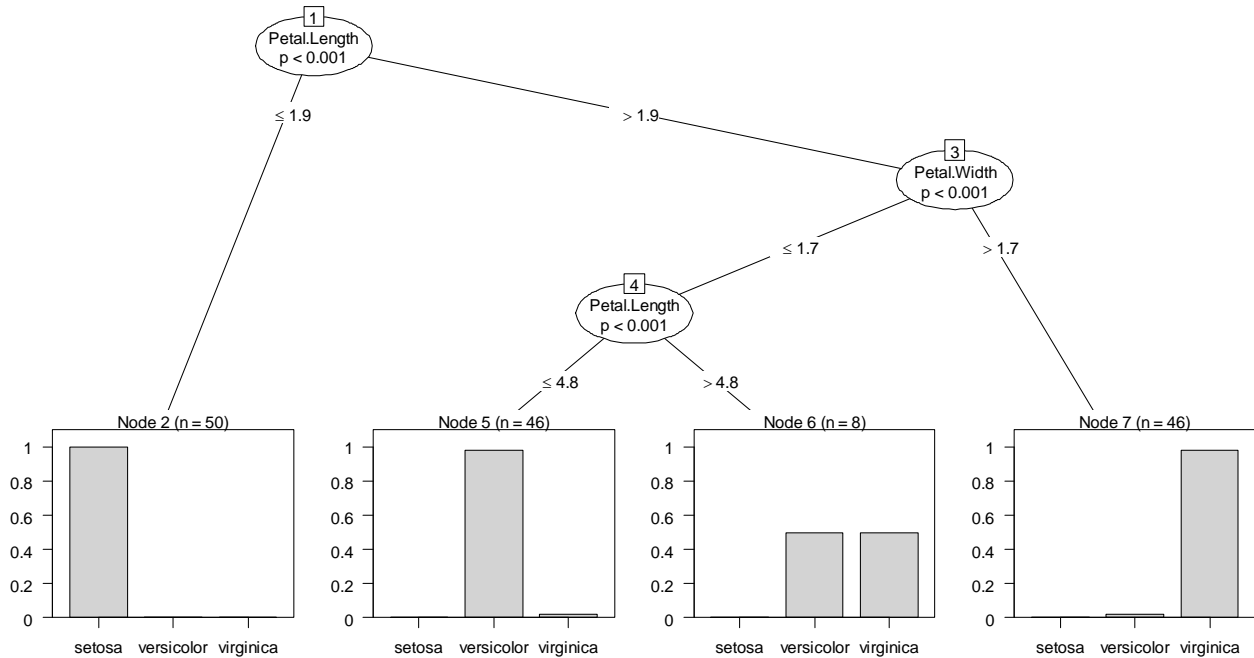


In the upper right we have the “virginicas”. At the lower left we have the “setosas”. The “versicolors” are in the middle. But looking only at the data, how many groups do we have?

- 2 groups: we can have the solution [“setosa”] and [“versicolor + virginica”] – they are well separated
- 3 groups: [“setosa”], [a group constituted only of “versicolors”] and [a group mainly made of “virginicas”, but with a mix of “versicolors + virginicas” in the lower left of the group]
- 4 groups: the same [“setosas”] and [“versicolors”] as before, [the mixed “versicolors + virginicas”] and finally [the group made only of “virginicas”].

The same effect can be seen in the decision tree below (Figure 2), constructed with a conditional inference procedure (Hothorn *et al.* (2006), see also *ctree* in *party* package in *R*):

Figure 2:
Decision Tree for Fisher's Iris data



The tree has segmented the sample into three bigger groups, one for each kind of Iris, and a fourth small mixed group as we mentioned before.

So, in lack of additional data (like the botanical ones in this case) the number of groups remains an open problem, which cannot be solved based only on statistical indicators as we show now.

We have applied a k-means algorithm to the Iris data as implemented in the *cclust* package in *R*. The same package provides us with a lot of indexes for selecting the number of groups. They are presented in Table 1 in the next page.

We are not going into the details of the indexes. They can be found in the documentation of the *cclust* package or in Weingessel *et al.* (1999). But as we can see in Table 1, each index points toward a different number of groups (different criteria, different number of groups).

Now let's solve the problem with a latent class algorithm (see Wedel and Kamakura (2000) and Leisch (2004)). Here the same local minima problem occurs, so we must solve the problem with several replications. Here we have used the *flexmix* package in *R*. Results are presented in Table 2. Latent Class seems to favor compact groups: the indexes point toward a 4-group solution.

Table 1:
Different indexes for 2 to 4 groups solutions in k- means (Iris data)

Indexes	2 groups	3 groups	4 groups
calinski	5,14E+08	5,62E+08	4,15E+08
db	4,62E+05	5,70E+05	5,51E+05
hartigan	1,24E+06	2,03E+06	2,14E+06
ratkowsky	5,46E+05	4,98E+05	4,44E+05
scott	3,10E+08	5,02E+08	5,55E+08
marriot	4,92E+11	3,06E+11	3,84E+11
ball	7,62E+07	2,63E+07	1,79E+07
trcovw	1,07E+09	2,59E+08	2,32E+08
tracew	1,54E+08	8,05E+07	7,31E+07
friedman	7,00E+06	1,96E+07	2,16E+07
rubin	7,88E+06	2,85E+07	4,04E+07
ssi	7,55E+05	9,98E+05	6,89E+05
likelihood	-1,49E+09	-1,49E+09	-1,49E+09

Table 2:
Different indexes for 2 to 4 groups solutions in latent class (Iris data)

Indexes	2 groups	3 groups	4 groups
Log Likelihood	-386.2157	-306.9185	-264.9656
AIC	806.4314	665.837	599.9312
BIC	857.6122	744.113	705.3034

There seems to be a dilemma between two criteria in segmentation: should we favor groups where an element would have a greater probability of belonging to a group or should we have compact groups? In the Iris data, greater probability seems to favor two groups and compactness seems to favor four groups.

In this work we are going to favor greater probability, so we are going to introduce what we have called the *Normalized Purity Index – NPI*. Note that a similar coefficient is used in R package *clue*.

THE NORMALIZED PURITY INDEX - NPI

One very common index for the assessment of the *impurity* of the distribution at a node in a decision tree as a classifier is the Gini index (see, for example, Venables and Ripley (2002)):

$$GiniIndex_i = 1 - \sum_k p_{ik}^2$$

Here *i* is the index of the elements and *k* is the index of the groups.

We see here that, for a given element i :

- if the probability of an element to belong to one group is 1, hence 0 (zero) for all the other groups, the index will be equal to 0 (zero)
- if the probability is the same for all groups (equal to $1/k$) then the index will be equal to $1 - k * (1/k)^2 = 1 - (1/k)$

Based on that result we can calculate a **purity** index that can be normalized, that is, will vary between 0 (equal probabilities for all groups) and 100 (probability equal to one for one group):

$$NPI_i = \frac{\sum p_{ik}^2 - (1/k)}{1 - (1/k)} * 100$$

Now we have, for a given element i :

- if the probability of an element to belong to one group is 1, hence 0 (zero) for all the other groups, the NPI will be equal to 100
- if the probability is the same for all groups (equal to $1/k$) then the NPI will be equal to 0 (zero)

Based on this idea, we can calculate not only the purity for the classification of a single element, but also for the whole clustering / segmentation. We could calculate, for example, a **pNPI75** index, that is, the percentage of elements in a segmentation that have at least a probability of 75% to belong to one of the groups.

We will see how to obtain this pNPI75 index shortly, but before that let's calculate this index for the Iris data in solutions of 2 to 4 groups.

We have obtained the 2 to 4 groups solutions with latent class (**flexmix** in **R**):

- two groups: pNPI75 = 100%
- three groups: pNPI75 = 93%
- four groups: pNPI75 = 95%

As we see, results are very close (all solutions are acceptable), but there is a little advantage for the two groups solution.

Now let's see a simple way to assess NPI for a probability of 0.75, or in general, the NPI for any probability value, for any element. We can have typical values as we illustrate for the three group case in Table 3 below:

Table 3:
NPI% values for 3-Group solutions

NPI	p(Group1)	p(Group2)	p(Group3)
100	1,000	0,000	0,000
73	0,900	0,100	0,000
72	0,900	0,050	0,050
52	0,800	0,200	0,000
49	0,800	0,100	0,100
44	0,750	0,250	0,000
39	0,750	0,125	0,125
30	0,700	0,150	0,150
23	0,650	0,175	0,175
16	0,600	0,200	0,200
11	0,550	0,225	0,225
25	0,500	0,50	0,00
6	0,500	0,250	0,250
0	0,333	0,333	0,333

As we can see, the smallest value for NPI for a probability of 0.75 for a group in a 3-group solution is 39. If we repeat the same exercise for 2, 4 to 10 and 20 groups we will have Table 4:

Table 4:
Threshold values for NPI (p=0.75)

# Groups	NPI75
2	25
3	39
4	44
5	47
6	49
7	50
8	51
9	52
10	52
20	54

So in a 5-group solution we can use, for every element, a NPI75 (NPI for p=0.75) threshold value of 47 to guarantee that this element has at least a probability of 0.75 of belonging to any group.

Based on that, we can calculate the percentage **pNPI75** of elements that have a probability of at least 0.75 of belonging to some group, that is, the purity of the solution.

In the 4-group solution we had, for the Iris data, a subset of 95% of elements that had at least 0.75 of probability of belonging to any group. It's easy to see that the same happens with the solution given by *ctree* given before.

The **pNPI75** is really a good measure of the distance among groups. We can illustrate that using another package from R called *clusterGeneration*. With this package we can simulate random clusters, controlling for the level of separation among them (Qiu and Joe (2006)). Below we have simulated three groups with different separations, and calculated the correspondent **pNPI75**:

Table 5:
Separation Index and pNPI75

Separation Index	pNPI75
0.02	13
0.04	36
0.06	45
0.08	55
0.10	65
0.15	75
0.20	84
0.30	95

HIT RATE AND KAPPA

As we deal with classifiers, it would be good if we take a look at measures of quality for them.

The **Hit Rate** is essentially the percentage of points correctly assigned by the classifier. If we are dealing with a clusterer and a classifier, we could look at the Hit Rate as a measure of agreement between the two devices.

On the other hand the **Kappa** coefficient (see Witten and Frank (2005)) or Ben-David (2006)) assesses the accuracy of the classifier discounting the predictions that could have occurred by chance.

Take for example the **confusion matrix** in Table 6 below, where we have in the rows the actual classification and, in the columns, the classes assigned by a naïve classifier that allocated all the elements to the greatest group:

Table 6:
Confusion Matrix for a Naïve Classifier

Actual Group	Group1 estimated	Group 2 estimated	Group 3 estimated
A	0	50	0
B	0	700	0
C	0	50	0

The hit rate for the table above is $700/800 = 88\%$. The Kappa coefficient is equal to 0 (zero)... So for an automated algorithm like ours the Kappa would be a good safeguard in the calculations.

Ben-David (2006) gives the equation for the Kappa:

$$K = \frac{N * \sum_{i=1}^I x_{ii} - \sum_{i=1}^I x_{i.} * x_{.i}}{N^2 - \sum_{i=1}^I x_{i.} * x_{.i}}$$

In the above equation x_{ii} is the count of cases of the main diagonal, N is the number of cases, I is the number of classes, and finally $x_{.i}$ and $x_{i.}$ are the column and row total counts, respectively.

Ideally Hit Rate and Kappa should converge, Kappa being less than or equal to the other coefficient.

As an example, let's see both coefficients in the use, as classifiers, of K-means (*cclust* in **R**), EM (*flexmix* in **R**) and a fuzzy clustering algorithm (see Wedel and Kamakura (2000)) and the *fanny* function in *cluster* package in **R**), on the Iris data, shown in Table 6:

Table 6:
Hit Rate and Kappa for the Iris Data

Method	Hit Rate	Kappa
K-means	89%	83%
Fuzzy	91%	87%
EM	94%	91%

From now on, as we need a method that can work on distance matrices, we have adopted the fuzzy clustering method.

BEHAVIORAL OR ATTITUDINAL DATA?

We will use again the Iris data to illustrate another issue that can be a dilemma.

If the behavioral data are really linked, "correlated" with the attitudinal data, then for small values of α we should get a solution that would work well for both behavioral and attitudinal data (remember the classifier is based only on behavioral variables).

But what if that doesn't occur? Table 7a below shows, for the Iris data, the application of *WDM* taking in the first set (the one that will be used as a predictor) only the variable *petal length*. The second set has the four usual sepal and petal variables.

Table 7a:
Hit Rate, Kappa and pNPI75 for Iris Data – Without Noise

# Groups	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.7$	$\alpha=0.8$	$\alpha=0.9$
Hit Rate									
2	92	92	95	95	95	97	97	100	100
3	91	93	94	94	97	100	100	100	100
4	88	89	90	94	97	99	100	100	100
Kappa									
2	83	83	88	90	90	94	94	100	100
3	87	89	91	91	96	100	100	100	100
4	83	84	86	92	96	98	100	100	100
pNPI75									
2	100	100	100	100	100	100	100	100	100
3	100	100	100	100	100	100	100	100	100
4	81	81	100	100	100	100	100	100	100

As we see, even for very small values of α we have very good solutions. But what if that is not the case? If we add noise to the variable that makes the first predictive set, “petal length + noise” will not be such a good predictor, so we can have a situation as depicted in Table 7b below:

Table 7b: Hit Rate,
Kappa and pNPI75 for Iris Data – With Noise

# Groups	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.7$	$\alpha=0.8$	$\alpha=0.9$
Hit Rate									
2	76	75	74	73	75	100	100	100	100
3	59	61	62	77	83	89	100	100	100
4	53	43	65	66	81	88	100	100	100
Kappa									
2	48	47	40	39	50	100	100	100	100
3	33	40	41	64	73	83	100	100	100
4	34	23	52	53	73	83	100	100	100
pNPI75									
2	81	81	27	27	37	100	100	100	100
3	5	9	9	74	91	100	100	100	100
4	5	11	22	29	79	100	100	100	100

We will need at least a value of $\alpha=0.5$, for a 3-group solution, or even $\alpha=0.6$, which is closer to real world problems. But we can even imagine worst cases where we will need $\alpha=0.8$ or $\alpha=0.9$...

We will always have a solution to the problem, but it may be the case that our solution is only behavioral. Here trying to insert an attitudinal point-of-view is a hopeless task. Here we should remember that Jones *et al.* (2006) and Wedel and Kamakura (2000) point out that data base variables are not usually good variables for purposes of classification.

A DISGUISED REAL WORLD EXAMPLE

Let's now turn to a real world example, naturally disguised. Table 8 below presents results for a 3 to 5 group solution for a problem with around 30 variables from the Company's database and around 40 attitudinal data. The Company's data comprise profitability, average ticket, Region, consumption etc.

Table 8:
Hit Rate, Kappa and pNPI75 for a real world example

# Groups	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.6$	$\alpha=0.7$	$\alpha=0.8$	$\alpha=0.9$
Hit Rate									
4	76	75	74	73	75	100	100	100	100
5	59	61	62	77	83	89	100	100	100
6	53	43	65	66	81	88	100	100	100
Kappa									
4	48	47	40	39	50	100	100	100	100
5	33	40	41	64	73	83	100	100	100
6	34	23	52	53	73	83	100	100	100
pNPI75									
4	81	81	27	27	37	100	100	100	100
5	5	9	9	74	91	100	100	100	100
6	5	11	22	29	79	100	100	100	100

Here $\alpha = 0.5$ for a 5-group solution would be absolutely appropriate. Figures 3 and 4 below show the Correspondence Analysis mapping for the Attitudinal and Behavioral variables, showing that based only on a Behavioral classifier we could have a very good Attitudinal differentiation.

Figure 3:
Correspondence Analysis mapping – Attitudinal

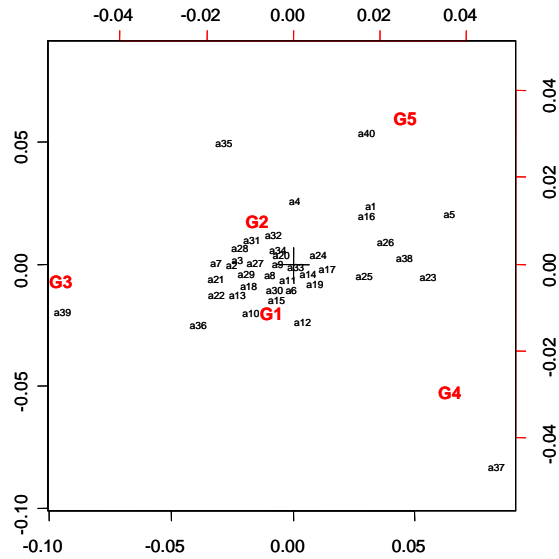
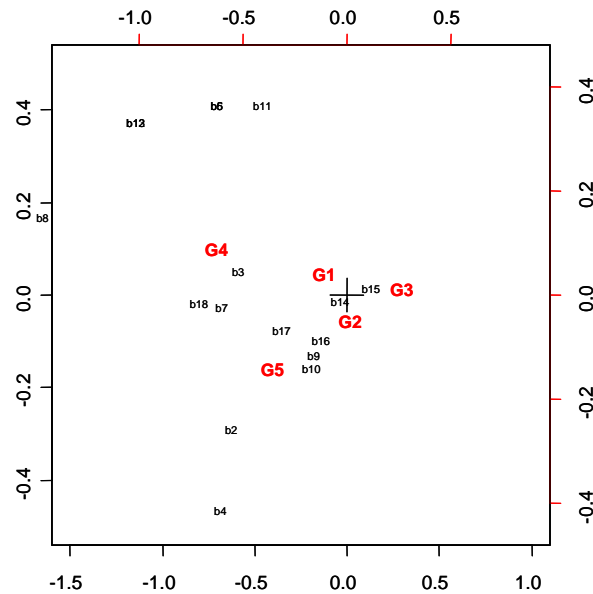


Figure 4:
Correspondence Analysis mapping – Behavioral



THE IF-THEN-ELSE RULES

We could think of another interesting application of WDM. Suppose now we have initially only one set of around 50 *needs* variables (need-based segmentation).

We could solve with cluster analysis the problem and, using some method (Random Forest, for example) assess the importance of each variable in the classification. We can then select the,

say, 15 most important variables and build a prediction set (the first distance matrix) with them. So now we have two distance matrices, and we can apply our method.

The resulting if-then-else rules could be like the ones in Figure 1 below, based on real-world data.

In Figure 2, Rule 1 states that if $v_{23} \leq 1.5$, we assign the observation to Group 1. We have 115 cases in this rule, and 87% of them belong to the original cluster analysis solution. It's also easy to see that the pNPI75 for the whole solution is 88%. We also reduced the original 50 variables to a set of 9 variables...

Figure 2:
IF-THEN-ELSE rules, for a pNPI75 of 88%

IF $v_{23} < 1.5$ ==> G1 / 87% / 115	Rule 1
ELSEIF $v_{29} \geq 1.5$	
IF $v_{34} \geq 3.5$	
IF $v_2 \geq 4.5$	
IF $v_4 < 2.5$ ==> G1 / 87% / 23	Rule 2
ELSE ==> G3 / 83% / 12	Rule 3
ELSEIF $v_{50} < 1.5$ ==> G1 / 76% / 17	Rule 4
ELSE ==> G2 / 82% / 122	Rule 5
IF $v_{18} \geq 3.5$ ==> G2 / 60% / 48	Rule 6
ELSEIF $v_{49} < 1.5$ ==> G1 / 55% / 11	Rule 7
ELSE ==> G4 / 82% / 71	Rule 8
ELSEIF $v_{50} \leq 3.5$ ==> G3 / 82% / 105	Rule 9
ELSEIF $v_{44} < 2.5$ ==> G1 / 81% / 27	Rule 10
ELSEIF $v_2 < 3.5$ ==> G2 / 69% / 23	Rule 11
ELSE ==> G3 / 61% / 26	Rule 12

IS A SINGLE DECISION TREE GOOD ENOUGH?

Another question may arise: how much are we losing when we decide for a single decision tree? Based on an artificial 3-group solution example, we applied WDM using several single and *bagging* and *boosting* ensemble methods. The methods and the corresponding **R** package are:

- A – AdaboostM1 – *RWeka*
- J – J48 – *RWeka*
- M – MultiboostAB – *RWeka*
- R – Random Forest – *randomForest*
- T – rpart – *rpart*

Witten and Frank (2005) describe these methods in general. A comparison of *bagging* and *boosting* methods is given in Rodríguez *et al.* (2006). Those authors created a method they have called Rotation Forests and they claim their technique would have a better performance in the solution in the so called *accuracy-diversity* dilemma. We have not tried it since we worked only with methods currently available in **R**.

All the methods behave well for the 3-group solution, but an interesting thing to be noted in Table 9 is the fact that the Adaboost methods worked “poorly” for the (wrong) 4 and 5-groups solutions. Although based only on a single experiment, this result suggests that at the development stage of the segmentation, the “T” solution should always be checked against an “A” and/or an “M” solution.

Table 9:
Hit Rate, Kappa and pNPI75 for several tree methods ($\alpha = 0.4$)

# Groups	A	J	M	R	T
Hit Rate					
3	100	100	100	100	100
4	87	96	88	100	97
5	63	100	63	100	100
Kappa					
3	100	100	100	100	100
4	76	92	78	100	94
5	47	100	47	100	100
pNPI75					
3	97	100	94	100	100
4	50	92	67	100	100
5	44	100	44	100	100

Another interesting set of techniques that should be considered for the classifier are the ones given in Pemberton and Powlett (2007). They could be applied if our segmentation was developed through non-scalar methods (max-diffs, for example).

CONCLUSION

We hope to have shown that WDM can be a very efficient tool to achieve the goals we have set for ourselves. Nevertheless, some points in the segmentation process have not been touched and some weaknesses should also be pointed out.

The first comment would be on *heterogeneity of use of scales* by different respondents. Here we would suggest the use of one of the following methods, all of them compatible with WDM:

- Bayesian methods (Rossi *et al.* (2005))
- MaxDiffs (Cohen *et al.* (2002, 2003a, 2003b), Chrzan (2004), Sá Lucas (2004), Bacon *et al.* (2007) and Hendrix and Drucker (2007))
- APEX method (Tang and Wiener (2006))

The second comment could be about the fact that the *use of too many variables* can cause difficulties for the cluster analysis solution or the classification problem. Here we could make the following suggestions at the cluster analysis level:

- Use a weighting variables method that can be used with mixed type variables, such as the one developed by Huang *et al.* (2005)
- Use a variable selection method like the one implemented in the R package *clustvarsel*

Alternatively one could work at the classifier level. Here the alternatives could be:

- The *ad-hoc* method we described here
- the if-the-else section above
- The “Iterative RELIEF” method created by Sun (2007)

Missing data are always a problem. Here we suggest:

- The norm, cat, EMV and Amelia packages in R
- Another interesting source of information can be the comparison of missing data methods and software made by Horton and Kleinman (2007)
- Another comparison of several methods is given in Alejandro and Pflughoeft (2007)

The main *Weakness* of the method presented here is the use of a dynamic clouds method such as k-means (fuzzy or not). Relying on methods that are strongly favorable to local minima is a dangerous procedure. Alternatively we could try:

- To solve k-means / fuzzy k-means / latent class several times (10-20 times) keeping the best solution
- Use the ant colony approach described in Kanade and Hall (2007)
- To try promising methods such as the one given by Ma *et al.* (2007)
- Graph coloring algorithms (Ulker *et al.* (2006))
- To use evolutionary algorithms (genetic algorithms, Jain *et al.* (2000))
- To use Archetypal Analysis (Elder and Pinnell (2003))
- Simulated Annealing algorithms (Jain *et al.* (2000), Wedel and Kamakura (2000), Venables and Ripley (2002))
 - The use of simulated annealing techniques in cluster analysis is not new (see, for example, Wedel and Kamakura (2000)), but these authors look at this technique as one of the most promising in the field.
- If a choice is involved, one could use Latent Class methods or Diener’s optimization model (Diener (2007)).
- To use **R** packages such as *clue* (see Retzer and Shan (2007)) and or *clusterSim* to automatically search for better clustering techniques
 - Cluster ensembles such as implemented in *clue* permit the use of several different algorithms that would finally be combined, so all the methods quoted here would be candidates ...

To conclude, we must comment that the extension of WDM to *m-way segmentations* is straightforward:

$$D = \sum_{i=1}^m \alpha_i D_i$$

where

$$\sum_{i=1}^m \alpha_i = 1$$

The predictors in the classifier may belong to the first matrix or a subset of them.

REFERENCES

- Alejandro, J. and Pflughoeft, K. (2007). Multiple Imputations as a Benchmark for Comparison within Models of Customer Satisfaction. Sawtooth Software Conference 2007 Proceedings.
- Ben-David, A. (2006). What's Wrong with Hit Ratio? IEEE Intelligent Systems, Vol 21, No. 6.
- Bacon, L., Lenk, P., Seryakova, K. and Veccia, E. (2007). Making MaxDiff More Informative: Statistical Data Fusion by way of Latent Variable Modeling. Sawtooth Software Conference 2007 Proceedings.
- Bryant, K., Windle, M. and West, S. (Eds.) The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. Washington D.C: American Psychological Association pp. 79-99.
- Buchta, C. (1999). Modeling Market Scenarios for Simulation Studies on the Joint Segmentation and Positioning Problem. Working Paper No. 59. Vienna University of Economics and Business Administration. <http://www.wu-wien.ac.at/am>.
- Chrzan, C. (2004). The Options Pricing Model. Sawtooth Software Conference 2004 Proceedings.
- Cohen, S., and Markowitz, P. (2002). Renewing Market Segmentation: Some New Tools to Correct Old Problems. ESOMAR 2002 Congress Proceedings, 595 – 612.
- Cohen, S., and Neira, L., (2003a). Measuring Preference for Product Benefits Across Countries: Overcoming Scale Usage Bias with Maximum Difference Scaling. ESOMAR 2003 Latin American Congress Proceedings, 333 – 352.
- Cohen, S. (2003b). Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation. Sawtooth Software Conference 2003 Proceedings, 61-74.
- Collins, L., Graham, J., Rousculp, S. and Hansen, W. (1997). Heavy Caffeine Use and the Beginning of the Substance Use Onset Process. An Illustration of Latent Transition Analysis.
- Diener, C. (2007). Segmentation using Choice Model Optimization. Sawtooth Software Conference 2007 Proceedings.
- Elder, A. and Pinnell, J. (2003). Archetypal Analysis: an Alternative Approach to Finding and Defining Segments. Sawtooth Software Conference 2003 Proceedings.
- Hendrix, P. and Drucker, S. (2007). Alternative Approaches to MaxDiff with Large Sets of Disparate Items. Sawtooth Software Conference 2007 Proceedings.
- Hothorn, T., Hornik, K. and Zeileis, A.(2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. American Statistical Association, Journal of Computational and Graphical Statistics, Vol 15 No. 3 pp. 651-674.
- Horton, N. and Kleinman, K., Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician, Vol 61, No. 1, 79-90.

- Huang, J., Ng, M., Rong, H. and Zichen, L., Automated Variable Weighting in k-Means Type Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 27, No. 5, 657-668.
- Jain, A., Murty, M. and Flynn, P. (1999). Data Clustering: A Review, *ACM Computing Surveys*, Vol 31, No. 3.
- Jones, U., Frazier, C., Murphy, C. and Wurst, J. (2006). Reverse Segmentation: an Alternative Approach. *Sawtooth Software Conference 2006 Proceedings*.
- Kanade, P. and Hall, L. (2007). Fuzzy Ants and Clustering. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, Vol 37, No. 5, 758-769.
- Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, Vol 11, Issue 8. <http://www.jstatsoft.org/>.
- Ma, Y., Derksen, H., Hong, W. and Wright, J. (2007). Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression,, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 29, No. 6, 1035-1051.
- Pemberton, J. and Powlettt, J. (2006). Identification of Segments Determined Through Non-scalar Methods. *Sawtooth Software Conference 2006 Proceedings*.
- Perkins, C., Kung, M., Lomeu, S., and Lineweber, D. (2007). Improving the Actionability of MEC Segmentation Models, *ART Forum, American Marketing Association*.
- Qiu, W. and Joe, H. (2006). Separation Index and Partial Membership for Clustering, *Computational Statistics and Data Analysis*, 41, 59-90.
- Ramaswamy, V. Chatterjee, R. and Cohen, S. (1996). Joint Segmentation on Distinct Interdependent Bases with Categorical Data, *Journal of Marketing Research*, Vol XXXIII, 337-350.
- Retzer, J. and Ming, S. (2007). Cluster Ensemble Analysis and Graphical Depiction of Cluster Partitions. *Sawtooth Software Conference 2007 Proceedings*.
- Rodríguez, J., Kuncheva, L. and Alonso, C. (2006). Rotation Forest: A New Classifier Ensemble Method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 28, No. 10, 1619-1630.
- Rossi, P., Allenby, G., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Wiley.
- Sá Lucas, L. (2004). Scale Development with Max-Diffs: A Case Study. *Sawtooth Software Conference 2004 Proceedings*.
- Sun, Y., (2007). Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 29, No. 6, 1035-1051.
- Tang, J. and Wiener, J. (2006). Cross-national Comparisons in Global Studies: an APEX Approach to Respondent Scale Usage Adjustment, *ART Forum, American Marketing Association*.
- Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer.

- Ulker, O., Ozcan, E. and Korkmaz, E. (2006). Linear Linkage Encoding in Grouping Problems: Applications of Graph Coloring and Timetabling, PATAT, 303-319
<http://cse.yeditepe.edu.tr/ARTI>.
- Wedel, M. and Kamakura W. (2000). Market Segmentation: Conceptual and Methodological Foundations. Kluwer Publishers.
- Weingessel, A., Dimitriadou, E. and Dolnicar, S. (1999). An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets, Working Paper No. 29. Vienna University of Economics and Business Administration. <http://www.wu-wien.ac.at/am>.
- Witten, I. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Elsevier.

DEFINING THE LINKAGES BETWEEN CULTURAL ICONS

*PATRICK MORIARTY
OTX*

*ROBERT MAXWELL
12 AMERICANS*

I. CONSUMERMAPS™ OVERVIEW

We believe that understanding how brands/cultural icons help consumers express who they are is a missing link in marketer's knowledge. ConsumerMaps™ was initiated to provide meaning to consumer's connection to popular culture: to understand why consumers like a brand/cultural icon, the reasons they identify with it, and as a way to create brand/cultural icon communities among those with similar brand identification.

ConsumerMaps™ rests on integrated theory from the fields of identity, marketing and media. Briefly, the integrated theory holds that consumers have multiple identities, some real and some aspirational, and that consumers consume brand symbols represented in the media—products, personalities, and programs—for identity reinforcement, guidance and aspiration. Today consumer relations with brands/cultural icons and self-identity are closely related; marketers need to understand the relationship between their brands and consumer's identity.

ConsumerMaps™ Theoretical Framework

This section pulls together a diverse amount of theory from a variety of disciplines as a basis for ConsumerMaps™. *First*, from identity theory it's been established that people have multiple identities, some important, others not so important, but all of which are expressed in some way; and, for each identity there is a real and an aspirational identity that people constantly negotiate. Identities are also created and shaped by social interactions; much of this interaction is symbolic in nature and is an ongoing negotiation between consumers, social groups and society. Also consumers are in a constant state of presenting themselves and in managing their identity impression with others. *Second*, from marketing theory, consumers create, cultivate and preserve their identities through possession and these possessions are frequently the instruments which people organize and construct meaning about their lives. Also, the definition of brands is undefined and *Third*, from media theory, consumers use media to support their identities as well as to provide guidance on living their identities. More specifically consumers use media to develop a variety of different identity relationships with personalities.

From these three areas ConsumerMaps™ the theoretical framework for ConsumerMaps™ was conceived. *First* using identity theory, media theory and marketing theory, it's hypothesized that people use symbols represented in the media as expressions of their identity. *Second* from media theory it's hypothesized that three symbolic elements that appear in popular media culture—products, personalities, and TV programs—are what needs to be measured. *Third*, from marketing theory it's hypothesized that anything—a product, personality, or TV program—can be a brand as long as it carries meaning to consumers.

Identity Theory as the Basis for Consumer Connection with Cultural Icons

The question of what people identify with has been debated for many years. What is identification? Why do some people identify with a particular product, celebrity or TV program, for example, while others don't? What is it in those things that people identify with? These are but a few of the many questions which social identity, marketing and media theorists have, and continue, to debate.

In understanding identity there are three key considerations. First the nature of the amount of identities we have. Second is that our identities are formed by interaction with others. Third is how we use symbols to communicate with one another about our group memberships. Understanding what constitutes identity provides a background to the rationale for ConsumerMaps™.

People Have a Self Consisting of Multiple Identities

While many philosophers and psychologists have debated the underpinnings of what constitutes the self, the majority of research and theoretical foundations on identity today are attributable to William James (1890), Charles Horton Cooley (1902), and George Herbert Mead (1934). They moved thinking of the self from a philosophical transcendent being to the idea that the self is established in everyday life. William James and George Herbert Mead are generally credited with developing the idea of the self as a psychological construct. At the heart of their approach is the belief that the individual interacts with the society at large in the construction of their identity and the self consists of multiple identities.

There are many approaches to dimensionalizing the types of identity constructs that exist. Thoits and Virshup (1997), in discussing the relationship between the "me" and "we" (i.e., the relationship between individual-level and collective-level identities in identity construction), divide the classification into the following general groupings.

1. Social Identities
 - A. Organizational (Little League member, church member)
 - B. Social roles (father, mother, brother)
 - C. Occupational roles (lawyer, researcher)
 - D. Social type of person (intellectual, leader)
 - F. Character type roles (optimist, caring)
 - G. The body (Note: author's addition)
2. Personal Identities
 - A. Gender
 - B. Race
 - C. Ethnicity
 - D. Age
 - E. Birthplace/Hometown
 - F. Physical Characteristics

In a more contemporary point-of-view Kleine & Kernan (2001), writing in the *Journal of Consumer Research* discuss the implications for marketers, "The social roles we ascribe to ourselves are the basis of our social identities and, collectively these identities form our global self—our overall sense of who we are." The self is made up of a group of different identities with many implications for marketers,

People Express Identities through Self-Presentation.

Irving Goffman (1959) summarizes the importance of self-presentation as a way to manage one's identity in a social world where symbols are constantly being negotiated, "Self-presentation is the intentional and tangible component of identity." ConsumerMaps™ takes the point of view that goods—products—are part of this. And also ConsumerMaps™ takes the point-of-view that this relates to individuals and groups.

Goffman's views have been instrumental in shaping identity theory on how people manage themselves and their multiple identities in multiple situations. He viewed the person as a social strategist where the outcome of their performance, or presentation, in a social situation is to deliver a desired impression. People act and behave differently in front of their wife vs. their boss. He believed that in these situations we are, in a sense, acting on a stage, and while doing so use a variety of symbols and techniques to deliver the desired impression. Symbols are used to negotiate a desired impression in an interaction with an audience. McAdams (1997) sums up this perspective,

Beginning with Mead (1934) and Goffman (1959), symbolic-interactionists and dramaturgical perspectives on the self have emphasized the ways in which individuals adopt multiple roles and multiple performances in order to negotiate meanings, status, and position in everyday social life. Social identities are linked to the particular exigencies of external role and situational demands, and as those demands change the corresponding identities change as well (McAdams 1997).

The Use of Symbolic Goods as Expressions of Identity

In both marketing and identity theory, it's believed consumers cultivate and preserve their identities via symbolic use of possessions (Belk 1988; Solomon 1982). McCracken also felt that goods are "an important instrument by which we capture, experiment with, and organize the meanings which we construct our lives (McCracken 1987, p. 122)." Also, the use of objects, as Cerrullo discusses in the state of identity theory, has been part of one's expression of identity. She notes the following studies of how people use art (Martorella 1989), products (Appadarai 1986; Goldman 1992; Hennion & Meadel 1993; and O'Barr 1994) and clothing (Rubenstein (1995) to project their identities. In short, products, or goods, are used as an expression of identity.

It's time media theory was brought into this. There is ample evidence in media theory dealing with people's identification with personalities, especially in wanting to be like someone—an aspiration—or an ideal identity. This is a case of someone using a personality, i.e., a "good," as a symbol of aspiration as much as one might purchase a Prada bag to express who they might be.

There is also evidence in TV program research where viewers seek social identity guidance from watching TV programs. TV programs are a "good," as they carry meaning; they are a symbol of how one should act and behave. As we'll discuss in uses and gratification research (Section II, 2), this is one of the reasons given by people for using television.

Finally, in media theory, identification with characters is looked at in a multi-dimensional framework which has somewhat inhibited the development of a comprehensive theory, as Cohen notes,

Although identification plays a major role in media research, the attempts to conceptualize the nature of identification and the theoretical treatment of this concept have been less than satisfactory. From the reviews of the literature on identification with film and television characters, it is evident that identification is understood in a variety of ways by different theorists and that this confusion has inhibited the development of a comprehensive theory of identification and its consequences.

As one can see, identification covers many areas. But the question is, why? Why do people identify with others. Here's what we believe and which will be substantiated in the remainder of this paper.

The Use of Products as a Form of *Personal* Identity

Goods—products—can be used as a symbolic form of identity in order to create an impression as suggested by Goffman. The idea of conspicuous consumption was developed by Veblen (1989) and has been connected to historical research for many years, e.g. “conspicuous and competitive consumption are especially important to the study of history of consumption because they play an important role in the growth of a consumer society.” (McCracken 1987) In a more contemporary vein and in the context of a consumer society, Schau & Gilly (2003) view consumption more generally as one “... of the most important ways in which people relate to each other socially is through the mediation of things.” Goods are viewed as symbolic when people focus more on the meaning of the good than the actual attributes (Levy 1959). Thus, goods can become the symbolic tools by which an individual communicates to another (Grubb & Grathwohl).

Aaker (1997) has used self-identity as an explanation in self-congruity theory, i.e., consumer's prefer products associated with an image that is similar to their self-concept (Belk 1988; Malhotra 1988; Sirgy 1982). However, “Instead of explicitly specifying the invoked identity, identity and social identity theory based consumer behavior focus on the influence of identity importance and commitment on the consumption of products. The idea is that self-identity is a valuable explanatory concept even though the specific invoked behavior stimulating the observed behavior may not be known.” (Pedersen Nysveen & Thorbjornsen 2003) This also demonstrates the increased attention, and believed importance, that identity issues are beginning to receive in marketing. It's best summed by Elliot Wattannasuwana (1998).

“The search of self-identity is a key determinant of postmodern consumption so it is essential for marketers to understand the concept and dynamics of self, the symbolic meaning of goods and the role played by brands.”

The link between consumption and presentation of self, as an expression of identity is commonly agreed upon. Consumers consume objects, i.e., presenting them to others in order to express an identity that someone will have to negotiate a meaning. Importantly, that expression of an identity can be an expression of a “real” identity or an identity one “aspires” to.

The Use of Products as an Expression of Communities

Goods can also be used to express identity, e.g., similarities with others, by creating groups or “brand communities” based on the consumption of similar goods. More specifically, these are groups of consumers who band together, around a product and, evidently, share the same meaning—the same meaning of the idea of the brand.

Some of the more well known brand communities that have been studied are Harley-Davidson, Apple/McIntosh, Jeep, Saab, BMW as well as entertainment programs such as Star Trek, Star Wars, X-Files and others. In reviewing the research, Muniz and Schau (2005) note the commonality by saying, “Brand communities appear to be defined, in one sense, by their capacity for powerful and transformative experiences.”

As Muniz (2005) notes, brands can not only signal group affiliation, but class, social standing, sexual orientation, a private subculture, etc. Brand communities are another form of consumer’s relationship with brands and are an example of another component of identity formation. We have discussed earlier, the importance of consumption in identity formation and expression. In an article discussing how consumers search for self and community in brands, they show how consumption practices are integral to personal and communal identity formation as well an expression of it. (Arnould Price 2001). More specifically, Shau Muniz (2002) comment about their review of the research on brand communities, “Moreover, members appear to derive an aspect of personal identity from their membership and participation in these communities.”

Marketers have become more interested in learning about, organizing, and facilitating brand communities (McAlexander, Schouten; Koenig 2002), which are “based on a structured set of relationships among admirers of a brand” (Muniz and O’Guinn 2001, p. 412). Many reasons underlie this interest, including the ability of brand communities to influence members’ perceptions and actions, often in persistent and broad-based fashions (Muniz and Schau 2005); to rapidly disseminate information (Brown, Kozinets, and Sherry 2003); to learn consumer evaluations of new offerings, competitive actions, and so forth; and to maximize opportunities to engage and collaborate with highly loyal customers (Franke and Shah 2003). In the present-day cluttered and often hostile marketing environment, many marketers believe that the facilitation of brand communities is both cost effective and powerful. (Algesheimer, Dholakia, Herman 2005).

The Use of Personalities as an Expression of Identity

Individual public figures—TV actors, movie stars, sports figures, politicians, etc. —are a significant part of popular culture. They are marketed by the media, entertainment and publicity industries, featured in the media and discussed by consumers. Many writers have documented this phenomena (See: Marshall 1997; Turner 2004). And as Joseph Epstein (2005) described it recently, it is pervasive in American culture,

Celebrity at this moment in America is epidemic, and it’s spreading fast, sometimes seeming as if nearly everyone has got it. Television provides celebrity dance contests, celebrities take part in reality shows, perfumes carry the names not merely of designers but of actors and singers. Without celebrities, whole sections of the *New York Times* and the *Washington Post* would have to close down. So pervasive has celebrity become in contemporary American life that one now begins to hear a good deal about a phenomenon known as the Culture of Celebrity.

The hero ... is always a contemporary. The hero is made by folklore, sacred texts, and history books, but the celebrity is a creature of gossip, of public opinion, of magazines, newspapers, and the ephemeral images of movie and television screen. The passage of time, which creates and establishes the hero, destroys the celebrity.

The hero in our culture has been replaced by the celebrity. Campbell (1988) has also pointed out the difference as heroes act to redeem society, whereas celebrities live only for themselves (p

xv). And Campbell noted a change in our culture where now we seem to worship celebrities, not heroes. He also observed that young people seek to be known, to have ‘name and fame,’ without any concept of having to give oneself for others (Fraser Brown 2002). Understanding why people identify with heroes is easy; however they are no longer prominent in popular culture marketing.

In discussing research on media characters and identification, Cohen (2001) sums up the state of knowledge,

Although identification plays a major role in media research, the attempts to conceptualize the nature of identification and the theoretical treatment of this concept have been less than satisfactory. From the reviews of the literature on identification with film and television characters, it is evident that identification is understood in a variety of ways by different theorists and that this confusion has inhibited the development of a comprehensive theory of identification and its consequences.

In summary, there is a substantial theoretical and empirical basis for how public figures persuade people to adopt attitudes, beliefs or behaviors. Albert Bandura’s work is probably the most valuable to understanding that identification takes place, i.e., identity can produce modeling and imitation. However, there are still many unanswered questions regarding how it works and why. Plus the research on it covers many areas. Cohen (2001) sums up the continuing need to figure what identification means,

Given the centrality of identification to media research, the need for a comprehensive theory of identification is clear. Such a theory must start with a definition of identification and measures that will enable researchers to accumulate evidence regarding the process of identification. The different concepts that have heretofore been equated with identification and used to measure it span behavioral, cognitive, and emotional concepts; encompass perceptions, attitudes, and desires; and include descriptions of a relational nature or of individual responses.

The Use of TV Programs as an Expression of Identity

Consumers spend considerable time viewing television programs—sitcoms, talk shows, reality contests, sporting events, etc.—as well as with websites, and other media content and properties. It’s hypothesized that these experiences represent a social interaction which consumers identify with in some way. While research is very limited in this area, a hypothetical point-of-view will be set forward for the development of the research to establish individual and social identity with mediatized group experiences.

Social identity theory suggests that we seek out that which supports our social identity (Abrams & Hogg 1990). It’s hypothesized that this is one of the factors that drives consumers to experience a TV program or other media content—a need to identify with a group or to express affiliation with a group. Social identity theory maintains that self-categorization, the process by which people categorize themselves into a group—a real process in the real world—will have similar applications in a mediatized entertainment environment. It’s hypothesized that media experiences are a part of this. After all, integration and social interaction (gaining insight into others, gaining a sense of belonging, connection, substitute of companions) and identification (reinforcement of personal values, identifying with others) are part of McQuail’s theory of why we use media.

McQuail's theory about why people use media seems to fit well into social identity theory as social identity is, "that part of the individual's self concept which derives from their knowledge of their membership of a social group or groups together with the value and emotional significance of that membership" (Tajfel, 1978, p. 63). It would seem that consumer's seek out mediated experiences, in part, for a sense of belonging to a group.

A TV program or any media property is, in a sense, a symbolic construction that has utility in terms of, for example, information or entertainment escapism. It's an idea that carries meaning. According to Jenkins (2003) symbols generate a sense of belonging to a group; it would seem the same could be said, for example, of a TV show. A TV program is, in a sense, a community. Anthony Cohen thought a lot about communities and the symbolic construction of them. Cohen hypothesized that a community encompasses beliefs of differences and similarity and that a community could be a symbolic construction and that membership in it means sharing similar symbolic things which create a sense of solidarity (Cohen 1986). When one looks at the sharing of symbolic things in such shows as Friends, Sex & the City, or the Sopranos, the argument would hold that TV programs provide a sense of community and social identity for many.

Harwood (1999), who provides one of the few studies on social identity in media, also echoes the concern about a lack of social identity research in media research surveyed the social related research in the field and found it unrelated to social identity theory. For example, he pointed to the work of Katz Guervitch & Haas (1973), which looked at how people used media to learn about others as well as how older people, as a group, used media (Mundorf Brownell 1990; Bliese 1986; Rosengren & Windahl 1989).

Harwood (1999) is one of the few who have broached social identity in media research. He looked at the relationship between social identity, in this case an age group, and television viewing gratifications. Briefly, he found that, "Young adults' selection of shows featuring young characters leads to increased age group identification." More specifically, he argued, "Respondents seek to view individuals with similar characteristics to themselves. However, the links from age identification and AIG to viewing indicate that this is more than a simple universal desire to view characters similar to oneself (Atkin 1985; Hoffner Cantor 1991). The desire varies with individual variation in endorsement of social identity measures. Hence, this result supports the idea that social identity reinforcement is sought by more highly identified viewers, but not by those less strongly identified."

In summary, Harwood's work sets the stage for the contention that group identification is associated with reasons individuals give for seeking out certain media experiences. As he states, "Those who expressed a stronger preference for the younger shows appeared to gain increased age identification as a result: Television viewing choices may serve identity reinforcement functions. The mere act of making a viewing choice may enhance one's sense of belonging in a group and be important to overall self-concept."

II. DEFINING CONSUMER IDENTITY FORMATION AS THE SOURCE FOR CONSUMERMAPS™

Much has been made of the mediatized culture we live in. Much has been made of the 500-channel TV environment, the thousands of commercials we consume a week, the blending of news and opinion, the interaction that the Internet has brought, the mobility of media and so

forth. Cultural theorists believe that culture has been turned upside down by this post-industrial, mediatized environment and humans now evolve with little sense of self or a core identity but instead seek popular cultural symbols as a means of negotiating who they are and where they fit in the culture.

It is not difficult to determine what the key dimensions around which produce the bulk of popular culture content. It's television, music, movies, product advertising, Internet content, books, newspapers, comic books, talk radio, etc. It's a large pool to choose from. However, from that pool ConsumerMaps™ has hypothesized that the three major dimensions of popular culture to examine are: products, personalities, and TV programs.

The rationale is as follows. *Products*— These are advertised in the media and each product presents itself with a different meaning that hopefully consumers will connect with. Product advertising is a major vehicle of identity and image messages. *Personalities*— These are people that appear in the media—movies, music, TV, talk radio—and each of them carries meaning. *TV content*— television is what consumers spend the most time with; it's a source of news and entertainment and the programming is rich in meaning and social guidance.

Understanding the relationships consumer's form with products and how products get to become brands has been a central aim in marketing research for some time. For example, one area where researchers have focused attention is on the emotional connection. Another area is the personality that a product represents. The image the product represents is another area. Also, the functional attributes a product possess has been another line of inquiry. Each of these inquires has been helpful to understanding the relationship between a product and a consumer and has helped to enlighten the definition of a brand.

However, there still is no clear definition on what a brand is. Aaker (1991), in defining brand equity, called it a mix of brand loyalty, awareness, perceived quality and other perceived associations and assets. More recently, McCracken (2006) provides another perspective on the amount of different definitions of it, "The brand is an elephant and we are all blind men. The designers have one idea of what a brand is. The Jungians another. The marketing managers, b-school professors, advertising creatives, account planners...everyone has a formal model, and a working one."

ConsumerMaps™ believes that a brand carries meaning, is an idea, that relates to consumer's identity. In other words, a consumer uses that brand to express their identity or to get guidance. And ConsumerMaps™ believes that while products have generally been perceived to be brands, the definition is broader. Mel Gibson is an idea and carries meaning to consumers. Green Day is an idea and carries a different meaning to consumers. The same for the Ford Ranger. Consumer's form a relationship with brands because of an idea imbedded in the brand that has a relation to some or one of their identities. Following is the ConsumerMaps™ definition of a brand,

(A brand is) a process of attaching an idea to a product. Decades ago that idea might have been strictly utilitarian: trustworthy, effective, a bargain. Over time, the ideas attached to products have become more elaborate, ambitious and even emotional. This is why, for example, current branding campaigns for beer or fast food often seem to be making some sort of statement about the nature of contemporary manhood. If a product is successfully tied to an idea, branding persuades people—consciously or not—to consume the idea by consuming the product. Even

companies like Apple and Nike, while celebrated for the tangible attributes of their products, work hard to associate themselves with abstract notions of non conformity or achievement. A potent brand becomes a form of identity in shorthand.

“Goods” become brands when they carry an idea that becomes a form of identity with which consumers form a relationship.

Fournier (1998) writing in the *Journal of Consumer Research* on consumers relationships to brands emphasizes the importance of understanding the context of identity in understanding a relationship, “Though they may operate below the level of conscious awareness, life themes are deeply rooted in personal history and are thus highly central to one’s core concept of self. A relationship may also deliver on important life projects or tasks. Life projects involve the construction, maintenance, and dissolution of key life roles that significantly alter one’s concept of self, as with role-changing events (e.g., college graduation), age-graded undertakings (e.g., retirement), or stage transitions (e.g., midlife crisis).” Fournier feels identity needs to be part of the equation in understanding the relationship consumers form with brands.

ConsumerMaps™ assumes people actively seek out in media that which they want, i.e., that which satisfies a need. Media competes for other options on people’s time. However, generally people choose to spend considerable time with media. The most research in this area has been done on television. It’s called usage and gratifications research.

This theoretical framework has been influential in media research. Its focus is on why people use media, and for what purpose, as opposed to what effects media might have on people. In other words, it assumes audiences actively choose media to satisfy certain needs, e.g., “what people do with media (Katz 1959).”

McQuail (1987) offers the following typology for why people watch television, the dominant entertainment media. ConsumerMaps™ feels it’s applicable to most media. There are four general uses of media according to media users:

- Information – keeping up-to-date with news, learning, self-education, advice
- Personal identity – reinforcement of personal values, identifying with others, behavioral models, insight into one’s self
- Integration and social interaction – gaining insight into others, gaining a sense of belonging, connection, substitute for companions
- Entertainment – escape, emotional release, arousal, filling time

In short, consumers use media for identity reinforcement and guidance.

III. TURNING THEORY INTO CONSUMERMAPS™: WHAT WAS DONE

With the above theoretical framework work ConsumerMaps™ developed two objectives in order to measure popular culture.

1. Objective One: To develop a measure that would allow greater understanding on how consumers define themselves across the range of personal and social dimensions.

2. Objective Two: To measure all elements of popular culture—products, personalities, and TV programs—on one scale which would allow greater meaning about how consumer's identified with them while at the same time allowing the ability to link each of them together so that brand communities could be constructed.

To accomplish these objectives, we have developed an online survey methodology collected over a series of three waves beginning in the fall of 2006 and continuing through the summer of 2007. The results discussed in this paper result from this study with the following characteristics:

- Three waves conducted quarterly from Q4 2006 thru Q2 2007
- Over 25,000 respondents
- Metrics for over 1,300 personalities, people, media properties, media content and consumer brands
- All cultural icons presented with a visual picture and well as literal name
- Average length of survey ~ 20 minutes

Evaluating Consumer Identity Construction

Our first step was to provide greater understanding regarding how and how strongly consumers define their individual identities across the range of personal and social dimensions. Are they more likely to define themselves by personal characteristics, or those that are driven by their social experiences? We acknowledge that personal characteristics are important, but our fundamental expectation is that social factors are generally more important. The level of importance associated with social factors is critical towards driving towards the value of understanding the influence of cultural icons as drivers of consumer behavior. Ultimately, understanding the nexus between personal and social characteristics and the extension to brand connection will help marketers build better relationships with consumers and to find the most effective means to reach/communicate with them and establish long term loyalties.

Measuring Core Identity

To measure consumers' core identities we utilized MaxDiff questioning prompting consumers to express the core elements of their identities in terms of importance. We use this method to tap into consumers' evaluation of their own personal identity construction, instead of revealing preferences for a product/service. In essence, we are asking consumers to evaluate their own identity much in the way that researchers try to uncover latent preferences for goods and services. We find this to be an interesting and useful application of MaxDiff as it is a practical approach for uncovering implicit as well as explicit individual preferences. Thus we use MaxDiff to get consumers to reveal the factors most determinant of their individual identity in a manner that enables us to determine both the most important factors and gain an understanding of the distribution of importance across the range of personal and social factors.

We view MaxDiff as an elegant & insightful method for deriving differentiation across highly integrated concepts like those driving one's own defining identity characteristics. We find that MaxDiff provides us with a user friendly methodology for individuals to tease apart the individual importance associated with their personal values. We believe that consumers have complex value profiles and it is generally difficult to gain differentiation using attribute based

approaches (discrete/scales). MaxDiff provides a straightforward method to derive value differentiation across individuals which delivers the capability to extend our analyses to look at value differentiation within individuals.

Using MaxDiff, we find that consumers do in fact place a high level of importance on social factors when we look at the characteristics that have the highest level of importance per individual. Figure 1 below, shows the distribution of most important identity characteristics across all consumers.

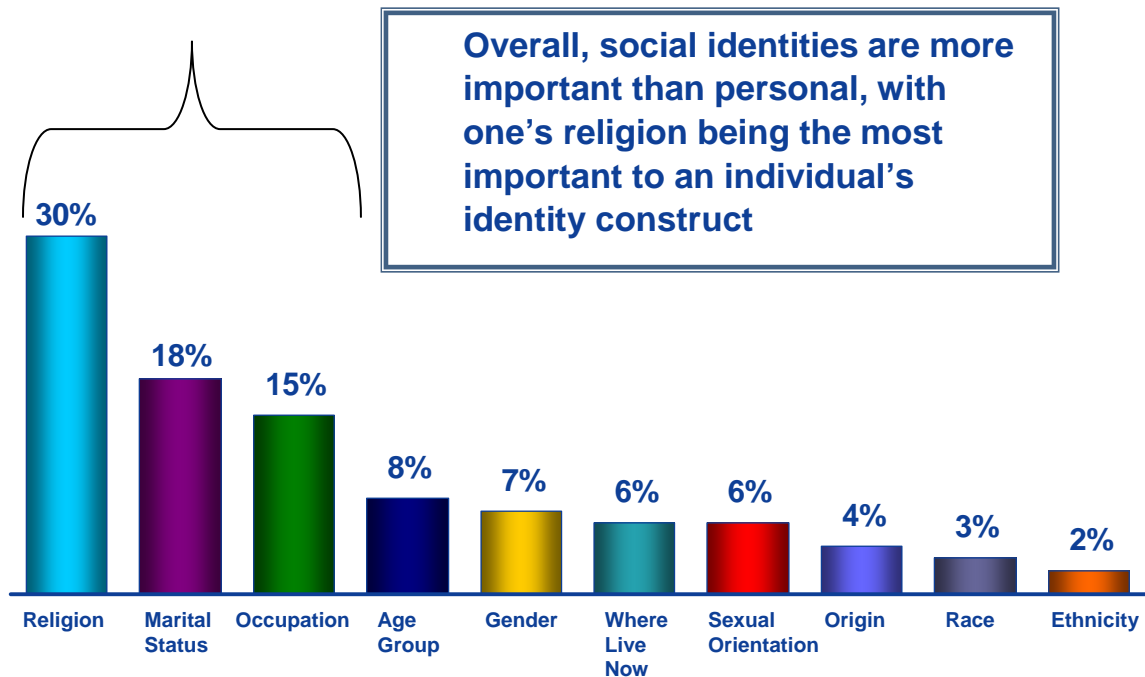


Figure 1:
Core Identity Distribution

One important point of note, is that figure 1 only shows the distribution of the most important identity factors across consumers, it does not reveal any of the rich information regarding the distribution of identity factors within individuals. The primary value of the analysis though, is that it helps to establish the overall importance of social characteristics as determinants of individual identities. This point is critical towards supporting the theory that individuals use social cues as descriptors of self and we believe that consumer consumption is a critical social cue.

Identity Drives Similar Affinity

Extending our evaluation further, we can now evaluate the differences that exist across groups of consumers when it comes to close or distant association with each of the social/personal characteristics. This is the point where we extend the importance of identity factors as drivers of affinity, thus we seek to show that groups of consumers with similar identities have distinct affinity for different cultural icons. We do this by evaluating consumer affinity evaluations for several well known politicians using the following method.

We assess consumer connections politicians using a 6 point “*Hate to Love*” scale. Our expectation is that while everyone has unique affinities across cultural icons, we will see that individuals with similar identity profiles will have similar connection profiles. Figure 2 below demonstrates this point with respect to several political figures with respect to personal and social connection:

Core Identity	Most Attractive Politician	Least Attractive Politician
• Religion	<i>George Bush</i>	<i>Rudy Giuliani</i>
• Age	<i>Barak Obama</i>	<i>Jesse Jackson</i>
• Race	<i>Barak Obama</i>	<i>John Kerry</i>
• Gender	<i>Bill Clinton</i>	<i>Jesse Jackson</i>
• Sexual Orientation	<i>John Edwards</i>	<i>George Bush</i>
• Occupation	<i>Bill Clinton</i>	<i>John Kerry</i>

Figure 2:
Core Identity Distribution

Using Affinity to Derive Linkages across Cultural Icons

Our next step was to provide greater understanding of consumer affinity across all cultural icons in our study. The key point here is to understand the affinity profiles for each individual and then evaluate the similarity and differences across groups of consumers. Similar to our evaluation of personal identity characteristics above, our expectation is that there are observable patterns of consumer affinity profiles that provide demonstrable linkages between cultural icons. The value of this assessment is that understanding the linkages will help marketers build better relationships with consumers and to leverage opportunities for messaging, cross-marketing and competitive landscape analysis.

To do this we had consumers evaluate each popular culture element separately using our proprietary 6 pt “Hate to Love” scale. Importantly, we constructed our scale such that the responses were relevant across all of the categories under evaluation: Personalities, Consumer Brands, Media Providers and Media Content. In short, we asked consumers to evaluate over 1300 cultural icons using one relatively simple affinity scale.

Affinity Scale, The Rationale:

We decided to use hate/love as the key measure for a variety of reasons. First is a commonly used expression among the public as evaluating a possession of some kind. Whether or not someone likes a public figure is probably the first step of identification with someone. It’s hard to imagine someone disliking a brand while wanting to emulate them for example. Second, liking is a universal expression of someone’s personal identification and affinity toward another, e.g., “I like him”. In identification with media figures it’s been one of the measurements frequently used. For example, Liebes & Katz (1990), in discussing identification between

viewers and TV characters cite three forms of character measurement: liking, being like (similarity) and wanting to be like (modeling). Also in television entertainment it is an indication of identification as Cohen (2001) notes, “Identification is often related to audience perceptions of liking, similarity and affinity to characters.”

Love, The Rationale:

We also chose love as a way to measure greater intensity than liking. We made the decision on similar grounds. First, it’s in consumer’s vernacular, e.g., “I love that program,” or “I love U2.” Consumers are accustomed to using it. Second the literature on love relative to consumption is persuasive, as Ahuvia (2005) indicates in his review of the literature, “In the use of products, Richins (1997) finds that love is a common consumption-related emotion. Love is so prevalent in consumption that when Schultz Kleine Kernan (1989) asked participants to list feelings that they experienced when they thought about objects with which they had an emotional attachment, love was the second most commonly listed emotion, superseded only by happiness. The people, and things, we love have a strong influence on our sense of who we are, on our self.”

RAPP Scores, the Basic Application of Affinity:

Using our proprietary six-point “Hate to Love” scale, we derive four key measures that describe the following key factors

Recognition: “Do they know what it is”

~
The proportion of the population who recognize the image and name of the icon presented to them

Atraction: “Do they love it or hate”

~
Derived score accounting for the aggregate love or hate felt for icons that consumers recognize

Presence: “Do they have an opinion about it”

~
The proportion of the population who have an opinion regarding an icon they recognize

Polarization: “The level of contrasting sentiment in the marketplace”

~
Derived score accounting for the ratio of people who like an icon versus those who do not like the icon

The RAPP scores represent our first level of assessment with respect to consumer evaluations across cultural icons. At the most basic level, we can use these measures to evaluate each cultural icon across distinct demographic groups or in relation to consumers with affinity for other icons. Figure 3 shows a very basic application with respect to understanding consumer sentiment for Hillary Clinton:

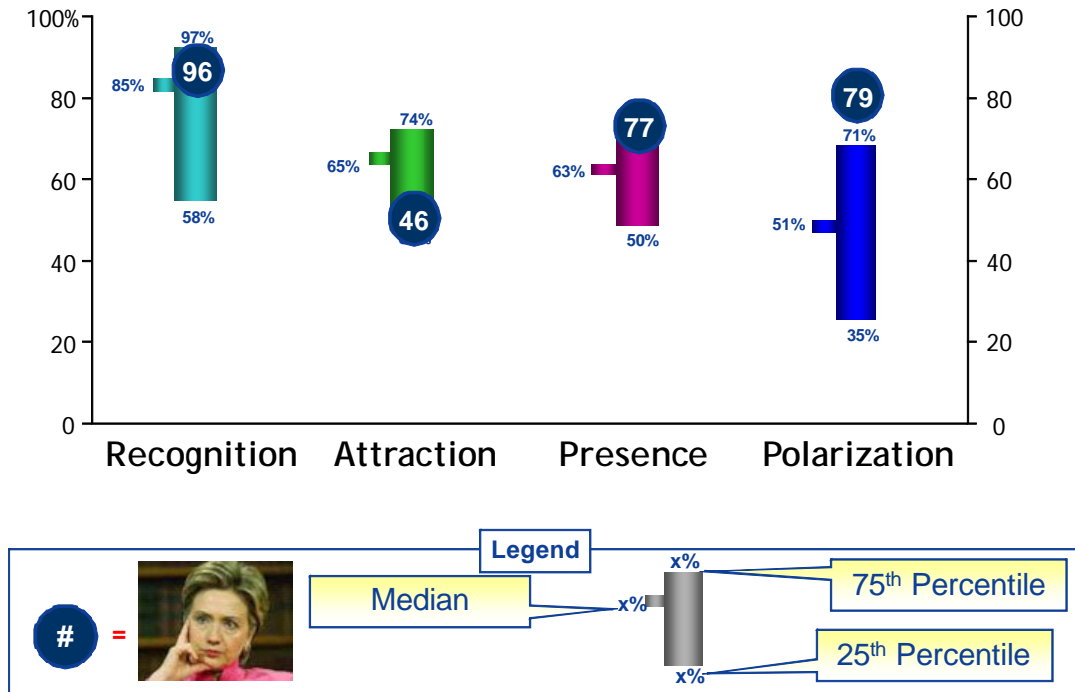


Figure 3:
RAPP Score Example- Hillary Clinton

In the case of Hillary Clinton, we find that she has a high level of recognition (96), low levels of attraction (46), high presence (77) and high polarization (79). This view comes from the total sample of 2,500 general population respondents. However, we can just as easily construct the same view across distinct consumer segments such as Democrats, Republicans, Women or even those who indicate they “Love” BMWs. Additionally, we have the capability to conduct similar evaluations across Hillary’s political competitors for comparative evaluations

Developing ConsumerMaps™ from the RAPP Scores:

While the RAPP scores are certainly valuable as an assessment tool, they represent a very basic component of ConsumerMaps™. The true extension of our effort is to deliver an assessment of the linkages that exist across cultural icons. As such, we seek to identify the cultural icons that that have the highest joint attraction. In other words, we are looking to determine the icons those who like a specific cultural icon also like. To do this we have developed a methodology to provide a similarity metric between all cultural icons that is meaningful both quantitatively and visually.

We develop the similarity metric and the resulting visual map by way of a proprietary multistage analytical process that can be generally described by the following stages:

- We derive identity groups from the MaxDiff exercise
- We collect demographic information for each respondent
- We run factor analysis across the demographics & Max-Diff identity groups to derive six dimensions of identity (mix of demographics & values)

- We calculate scores for each pair of icons identifying the joint probability of similar attraction evaluations based on the aggregate consumer evaluations
- We use Correspondence Analysis to generate coordinates for each identity dimension across over 100 consumer segments (demographic/attitudinal/identity)
- The resulting map displays the linkages between cultural icons based on consumer preferences, level of connection and commonalities across similar minded groups

The end result of this analytical process is the derivation of a ConsumerMap™ for the universe of cultural icons in our study (over 1300 at this time). We can then focus in on any specific cultural icon to evaluate their nearest neighbors and the key demographic groups that have the highest levels of attraction for that icon. Figure 4 below, provides a visual representation of ConsumerMaps™ using our example of Hillary Clinton.

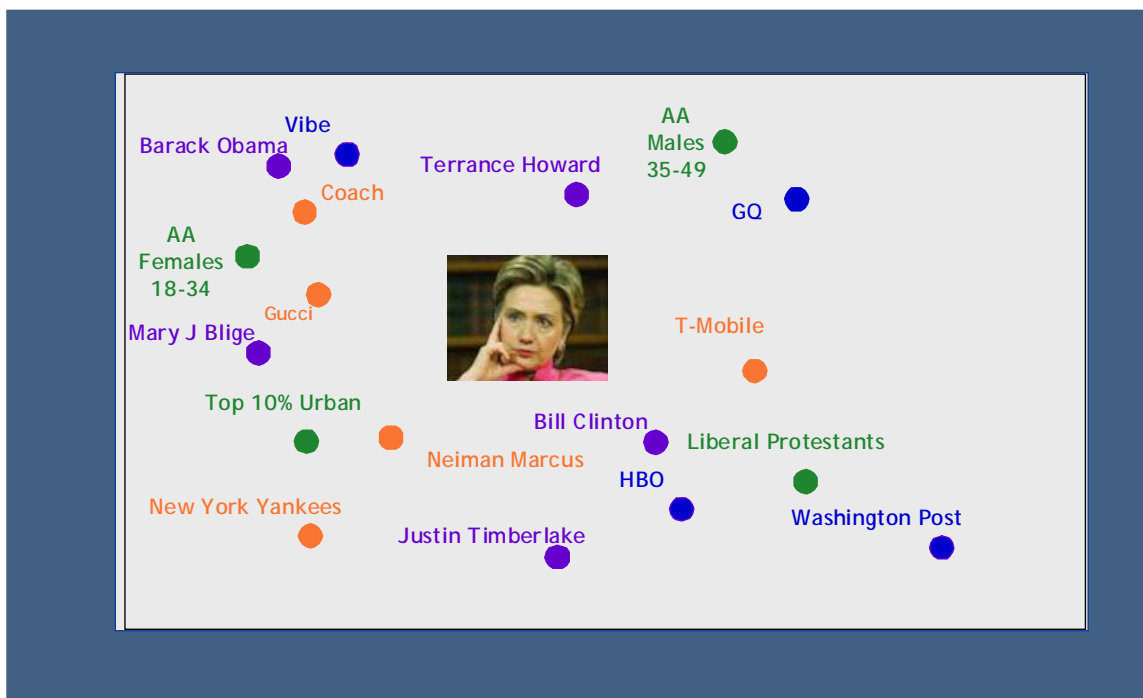


Figure 4:
ConsumerMap™ for Hillary Clinton

In the case of Hillary Clinton, we can see the key demographic groups, consumer brands, media elements and people that are most closely linked with the presidential candidate. Hillary Clinton has strong attraction with young urban populations and more liberal groups. Not surprisingly, Senator Clinton is linked tightly with other political figures Bill Clinton and Barack Obama. Interestingly though, she is also closely linked with several people associated with the young urban scene (The singers Mary J Blige and Justin Timberlake, as well as the actor Terence Howard). She is also linked with an array of brands and media providers that are generally viewed as more liberal and/or associated with urban culture. Finally, the association with the New York Yankees, is the only clear linkage between Mrs. Clinton and the constituents she primarily represents in the US Senate.

We can also develop maps for competitors, in order to do comparative analysis that may be useful for both strategic and tactical evaluation. Figure 5 provides an illustration of this with a ConsumerMap™ for Rudy Giuliani, which shows a much different collection of nearest neighbors and thus leads to a very different understanding of the consumers most highly attracted to Mr. Giuliani.

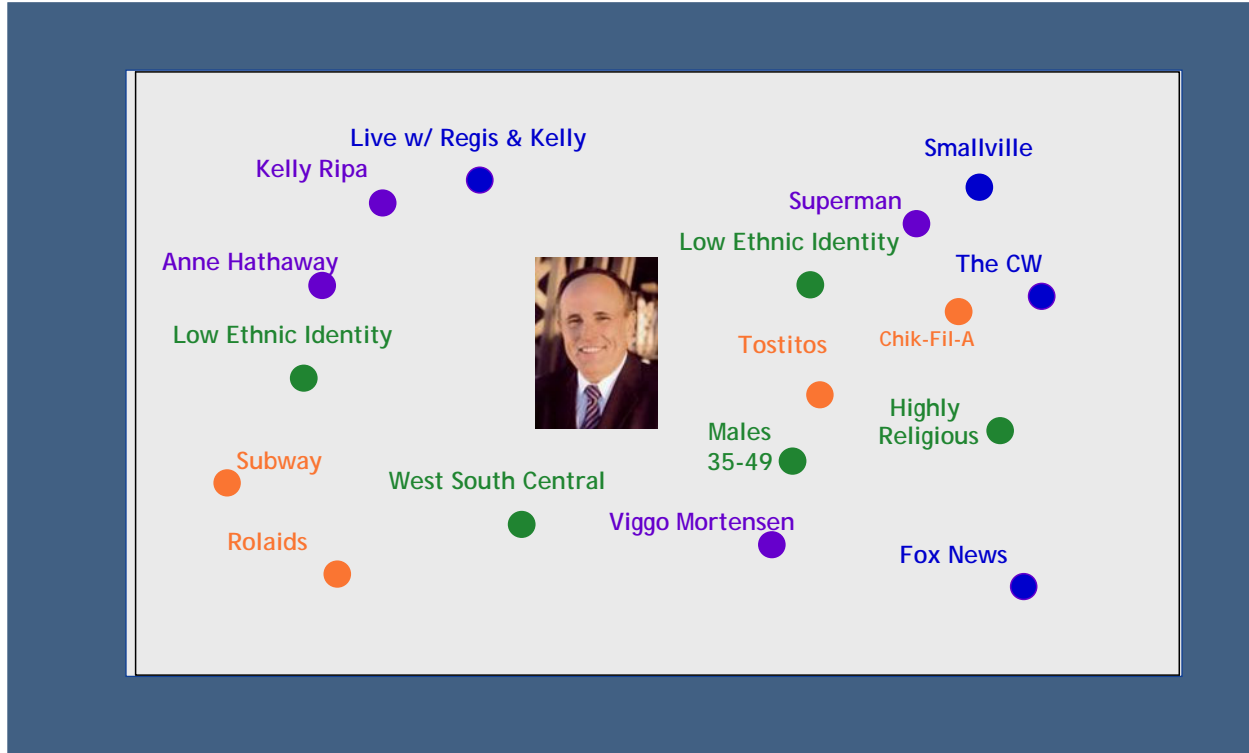


Figure 5:
ConsumerMap™ for Rudy Giuliani

The primary value of the ConsumerMaps™ for Hillary Clinton and Rudy Giuliani displayed above is that they bring to life the essence of how consumers navigate through a very crowded mediatized world. Consumers live in a highly dynamic world that is integrated and interactive. Most approaches to evaluating brands do not account for the dynamic and volatile nature of the marketplace. In ConsumerMaps™, we have developed a methodology that delivers traditional evaluative measures and accounts for the dynamic integrated nature of the marketplace. This study provides the foundational assessment metrics for cultural icons (RAPP scores), but more importantly enables us to understand the linkages that exist across cultural icons by way of similarity measures and map development. The end result is a methodology that will provide marketers with the ability to understand the core metrics of brand health and understand how consumers view their brands relative to other cultural icons in the marketplace.

IV. CLOSING SUMMARY

We developed ConsumerMaps™ because we believe that brands/cultural icons do several things. *First*, they carry imbedded symbolic meaning. Meaning comes from marketers; but also, importantly, consumers create meaning in an ongoing interactive negotiation with others—how brands/cultural icons are talked about and expressed. *Second*, brands/cultural icons are at the center of our mediatized and commercialized popular culture. *Third*, ConsumerMaps™ believes Ford trucks, Prada, U2, Tom Cruise, and Verizon all are attempting to be brands—attempting to carry meaning beyond their physical attributes. When products carry meaning beyond those attributes and are used as a form of communication about one’s identity to oneself or others, they become brands. ConsumerMaps™ believes there are three main brand categories in popular culture: products, personalities and programs. Within this context, ConsumerMaps™ provides a means to understand how much one identifies with (likes) a brand, how that brand is used to express themselves or provide social identity guidance, and how brand identification can be leveraged into brand communities.

ConsumerMaps™ consists of three separate offerings for the following brand categories: products, personalities and TV programs. 1. *Brand Attraction Ratings*, 2. *Personal Identity Expression*, and 3. *Brand Linkages*. *Brand Attraction Ratings* provide marketers with measures of awareness, likeability, salience (a factor of the two) and polarization for each brand. *Personal Identity Expression* tells how strongly a brand is used in expressions of identity: for personalities—who they want to be; for products—who they are and/or who they want to be; for programs—as a model of social guidance. *Brand Communities* tells how consumers who share similar brand identities can be brought together.

CLUSTER ENSEMBLE ANALYSIS AND GRAPHICAL DEPICTION OF CLUSTER PARTITIONS

*JOSEPH RETZER
MING SHAN
MARITZ RESEARCH*

INTRODUCTION

Learning ensembles as a general approach for machine learning have shown great promise. The basic idea is to generate multiple solutions and rely on some mechanism to combine the knowledge to achieve a final result better than one based on a single solution. Strength can be gained on the weaker individual solutions by either having them complement each other (e.g. boosting) or using them as the basis to form plurality vote as the final solution (e.g. bagging). The ensemble approach creates a consensus solution that captures data complexity which may otherwise be difficult or impossible using any single model.

While various well-known ensemble methods such as boosting, bagging and Random Forests have been proven highly effective in improving model performance, these learning ensembles all take place within the domain of supervised algorithms. However, the same idea is equally applicable for unsupervised learning which is typically associated with clustering techniques. In contrast to supervised learning, where the group label for each individual case is known and the main goal is to improve the prediction of classification and understand what is driving it, clustering is aimed at identifying group membership without the use of pre-existing labels. Cluster ensembles, as the name suggests, employs multiple “base” cluster solutions to reach the final clustering solution called the “consensus” solution. The approach has shown benefits specifically related to its ability of improving both accuracy and stability over a single partition. We demonstrate the potential of this relatively new clustering technique introduced by Strehl, A. & J. Gosh (2002), as a potent segmentation method for market research. We also review some ways of using special graphical techniques to evaluate cluster solutions.

DESCRIPTION OF METHOD

Cluster ensembles begin by generating multiple cluster solutions from the same data set. These solutions can be produced in various ways. For example, they could be based on a single base learner (e.g. k-means) with different initialization values. Alternatively completely different and independent clustering algorithms may be employed to generate ensemble member partitions. The primary focus of cluster ensembles however is to address the problem of combining these different solutions to create a consensus.

Consider P_1, P_2, \dots, P_L to be a set of partitions of data set Z , each from a clustering algorithm. The goal is to find a P^* based on P_1-P_L which best represents the structure of Z . P^* is the combined decision referred to as the consensus. Choice of the base algorithm(s) is the decision of the analyst. An important requirement is that there be diversity among individual solutions. If not, the ensemble result will mirror the individual solutions.

1. Generating partitions

Given a pre-specified number of clusters, there are many ways to generate diversified individual solutions. We briefly introduce some typical approaches:

a) **Generating partitions through random initializations.**

One of the key challenges for solving a clustering problem is often the lack of a clear-cut boundary among the limited number of clusters. This is perhaps even more so in the case of customer segmentation in market research where things from behavioral complexity and uncertainty of the customers to scale usage heterogeneity and measurement errors introduced by the survey instruments can all add ambiguity to the data. Setting data issue aside, it is well known that different runs on the same data by exactly the same method can often result in very different cluster solutions. We take k-means, perhaps one of the most frequently used methods, as an example here. It starts by selecting random seeds as the base for forming a preliminary cluster membership for all other cases in the data set and then step by step improving cluster membership structure by adjusting membership assignments for individual cases based on some criteria until the final stability is achieved. Typically, the criteria involve minimizing the total intra-cluster variance so that each member has the smallest distance to its current cluster center.

In the case of trying to improve on a single traditional solution approach, one could aim at working with the seeding structure to achieve better solutions (e.g. Arthur and Vassilvskii 2007). Sawtooth Software's convergence clustering algorithm also specifically addresses this type of problem by choosing more sensible seeds and evaluate cluster reproducibility of the different solutions so that the chance of settling on a final solution that is more biased from the optimal one can be reduced. In contrast, cluster ensemble addresses the uncertainty by taking advantage of such phenomenon of diversity through the creation of many different possible partitions resulting from random initializations. In a certain sense, the instability of a single partition due to the fuzziness of the data or method dependency becomes a welcoming feature to ensemble approach for achieving a more robust result.

b) **Sub-sampling/re-sampling.**

In this approach, individual partitions in the ensemble are generated by first drawing sub-samples from the given data set and then conducting clustering only on these subsets of data. Sub-sampling can be done either without replacement or with replacement (i.e., bootstrap). It has been shown that clustering ensembles based on small sub-samples generally can capture a meaningful consensus partition for an entire set of data (Minaei-Bidgoli etc. 2004). This approach can be especially advantageous when clustering ensemble is conducted on very large data set and computational cost and time becomes a concern.

c) **Use different types of clustering algorithms.**

As seen above, the input matters to cluster ensemble is the specification of how each case is assigned to a cluster rather than the detailed clustering mechanism by which the actual partition is obtained. For real problems, the choice of the actual methods depends upon a range of factors from the data type and size in hand to the availability of the software tools. Sometimes, one may not be able to decide on the most suitable method for a given problem. To harness the strength of a diversity of methods, one may consider applying more than one base clustering algorithm to generate individual solutions.

d) Use subsets of features.

Another way of creating different base solutions is to take different subsets of variables. Particularly from a practice perspective, it is reasonable and sometimes even beneficial to choose a subset rather than all variables from the data set, conduct independent clustering analysis and finally combine the results. For example, a firm might be interested in separately exploring a specific customer demographics segmentation solution and in addition, attitudinal or needs based segmentation solutions. Ensemble analysis allows one to bridge these separate solutions based on different subsets of features to create a final consensus.

e) Randomly choose different number of clusters.

One of the many challenges for data clustering involves making a decision on the number of clusters, “ k ”. In practice, exploring and evaluating multiple solutions with different values of “ k ” often achieves this goal. Similarly, one approach for generating multiple ensemble measures is to create partitions based on randomly selected values of “ k ”. This approach has been shown to be a very effective heuristic. In theory, we may generate partitions based on all possible values of “ k ” i.e., $k = 2, 3, K, n$.

f) Project data in random affine subspaces.

This approach randomly generates a linear transformation matrix in a fashion analogous to principle components analysis (PCA). Unlike PCA, however, where the matrix is generated with a specific goal in mind, i.e., to maximize component variance, random affine subspace linear transformations are purely random. Empirical studies suggest this approach is inferior in some respects to other methods.

2. Arriving at consensus

The goal of this step is to reach a final single clustering result based on a set of diversified individual partition solutions. As previously noted, the consensus solution integrates the inputs from each individual solution and may better represent the underlying data structure. This step is the core of ensemble clustering and has been subject to extensive research. We briefly introduce some common approaches. The reader may refer to sources such as those by Day (1986) and Topchy etc. (2004) for more information.

a) Direct method.

One major source of complexity in obtaining a consensus partition is due to the fact that labels attached to cluster solutions from multiple partitions are not fixed. For example, a cluster labeled “1” in a given solution may be best reflected by cluster “3” in another. Addressing this absence of explicit correspondence between the labels becomes a critical task. A permutation of the cluster labels through re-labeling is necessary to compare solutions and arrive at a consensus. The final cluster assignment is then determined by finding a consensus solution with minimum distance, on average, to all other permuted partitions.

b) Feature based method.

In the feature-based approach, each underlying cluster solution is considered to be reflective of some “feature” of the data. As such, each data vector partition may be viewed as analogous to an underlying basis variable employed in the initial cluster analyses. Given this perspective, we may then use the ensemble member partitions as attributes in a mixture model segmentation analysis. The mixture model will then produce the consensus clustering.

c) Hyper-graph method.

In this approach, hyper-edges on a graph with N vertices are created to represent all the clusters in the ensemble partitions. Each hyper-edge describes a set of objects belonging to the same cluster. A consensus function is formulated as a solution to k -way min-cut hyper-graph partitioning problem. Each connected component after the cut corresponds to a cluster in the consensus partition. However, the cuts only remove hyper-edges as a whole. Hyper-graph algorithms seem to work the best for nearly balanced clusters.

d) Pair-wise method.

The pair-wise method begins by depicting each underlying cluster solution with a similarity matrix reflecting joint membership in the various segments. The resultant matrices are averaged to create an ensemble similarity matrix. Hierarchical cluster methods are then applied to produce the segment assignments.

3. Determining the optimal number of clusters

A critical decision on the correct number of clusters must be made based on the evidence generated by the clustering process. This can directly impact the interpretation and usage of the consensus clustering solution. A large amount of research is available on this topic and as such, numerous options exist. There is little consensus on either the approach or criteria however. For practitioners, a graphical approach based on observing various clustering scenarios can be powerful and intuitive. In our view, their applications are still lacking however. Our goal is to introduce some graphical tools to aid in intuitive evaluation of the final consensus solution. Note that these tools could be applied to conventional clustering methods as well. Our focus is on a particular method referred to as a “silhouette plot” (Rousseeuw, 1987). Before going into detail, we first offer a few general observations on the relationship between the number of clusters and the quality of the solutions.

- This issue can be partly addressed by evaluating cluster solution using either geometric properties (e.g., within cluster concentration and between cluster separation) and/or solution stability. Here we assume the correct number of clusters is a point of stability for clustering algorithm. As illustrated later, the consensus solution may also be evaluated on the basis of diversity.
- The geometric approach assumes specific cluster shape. If clusters are not of the shape assumed by algorithm, they could be missed.
- Stability makes no implied guess about cluster shape but also does not guarantee correct solution.

SILHOUETTE PLOTS

Two critical tasks in cluster analysis are: deciding on the appropriate number of clusters and distinguishing a bad cluster from a good one. Silhouette plots provide key information about both. They display the entire clustering by plotting all silhouettes into a single diagram. Silhouettes show which objects lay within the cluster and which objects merely hold an intermediate position. They are an effective tool for depicting cluster solution quality as well as to make comparisons among multiple potential solutions.

Only two things are required in order to construct silhouettes: the clustering results obtained from an ensemble (or any other clustering algorithm) and the pair-wise distances among all objects. The later is typically represented in the form of a dissimilarity matrix. Consider any object i of the data set, and let A denote the cluster to which it is assigned, and then calculate:

$$a(i) = \text{average dissimilarity of } i \text{ to all other objects of } A$$

Now consider any cluster C different from A and define:

$$d(i,C) = \text{average dissimilarity of } i \text{ to all objects of } C$$

Compute $d(i,C)$ for all clusters C other than A and select the smallest one to denote it as:

$$b(i,C) = \min\{d(i,C)\}, \text{ where } C \neq A$$

Let B denote the cluster which attains the minimum i.e., $d(i,B) = b(i)$, which is called the neighbor of object i . Define $s(i)$ as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The value $s(i)$ indicates how well the i^{th} object fits into its assigned cluster vs. its closest neighbor. Specifically, $a(i)$ measures how dissimilar object i is, on average, to objects in its own cluster and $b(i)$ measures how dissimilar it is, on average, to the next closest clusters objects. The higher $b(i)$ and a lower $a(i)$ produces a higher positive $s(i)$ and suggests the object fits with its current cluster well rather than needs to be considered for a reassignment to one of the other clusters. On the other hand, a higher $a(i)$ and lower $b(i)$ indicates that the object does not appear to have a lot of similarity with the other objects in its current cluster and it may be rather closer to some other clusters. In theory, $s(i)$ cannot be larger than 1 or less than -1. A s value close to 1 means the object falls into the current cluster with almost no ambiguity. A s value close to 0 means the object is about as close to another cluster as to its own. These are likely to be those objects located near the borders between different clusters. On the other extreme, if a s value is close to -1, it may be preferable to have the case switched to a different cluster.

The value $s(i)$ is the key for building the silhouette plot. Silhouette plots can be viewed as a horizontal bar chart on these individual $s(i)$ values for all of the objects with no space between the bars. Silhouettes corresponding to each cluster are arranged from the top to the bottom by the unique cluster numbers or labels. Within each cluster, the individual objects are ranked from high to low according to the $s(i)$ value. The height of the silhouette reflects the size of the cluster, where the width of the silhouette reflects the quality of the cluster. A silhouette stretching wider to the positive side is more desirable as it reflects a cluster with more distinguished membership grouping from the rest. On the contrary, a narrower silhouette or one with a section of even negative silhouette value indicates a relatively larger degree of uncertainty in term of the classification of those objects and gives a sign of a weaker cluster.

Besides visual inspection, one can also rely on the within cluster or even the overall average of the silhouette width to determine the quality of the cluster solutions. One way of making a decision on the number of clusters (k) is to simply compare the overall average silhouette width (SC) among solutions and choose the one having the highest value.

$$SC = \max_k \bar{s}(k) \quad k = 2, 3, \dots, N$$

Figure 1 is a silhouette plot of 4 cluster solutions on 4381 cases. In this particular case, the size and average silhouette width of each cluster is printed on the right hand side to aid better interpretation. Cluster 2 has the widest silhouette and a very small proportion of it appears to be close to 0. It is the best cluster among all four. This is confirmed by the highest average silhouette of 0.52. Its relative shorter height indicates a smaller cluster size. Cluster 3 is similar in size but inferior in quality. Both the first and the last clusters are relative large in size. But the 4th cluster has a very short silhouette width and a section of it has negative silhouette value, suggesting a questionable cluster formed by those objects.

SC is a dimensionless measure of the extent of clustering structuring. Table 1 gives a general guideline for determining cluster solution quality when examining silhouette plots.

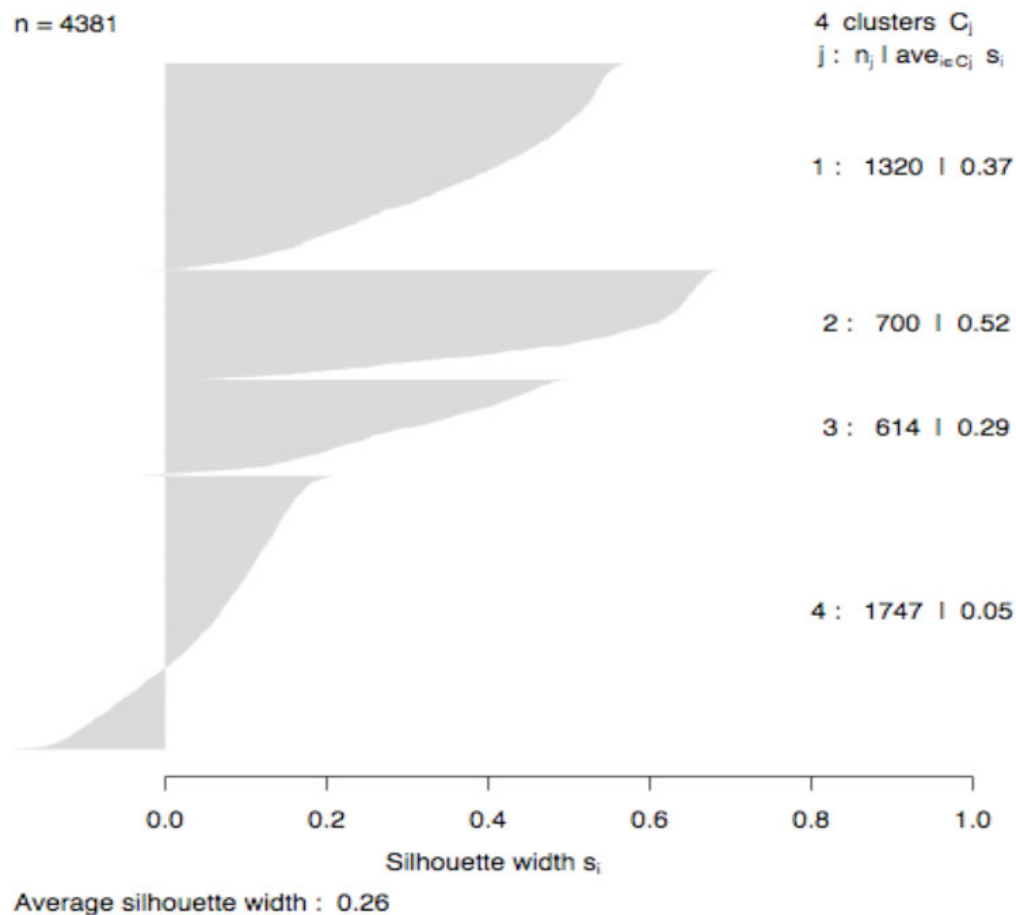


Figure 1.
 An example of silhouette plot with 4 cluster solution.

Range of SC	Interpretation
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	Structure is weak and possibly artificial
≤ 0.25	No substantial structure found

Table 1.
Interpretation of overall average silhouette width

4. Illustration: comparing cluster ensemble solutions with known labels

We use the Cassini data studied by Dimitriadou *et al.* (2002) and Leisch (1999) to compare the effectiveness of using cluster ensemble over single clustering solution in uncovering non-spherical clusters. Figure 2 shows the scatter plot¹ and the silhouette plot of this exactly two-dimensional data. Each cluster contains 1000 data points. The three clusters are visually distinguished one from another (Editor's note: cluster membership is represented by different colors, which cannot be seen in the black-and-white, hard copy proceedings. Please refer to the electronic version of these proceedings to view these graphics in color). Let's now compare single clustering solutions vs. cluster ensemble.

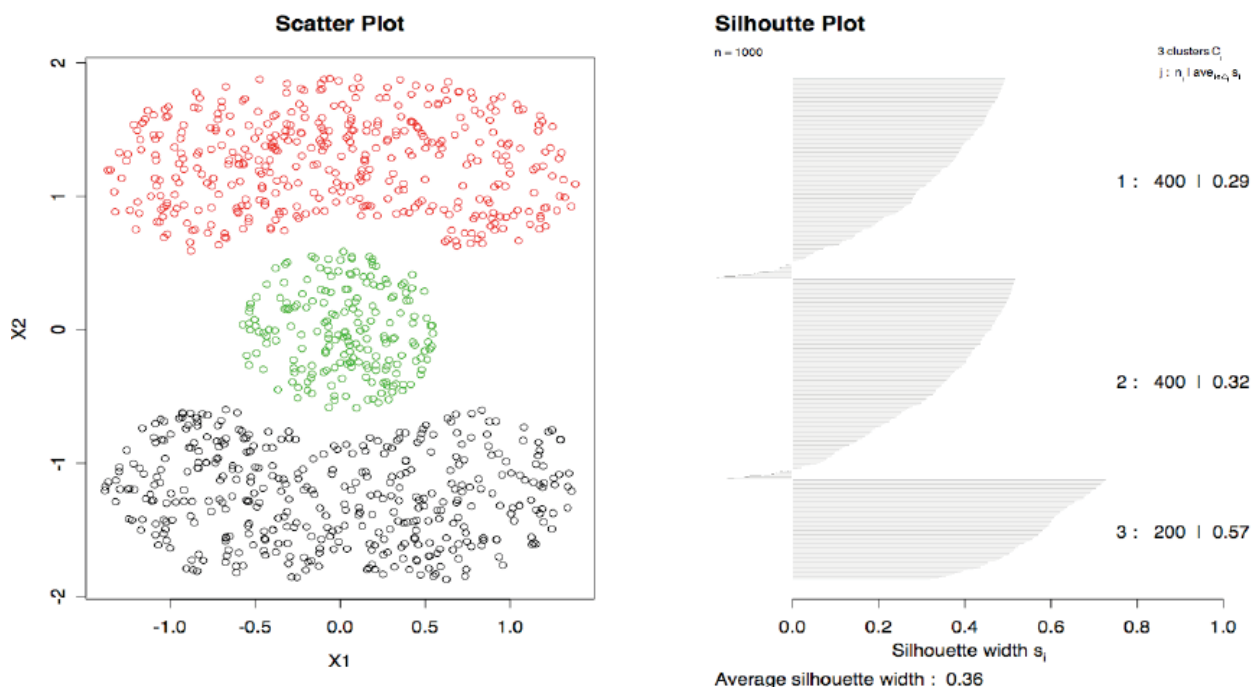


Figure 2.
Cassini data

¹ The scatter plots in figures 2-5 require color representation to display the group memberships, whereas these proceedings are printed in black-and-white. The interested reader can refer to electronic color renditions of this paper in the CD of these proceedings distributed by Sawtooth Software or in the PDF document downloadable from the Technical Papers Library at www.sawtoothsoftware.com.

Figure 3 shows 6 out of the 9 different individual cluster solutions using a mix of two different clustering algorithms: conjugate convex functions and k-means. None of these single solutions stands out as effective. Part of the reason that *k*-means algorithm by itself is not able to detect the natural clusters is its implicit assumption of hyper-spherical clusters.

One way to examine diversification of the individual solutions of a clustering ensemble is to visualize pair-wise distances among them using a dendrogram. The left panel of Figure 4 depicts the similarity among 9 different individual solutions with cluster numbers varying from 3 to 4. The scatter plot in right panel suggests some improvement based on the ensemble of these 9 solutions. However, the solution still differs strongly from the known clustering structure. Specifically, it matches the bottom cluster well but generates incorrect separation between the top and the center groups.

When we further increase to 28 the number of diversified individual solutions (by varying both cluster numbers and the base clustering algorithms) the new ensemble solution appears to have recovered most of the original structure. The results are shown in Figure 5.

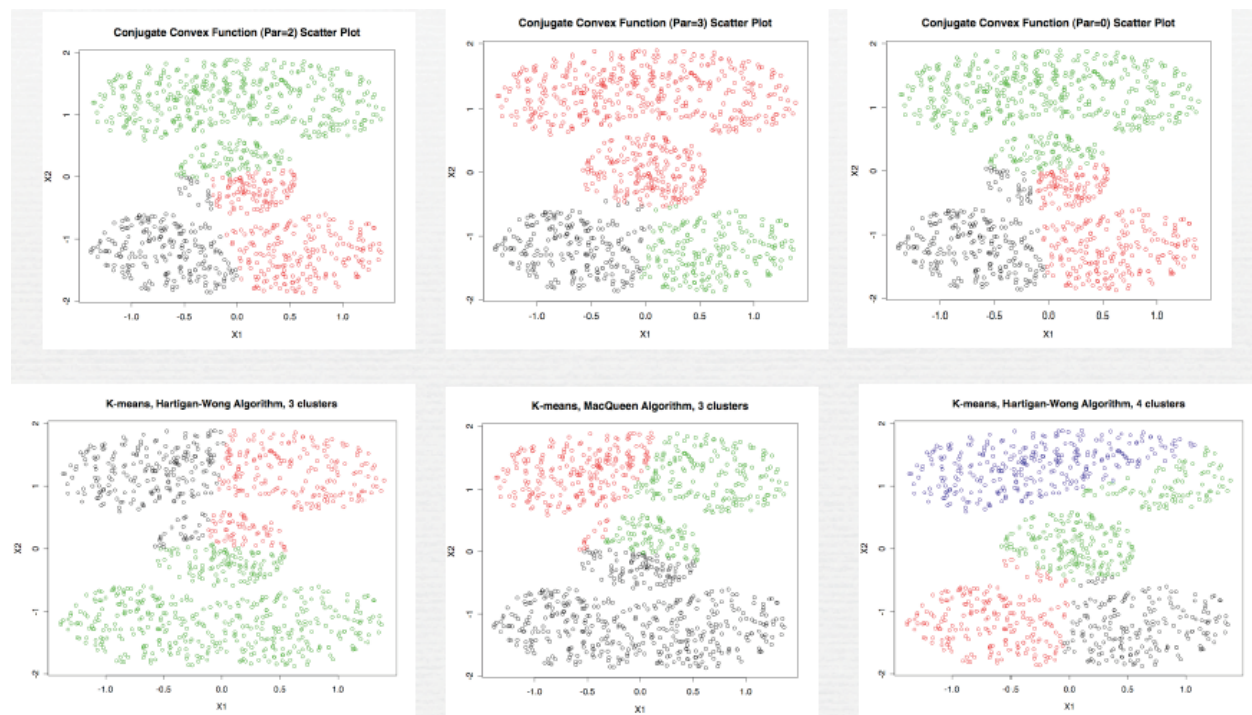


Figure 3.
Six individual solutions based on conjugate convex functions and k-means

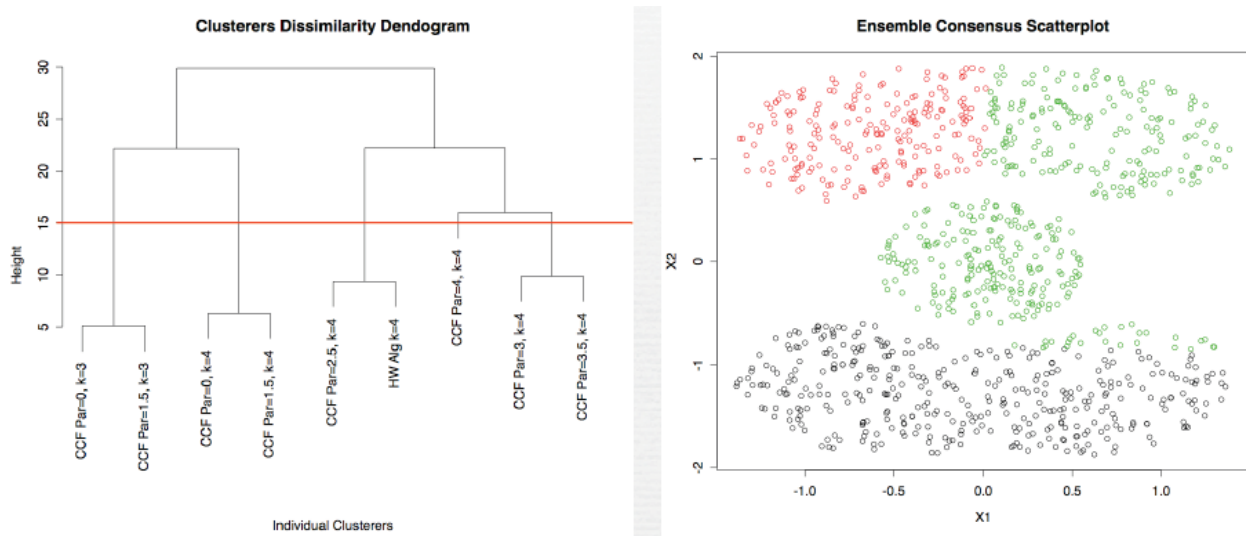


Figure 4.
An ensemble of 9 individual clustering solutions

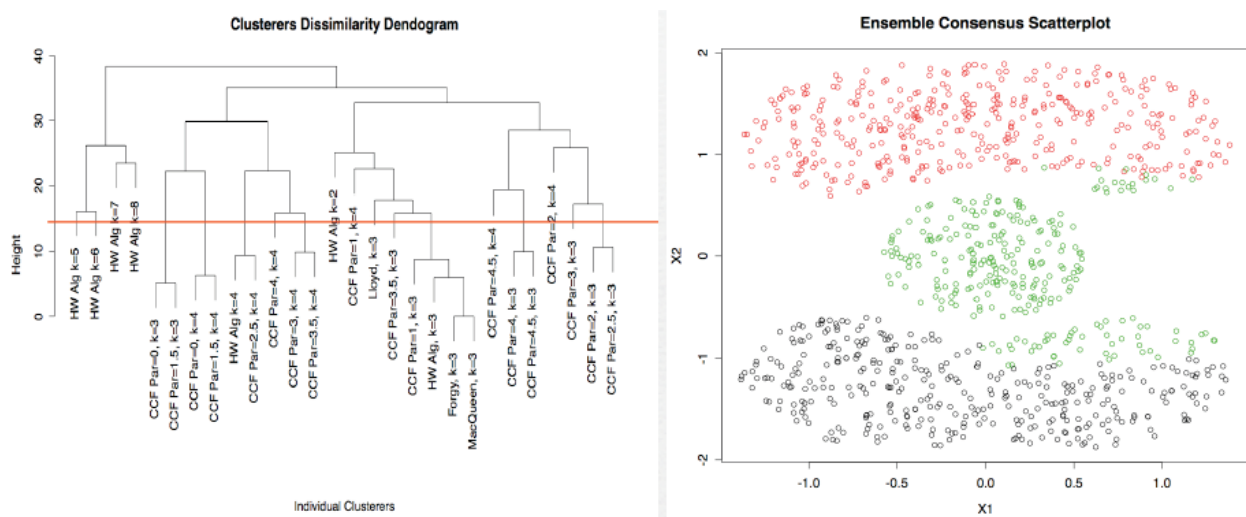


Figure 5.
An ensemble of 28 individual clustering solutions

Major Benefits and other Potentials in Application

Cluster ensemble analysis has been shown to be capable of producing more robust solutions than a single clustering. The major benefits and potential practical expansions can be multi-fold. These benefits may be summarized as follows:

First and the foremost, it helps to improve the quality and robustness of the solution by better averaging performance across domains and data sets. This is also evidenced in examples where demonstrated gains from ensemble methods have been realized in the area of classification and regression analysis. Note that of critical importance for achieving this robustness is that each learner (or clusterer) should be strong with different inductive biases.

Cluster ensembles can improve the stability of the clustering. In other words, solutions with lower sensitivity to noise, outliers or sampling variations may be obtained

For large problems, we can leverage multi computer power by pursuing parallel clustering of data subsets and subsequently combining the results.

Ensemble methods may lead to novel findings not attainable from single clustering solutions.

For situations where access to data features is limited, a primary potential benefit pertains to knowledge reuse. For instance, one can merge previously existing cluster solutions with no knowledge of the algorithm or even basis variables used to create the partition. An example would be merging clustering results on customer demographics, with partitions based on credit ratings with satisfaction data partitions. Another example could be to incorporate legacy-clustering solutions provided by human experts or other entities using proprietary techniques.

Perhaps somewhat less relevant to a typical market research situation but potentially of great use in some other larger scale application, cluster ensemble offers the opportunity of distributed computing. Specifically, it is possible to conduct separate clustering on data stored at different geographical locations due to organizational or operational constraints. These multiple partitions could then be combined using ensemble analysis.

Lastly, for the purpose of privacy preservation, separate entities may cluster on subsets of features or attributes and share only cluster labels.

A General Framework for Cluster Ensemble Analysis

In this paper, we demonstrate cluster ensemble analysis as a general approach for improving clustering over traditional single partition approaches. Cluster ensemble analysis begins by generating multiple diverse cluster partitions. Next a consensus is formed using one of a number of various algorithms. Lastly, visualization tools may be used to aid in arriving at and assessing the optimal solution. Figure 6 summarizes this general framework.

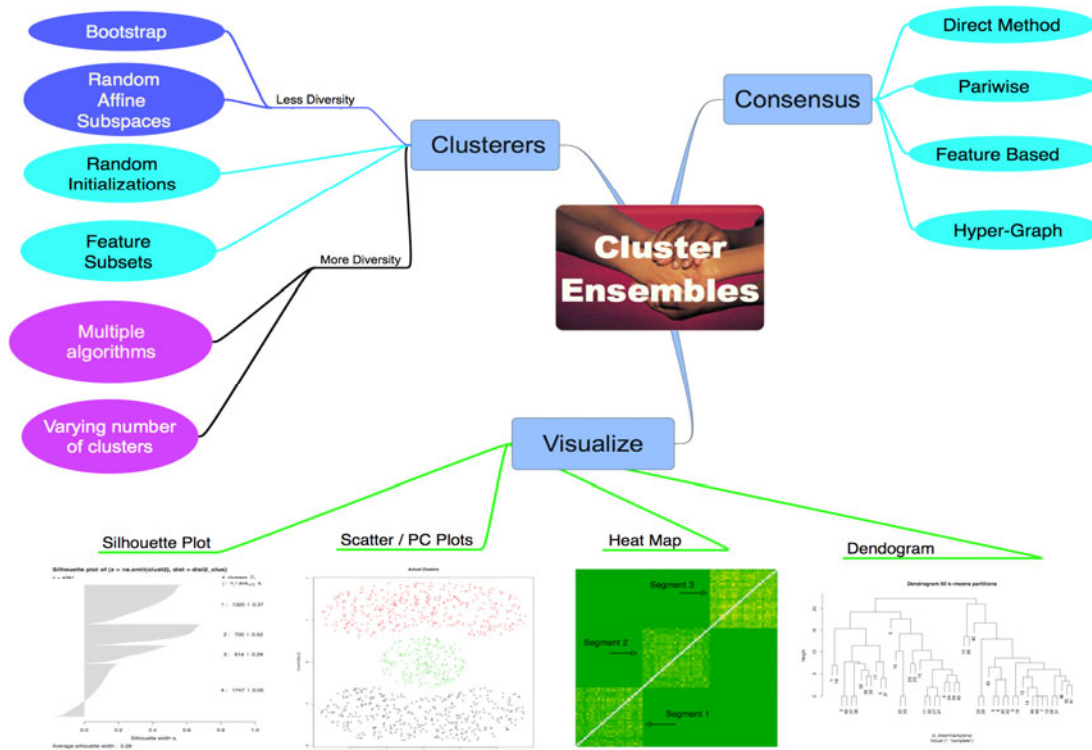


Figure 6.
A general framework for cluster ensemble analysis

CONCLUSIONS

Cluster ensemble analysis is a computationally intense data mining technique. It is not a new clustering algorithm in itself but rather builds on other pre-existing algorithms. Using these algorithms, a consensus cluster solution is created which is superior to its individual underlying components.

This paper has demonstrated cluster ensemble analysis and reviewed various approaches for both generating individual cluster solutions as well as creating a final consensus solution. It has also presented evidence of cluster ensemble analysis' ability to detect non-spherical clusters.

Future work will involve empirical comparisons of cluster ensemble analysis to alternatives as applied to marketing data. In addition, construction and evaluation of appropriate measures of cluster quality and stability will also be pursued.

REFERENCES

- Arthur D. & S. Vassilvitskii (2007), "k-means++ The Advantages of Careful Seeding." Symposium on Discrete Algorithms (SODA).
- Day, W. H. E. (1986), "Foreword: Comparison and consensus of classifications." *Journal of Classification*, 3,183-185.
- Dimitriadou E., A. Weingessel, & K. Hornik (2002), "A combination scheme for fuzzy clustering." *International Journal of Pattern Recognition and Artificial Intelligence*, 16(7), 901–912.
- Gordon, A. D. (1999), "Classification." Chapman & Hall/CRC.
- Hornik, K. (2007), "A CLUE for CLUster Ensembles." R package version 0.3-18. URL <http://cran.r-project.org/doc/vignettes/clue/clue.pdf>.
- Kaufman, L. & P. J. Rousseeuw (2005), "Finding Groups in Data, An Introduction to Cluster Analysis." Wiley-Interscience. Kuncheva.
- Leisch F. (1999), "Bagged clustering." Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science."
- Rousseeuw, P. J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Kuncheva, L. I. & D. P. Vetrov (2006), "Evaluation of stability of k-Means cluster ensembles with respect to random initialization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, 11, November.
- Minaei-Bidgoli, B., A. Topchy & W. Punch (2004), "Ensembles of Partitions via Data Resampling." *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, Vol. 2, 188-191.
- Strehl, A. & J. Gosh (2002), "Cluster Ensembles - A knowledge reuse framework for combining multiple partitions." *Journal of Machine Learning Research*, 3, 583-617.
- Topchy, A., A. Jain & W. Punch (2004), "A mixture model for clustering ensembles." *Proc. of the SIAM conference on Data Mining*, 379-390.
- Weingessel, A., E. Dimitriadou, & K. Hornik (2003), "An ensemble method for clustering." DSC Working Papers.
- Xiaoli, F. & C. Brodley (2003), "Random projection for high dimensional data clustering: a cluster ensemble approach." *Proc. of the twentieth conference on machine learning*, Washington D.C.

MODELING HEALTH SERVICE PREFERENCES USING DISCRETE CHOICE CONJOINT EXPERIMENTS: THE INFLUENCE OF INFORMANT MOOD

CHARLES E. CUNNINGHAM,
HEATHER RIMAS,
KEN DEAL,
MCMMASTER UNIVERSITY

This paper considers the rationale for the use of conjoint analysis to involve patients, family members, and the broader public in the health service planning process. Next we discuss the ways in which the depressed mood which accompanies many acute and chronic health problems might influence the information processing and decision making mechanisms that mediate performance on discrete choice conjoint experiments (Bakken, 2007). Finally, we present the results of studies designed to determine the extent to which depressed mood might influence the reliability of responses on discrete choice conjoint surveys, the validity of share of preference simulations, and the broader service preferences of parents seeking children's mental health services.

USING DISCRETE CHOICE EXPERIMENTS TO MODEL HEALTH SERVICES

Involving patients and the broader public in health service planning decisions is a basic tenant of patient-centred care and, increasingly, a matter of public policy (Ryan *et al.*, 2001). A large number of studies have used stated preference methods, such as conjoint analysis or discrete choice conjoint experiments, to understand the service delivery preferences of patients with a wide range of medical problems. These methods have been applied to prenatal care (Lewis, Cullinane, Carlin, & Halliday, 2006), endocrinology (Ahmed, Blamires, & Smith, 2007), oncology (Basen-Engquist *et al.*, 2007), HIV (Beusterien, Dziekan, Flood, Harding, & Jordan, 2005; Phillips, Maddala, & Johnson, 2002), osteopathy (Fraenkel, Gulanski, & Wittink, 2007) psychiatry (Dwight-Johnson, Lagomasino, Aisenberg, & Hay, 2004; F. R. Johnson *et al.*, 2007), and children's mental health services (C. E. Cunningham, Buchanan, & Deal, 2003; Spoth & Redmond, 1993). They have also been extended to study clinical decision making processes (Chinburapa *et al.*, 1993; McGregor, Harris, Furuno, Bradham, & Perencevich, 2007; Oudhoff, Timmermans, Knol, Bijnen, & Van der Wal, 2007; Wigton, Hoellerich, & Patil, 1986), medical education (C. E. Cunningham, Deal, Neville, Rimas, & Lohfeld, 2006), and the service delivery preferences of health care professionals (Caldon, Walters, Ratcliffe, & Reed, 2007; Chinburapa *et al.*, 1993; Oudhoff *et al.*, 2007).

Several factors make discrete choice conjoint experiments especially useful as health service planning tools (Ryan *et al.*, 2001). First, well designed discrete choice conjoint experiments approximate the complex conditions under which care providers and their patients make real world health service decisions. Patients, like service planners, must make decisions regarding treatment options with a competing combination of potential benefits, risks, costs, and logistical demands. Discrete conjoint experiments allow the views of both health service providers and their patients to inform the service planning process. This is particularly important because the

health service delivery preferences of providers and patients are often quite different (C. E. Cunningham *et al.*, 2003; Pieterse, Stiggelbout, Baas-Thijssen, van de Velde, & Marijnen, 2007).

Second, patients often fail to use or adhere to potentially effective health and mental health services (McDonald, Garg, & Haynes, 2002; Vermeire, Hearnshaw, Van Royen, & Denekens, 2001). Health services which consider the design preferences of patients may improve utilization and adherence by reducing the barriers that inadvertently undermine the outcome of potentially effective treatments (C. E. Cunningham *et al.*, 2000; Kazdin, Holland, & Crowley, 1997; Owens *et al.*, 2002).

Finally, the development of health services is a complex and often expensive process. Planners must consider the costs of different service delivery options, weigh the outcomes of alternative treatments, anticipate the relative risk of short and long term side effects, evaluate the relative strength of the scientific evidence supporting different options, and consider the preferences of patients, family members, and the broader public. By understanding the relative importance of the attributes composing a potential service, and simulating the response of patients and care providers to different service attribute combinations, planners can reduce the probability of costly health service design failures.

INFORMATION PROCESSING DEFICITS ASSOCIATED WITH DEPRESSIVE DISORDERS

A considerable body of evidence suggests that demographic characteristics such as age, education, and illness acuity influence health service preferences (N. K. Arora & McHorney, 2000; Dwight-Johnson, Sherbourne, Liao, & Wells, 2000; F. R. Johnson *et al.*, 2007). Although contextual factors influencing the mood of informants affect their choices (Caruso & Shafir, 2006), we know relatively little about the ways in which the mental health problems associated with many acute and chronic health conditions might influence service preferences. We know even less about the ways in which mental health problems might interact with the different methods used to assess health service preferences.

The prevalence of depressive disorders, for example, is substantially higher among patients with a wide range of acute and chronic diseases (Katon, Lin, & Kroenke, 2007; Lane, Carroll, Ring, Beevers, & Lip, 2002). There are a variety of mechanisms via which depression might influence different stages of the information and decision making processes active in the context of discrete choice conjoint experiments (Bakken, 2007). The visual and spatial recognition deficits associated with depressive disorders (Rubinsztein, Michael, Underwood, Tempest, & Sahakian, 2006), for example, might affect the early processing of information presented via either text or more complex visual displays.

Depressive disorders have also been associated with both mood congruent attentional biases (Erickson *et al.*, 2005; Gotlib *et al.*, 2004; Gotlib, Krasnoperova, Yue, & Joormann, 2004) and an affective bias for negative stimuli (Murphy *et al.*, 1999). The choice tasks included in discrete choice conjoint experiments examining health service options often ask patients to trade the potential benefits of treatment against the inevitable risk of harm. Preferences may shift as a function of the extent to which depressed informants selectively attend to negative attributes of a treatment or service.

Discrete choice conjoint experiments place considerable demands on the effortful processing required to scan and consider the relative importance of health service attributes composing

competing choice task concepts. The processing demands of discrete choice conjoint experiments increase as a function of attribute complexity, the number of attributes included in full profile studies, or the number of attributes presented in each partial profile choice task (Patterson & Chrzan, 2004). Depression might be expected to limit the effort patients deploy to more complex choice task decisions. Moreover, the demands of discrete choice conjoint experiments, in combination with information processing deficits associated with depression, may reduce the threshold at which simplifying heuristics are activated (Payne, Bettman, & Johnson, E. J., 1993).

The routine management of chronic health conditions or specialized diagnostic procedures, such as colonoscopy, can be a source of considerable discomfort (Janz *et al.*, 2007). The treatment of acute illnesses, such as heart attacks or cancer, is associated with an increased risk of post traumatic stress disorder and depression (Spindler & Pedersen, 2005). The episodic autobiographical memory impairments (Moore & Zoellner, 2007) and mood-congruent memory biases associated with depressive disorders (Watkins, Vache, Verney, Muller, & Mathews, 1996), may alter the recall of health service experiences that influence the preferences mediating response to discrete choice conjoint experiments.

Two stage information processing models suggest that discrete choice conjoint experiments would activate both fast, automatic associative processing (stage 1) and more effortful, reflective (stage 2) processes (Beevers, 2005). According to this model, health service delivery attributes with stressful content (e.g. the possibility of poor health outcomes) would activate stage 1's automatic emotional processing (Beevers, 2005). Although automatic associations or negative processing biases could, with effort, be overridden by more conscious stage two processes, the resources available for more reflective processing are limited and may be impaired in depressed individuals (Beevers, 2005). Depression for example, is associated with task irrelevance, intrusive thoughts that may distract informants and deplete the cognitive resources necessary to engage in more effortful reflective stage 2 processing (Beevers, 2005).

To the extent that depression prevents the regulation of emotional responses to the attributes presented in a discrete choice conjoint task, depressed individuals might develop counterproductive decision making heuristics. Informants might act to alter their mood states by avoiding stressful, though potentially effective, service options, or choosing options that might improve their mood (Caruso & Shafir, 2006). In addition to influencing performance on discrete choice conjoint experiments, the extent to which depression disrupts the stage 2 correction of automatic emotional processing may increase vulnerability to depression, prompt counterproductive service delivery decisions, and contribute to poorer longer term outcomes (Beevers, 2005).

Finally, depression is associated with set shifting deficits and a tendency to dichotomous thinking that might reduce the flexibility needed to adjust choices to differing attribute level combinations in discrete choice tasks (Rubinsztein *et al.*, 2006).

THE PRESENT STUDY

This study examined the influence of depressed mood on performance in discrete choice conjoint experiments in the context of a larger program of research examining the information preferences of parents seeking help for children with mental health problems. The growing body of scientific knowledge regarding the etiology, prevalence, treatment, and longitudinal course of children's mental health problems is, increasingly, available in the public domain as brochures, books, video tapes, public lectures, workshops, internet sites, and direct pharmaceutical advertisements. Indeed, the development and provision of information is a component of professional practice guidelines (Leslie, Weckerly, Plemmons, Landsverk, & Eastman, 2004) and hospital accreditation processes. Although the quality of the health information available to the public is often poor (Bowskill, Clatworthy, Parham, Rank, & Horne, 2007; Coulter, Entwistle, & Gilbert, 1999; Godolphin, Towle, & McKendry, 2001; Lissman & Boehnlein, 2001) high quality information services offer considerable benefit to families of children with mental health problems. To make informed decisions regarding services for their children, for example, parents must consider the risks and benefits of different treatments, the logistical demands of each option, and the outcomes associated with the decision not to treat (Wills & Holmes-Rovner, 2006). Parents who are better informed regarding the etiology, developmental course, and treatment of childhood mental health problems are more likely to utilize available service options (Andrews, J. N., Swank, P. R., Foorman, B., Fletchers, J. M., 1995; Corkum, Rimer, & Schachar, 1999; Johnston, Seipp, Hommersen, Hoza, & Fine, 2005).

Information can also be integrated into tools which help patients make evidence-informed decisions regarding complex treatment options. Decision aids improve knowledge regarding treatments, contribute to more realistic outcome expectations, reduce the discrepancy between patient expectations and health service priorities, lower decisional conflict, and improve health outcomes (O'Connor *et al.*, 2007; Ruland, 1999).

Finally, information can be translated into books, videotapes, or internet sites designed to teach solutions to common childhood problems. Systematic reviews of the randomized trials in this area suggest that these types of media-based interventions can yield moderate improvements in childhood behavioral problems (Montgomery, Bjornstad, & Dennis, 2006).

We consider four ways in which the information processing biases associated with depressed mood might influence the response of informants to discrete choice conjoint experiments. First, we postulated that depressed mood would reduce informant reliability as measured by a consistent response to identical hold-out tasks (R. M. Johnson, 1997). Second, we predicted that depressed mood would increase the percentage of informants making illogical health information choices on hold out tasks. Given less consistent responses to hold out tasks and more illogical choices, we predicted that simulations of the choices depressed informants made on hold out tasks would be less accurate than those of non depressed parents. Finally, we predicted that depressed mood would be associated with unique information service preferences and different segment membership.

METHODS

Participants

Interviewers at four central children's mental health service intake sites in Ontario, Canada asked parents of 6 to 18 year olds to consider participating in a study of their information preferences. The research team contacted those parents agreeing to participate, provided a description of the study's rationale and methods, and answered questions. Of these, 1180 completed a useable discrete choice conjoint survey, a total return rate of 49%. As a study conducted according to university research ethics board protocols, participants endorsed consent forms assuring confidentiality, the option of refusing to participate, or the right to withdraw without consequence.

Discrete Choice Conjoint Survey Design

We derived 20 children's mental health information content, process, and outcome attributes from 6 focus groups with parents of children with mental health problems (3 groups with fathers and 3 groups with mothers). We composed attributes that did not require the prohibition of incompatible attribute level combinations (Orme, 2006) and operationalized all attributes with 4 levels to avoid the biasing effects of attributes with a greater number of levels (Orme, 2006). We asked several parents and children's mental health professionals to complete and comment on a pilot version of the survey and used this feedback to revise the wording of the study's attributes (Ryan & Gerard, 2003).

While increasing the number of attributes presented in each choice concept improves mathematical efficiency, informant efficiency declines (Patterson & Chrzan, 2004). Pilot studies and previous research at our Patient-Centred Service Research Unit confirm that it is virtually impossible for informants to process full profile choice tasks composed of health service attributes which are complex, abstract, or unfamiliar. In order to maximize informant efficiency, we used a partial profile CBC design with 30 choice tasks, each presenting 3 concepts defined by two attribute levels (G. Allenby *et al.*, 2005). To obtain an additional increase in informant efficiency, and to increase the percentage of parents who completed the survey, we inserted 4 spacers depicting progress after choice tasks 1, 9, 19, and 30. The D-efficiency (Kuhfeld, Tobias, & Garratt, 1994) of this design was .99, approaching an optimal level of 1.0.

Hold Out Tasks

We included identical hold out choice tasks at positions 2 and 21. We attempted to avoid concepts which would dominate choices among the three options presented in each holdout choice task, to compose concepts which were not equally attractive, and to include concepts with differential appeal to the segments that we thought would emerge (R. M. Johnson, 1997). We used hold out tasks in three ways. First, we identified respondents making what the research team judged to be illogical choices (R. M. Johnson, 1997). Inconsistent or illogical responses to hold out tasks can help determine whether informants understood the task, were adequately motivated, or were simply responding randomly (R. M. Johnson, 1997). Second, as a measure of reliability, we determined the extent to which informants responded consistently when identical choice tasks were presented (R. M. Johnson, 1997). Finally, we used randomized first choice simulations to predict hold out task choices (R. M. Johnson, 1997) Although hold out tasks provide an opportunity to improve predictions by adjusting the scale parameter (R. M. Johnson, 1997), our simulations suggested that rescaling was not required.

Supplementary Questions

Child Characteristics. Before beginning the discrete choice conjoint survey, parents completed the Brief Child and Family Phone Interview (BCFPI), a standardized intake screening measure used by all provincially funded children's mental health service providers in Ontario (C. E. Cunningham, Pettingill, & Boyle, 2007). Using Likert scales, parents rated the following 6-item children's mental health scales (1) attention, impulsivity, and overactivity, (2) oppositional behavior, (3) conduct problems, (4) separation anxiety disorders, (5) generalized anxiety disorders, and (6) child mood disorders. Parents also rated 7 items on the extent to which their child's problems adversely affected: (1) social activities, (2) relationships with parents, teachers, and peers, and (3) school performance and achievement. Finally, parents rated 7 items on the extent to which the child's difficulties adversely affected: (1) family activities, and (2) conflict and anxiety regarding the child's difficulties.

Parental depression. After completing the discrete choice conjoint survey, parents reported symptoms of depression on the 6-item parental mood scale of the BCFPI. The internal consistency of this scale was 0.86. For the analyses presented here, parents reporting BCFPI parental mood t-scores equal to or greater than 70 (above the 98th percentile for the general population), were categorized as having high depression scores.

Data Analysis

We used Sawtooth Software's CBC/HB Hierarchical Bayes Estimation version 3.0 to compute choice-based preference parameter estimates for each participant (G. M. Allenby, Arora, & Gintner, 1995; N. Arora, Allenby, & Ginter, 1998; Lenk, DeSarbo, Green, & Young, 1996; Orme, 2006). Next we used Latent Class Version 3 (Sawtooth Software Inc.) to identify segments of parents with similar children's mental health information preferences (DeSarbo, Ramaswamy, & Cohen, 1995). The latent class formula finds clusters of respondents with similar preferences and utilities, simultaneously computing the probability of membership in each segment (DeSarbo *et al.*, 1995). We replicated the latent class solution 5 times beginning at random starting points, assuming convergence when log-likelihood decreased by less than 0.01, and accepting the solution with the best fit and the lowest Consistent Akaike Information Criterion (CAIC). We used multinomial logit to fit a set of utilities to each participant's choice task data and standardized (zero-centred) the part-worth utility values setting the average range within the utility values of all attributes to 100 (Orme, 2006). We computed standardized importance scores by converting the range of each attribute's part-worth utility values to a percentage of the sum of the utility value ranges of all attributes (Orme, 2006).

We used independent samples t-tests corrected for heterogeneity of variance to compare the Brief Child and Family Phone Interview's child mental health, child functioning, and family functioning scores of parents with and without high depression scores. We computed chi square analyses to compare the proportion of depressed and non-depressed parents who responded consistently to hold out tasks, preferred different service options, and whose hold out choices we predicted accurately.

Finally, we computed Randomized First Choice Simulations to determine the response of depressed and non-depressed parents to three children's mental health information service options: (1) a *Waiting List as usual* option in which therapists provided evidence-based information at parental request, (2) an *Information Model* in which therapists recommended evidence-based materials that helped parents understand their child's difficulties, or (3) an *Active*

Learning model in which therapists recommended evidence-based interactive materials that, with the help of weekly phone calls from a therapist, helped parents develop solutions to their child’s behavioral and emotional problems. Randomized first choice share of preference simulations employ a maximum utility rule assuming that participants choose the option with the highest overall utility. This approach to simulation reduces concerns regarding IIA (red bus blue bus) problems and improves choice share predictions by estimating both attribute and product variability (Orme, 2006).

RESULTS

Question 1: Did Depression Reduce Consistent Responding to Hold Out Tasks.

Table 1 shows that 22% of our participants met the criteria for inclusion in the depressed group. The proportion of depressed versus non-depressed parents who responded inconsistently to identical hold out tasks did not differ significantly, $X^2(2, 1074) = .046, p = .829$ (Table 1). Although depression was not associated with inconsistent response patterns, 42.4% of the parents in this study failed to respond consistently to repeated hold out tasks, an issue we discuss below.

Table 1

Percentage of parents in the non depressed and depressed groups responding “logically” to hold out tasks 1 and 2, consistently across hold out tasks 1 and 2, in the Overwhelmed segment, preferring a waiting list option rather than information option (), and whose hold out choices were accurately predicted

Measure	Non Depressed	Depressed	X^2	<i>p</i>
Sample Size	845	235		
Consistent Response to Hold Out Tasks	57.8	57.0	.05	.829
“Logical” Response to Hold Out Task 1	25.8	24.7	.41	.815
“Logical” Response to Hold Out Task 2	40.9	32.3	6.36	.042
Percent in Overwhelmed Segment	11.8	24.3	25.14	<.001
Percent Preferring to Wait	18.9	32.0	19.31	<.001
Hit Rate for Hold Out Task 1	46.0	49.1	.74	.389
Hit Rate for Hold Out Task 2	56.9	55.2	.22	.638

Question 2: Did Depression Increase Illogical Responses to Hold Out Tasks?

Focus group discussions, a previous study of the service delivery preferences of hospital patients, and the a priori assumptions of the research team suggested that the logical response to our hold out tasks was to choose a concept in which all parents were automatically given information about children’s mental health problems that was specific to their child and family. Table 1 shows that, at the first choice task, the proportion of depressed and non depressed parents making a logical choice did not differ, $X^2(2, 1077) = .41, p = .815$. At the second choice task, in contrast, depressed parents were more likely than non depressed parents to choose an illogical concept in which information was not specific to their child or family and was provided only if parents asked, $X^2(2, 1077) = 6.36, p = 0.042$. Although the proportion of parents making a “logical” choice increased at the second hold out, a significant percentage of parents preferred alternative concepts.

Question 3: Did Depression Influence Segment Membership

Latent Class segmentation analysis yielded a 3 group solution. A total of 42% of the sample were members of an *Action* segment that preferred materials providing brief, evidence-based, step-by-step, solutions to behavioral or emotional problems. Parents in the Action segment preferred that information was supported by calls from a therapist. An additional 40% of the sample was in an *Information* segment preferring materials which helped them understand but not solve their child's problems. Finally, 17% of parents were members of an *Overwhelmed* segment that chose not to utilize materials providing active learning, step-by-step solutions to child mental health problems. These parents did not choose therapist assistance or group support. Table 1 shows that parents with high depression scores were twice as likely as non depressed parents to be members of the Overwhelmed segment, $X^2(2, 1080) = 25.14, p < 0.001$.

Figure 1 depicts standardized (zero-centered) utility values for the Action, Information, and Overwhelmed Segments for an attribute describing the content of the information in children's mental health materials. Parents in the Action segment preferred information which provided a step-by-step guide to solutions to their child's behavioral difficulties. The Information segment preferred materials helping them understand their child's problems. Most parents in the Overwhelmed segment, in contrast, preferred no information helping them solve their child's problems. A similar pattern was observed for information that might help parents develop solutions to emotional problems, advocate on their child's behalf, or consider whether medication might be helpful for their child.

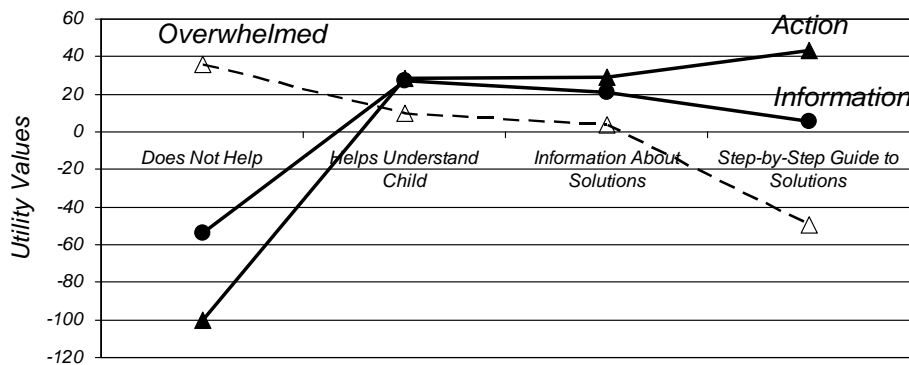


Figure 1: Zero-centered utility values for the Action, Information, and Overwhelmed segments (n = 1028). Attribute levels included information which did not help understand or solve behavior problems, information which helped parents understand their child's behavior problems, materials providing information about solutions, and information teaching step-by-step solutions to child problems.

Randomized first choice simulations of the *Waiting List, Information, and Active Learning* service options suggested that depressed parents were more likely than non-depressed parents to simply wait for children's mental health services rather than using an evidence-based Information or Action Learning option, $X^2(2, 1025) = 18.31, p < 0.001$.

Question 4: Did Depression Influence Related Information Preference Measures

Depressed parents reported that their children had more difficulties with inattention, overactivity, and impulsivity, $t(1013) = -3.18, p = 0.002$, were more oppositional and defiant,

$t(1013) = -3.11, p = 0.002$, engaged in more conduct problems, $t(1013) = -3.32, p = 0.001$, and evidence more symptoms of depression, $t(1013) = -5.5, p < 0.001$, than children of nondepressed parents. Parents felt these problem had a greater negative impact on their child's participation in social activities, $t(1013) = 5.03, p < 0.001$, social relationships, $t(1013) = -3.03, p = 0.003$, and school performance, $t(1013) = -2.54, p < 0.001$. In addition, these problems had a greater negative impact on family activities, $t(1013) = -3.64, p < 0.001$, and were a greater source of conflict and anxiety to the families of depressed versus non depressed parents, $t(1013) = -3.37, p < 0.001$.

Depressed participants felt they encountered more barriers in their attempts to obtain information about their child's problems. They felt that children's mental health information cost too much, $t(1077) = 2.84, p = 0.005$, that information was inconveniently located, $t(1077) = -3.12, p = 0.002$, and that professionals were reluctant to provide information, $t(1077) = 5.03, p < 0.001$. In comparison to non depressed parents, parents with high depression scores felt that information was more likely to leave them feeling stressed, $t(1077) = 3.47, p = 0.001$, and guilty, $t(1077) = 6.02, p < 0.001$.

Question 5: Did Depression Influence the Validity of Hold Out Task Simulations?

We simulated the response of depressed and non depressed parents to hold out tasks and computed hit rates (Orme, 2006). The mean absolute error (MAE) was 8.9 for hold out task 1 and 3.6 for hold out task 2. Table 1 shows that the hit rates for the depressed and nondepressed groups for did not differ at either the first, $X^2(2, 1072) = .74, p = 0.389$, or second holdout task 2, $X^2(2, 1072) = .22, p = 0.638$, did not differ. Hit rates were higher at holdout task 2 than holdout task 1.

DISCUSSION

Although depression influences selective attention, memory, and decision making (Beevers, 2005), significantly elevated depression scores were not associated with an increase in inconsistent responses to two identical hold out tasks. Moreover, we observed only a slight increase in illogical responses to the second holdout task presented in this study. Depression was, however, associated with a significant shift in the health information preferences of parents of children with mental health problems.

Depressed parents were more likely to be members of an Overwhelmed segment that preferred to simply wait for children's mental health services. In contrast to parents in the Action segment or Information segment, those in the Overwhelmed segment did not choose to use interactive materials designed to help them understand and develop step-by-step, therapist-supported solutions to their child's behavioral or emotional problems. Although randomized trials suggest these options would yield a considerable improvement in their child's difficulties (Montgomery *et al.*, 2006), parents in the overwhelmed segment preferred to simply wait for more traditional clinical services.

Our findings are consistent with studies suggesting that the information processing biases associated with depression may prompt counterproductive health service decisions and increase vulnerability to depression (Beevers, 2005). Although depressed parents felt their children presented more behavioral and emotional problems, they were less likely to choose the types of evidence-based information services that might help them understand or contribute to the

solution of their child's difficulties (Montgomery *et al.*, 2006). Although these decisions seem counter intuitive, focus group discussions suggested that information regarding children's mental health problems exposes parents to at least three potential threats. First, children's mental health information may prompt guilt regarding the role parents might have played in the etiology of their child's difficulties. In this study, for example, depressed parents were more likely than non depressed parents to report that children's mental health information left them feeling more guilty. Second, information might elicit anxiety regarding the longer term risks associated with childhood mental health problems. These might include rejection by peers, academic failures, delinquency, adult psychiatric disorders, and longer term impairments in social and vocational functioning. Again, depressed parents were more likely than non depressed parents to report that information was a source of stress. Third, depression is associated with a negative interpersonal information bias. In information processing paradigms, for example, depressed individuals show an attentional bias to sad faces (Gotlib, Krasnoperova *et al.*, 2004). Several of the attributes included in this study depicted the presentation of information to groups of parents, the opportunity to meet with other parents working through self-paced educational materials, or the option of weekly telephone help from a therapist. Given negative social information processing biases (Gotlib *et al.*, 2004; Gotlib, Krasnoperova *et al.*, 2004), social skills deficits (Segrin, 2000), and a tendency to underestimate their interpersonal competence (Gotlib & Meltzer, 1987), it is not surprising that depressed parents might be hesitant to pursue information in a social context.

While the information processing biases associated with depression may predispose parents to respond to negative attributes of the information sources depicted in our study (Kerr, Scott, & Phillips, 2005; Rinck & Becker, 2005), depression may also disrupt the effortful, reflective processes needed to put this information in context. Depressed parents, for example, might have difficulty considering the protective factors that may reduce longer term risk, the ways in which parents can assist their children, or evidence that many children recover from early mental health difficulties.

In addition to altering service delivery preferences, the information processing biases associated with parental depression can contribute to a negatively distorted evaluation of their child's difficulties (Chi & Hinshaw, 2002). In the present study, depressed parents felt their child had more behavioural and emotional difficulties, more functional impairments, and a more adverse impact on family functioning. A distorted evaluation of the child's difficulties can contribute to an approach to parenting that adversely effects the child's development and adjustment (Elgar, Mills, McGrath, Waschbusch, & Brownridge, 2007; Maughan, Cicchetti, Toth, & Rogosch, 2007).

The materials depicted in this study require parents to devote time to acquiring, reading, reviewing, and applying new information to the solution of their child's problems. Depression, and the difficulties associated with the management of more challenging children, may reduce the resources parents are able to devote to this process. Our focus groups suggested that these time requirements represent insurmountable barriers to some parents. Not surprisingly, depressed parents in this study were more likely to feel that children's mental health information was inconveniently located, expensive, and difficult to obtain from health professionals.

A number of investigators have suggested that the complexity of discrete choice conjoint experiments activates the decision making heuristics which influence real world choices

(Phillips, Johnson, & Maddala, 2002). It has also been suggested that complex choice tasks reduce the social desirability biases that influence responses to single question ratings (Phillips *et al.*, 2002). Indeed, the service preferences observed in our DCE study, were not evident on simple questions where most parents stated they were interested in books, videos, or parenting groups. Our study's utility values and simulations, in contrast, suggested that a larger percentage of depressed parents would not use these resources. These simulation are consistent with studies suggesting that depression and associated stressors reduce the utilization of children's mental health information services (C. E. Cunningham, Bremner, & Boyle, 1995).

Results show that, while depression influenced segment membership, the preferences of depressed parents were associated with a much a broader range of closely-related factors. Depressed parents, for example, reported more child behavior problems and more impairments in their children's activities, social relationships, and academic performance. Depressed parents felt their children's difficulties had a greater impact on family activities, and were a greater source of conflict and anxiety. All of these factors may contribute to the service preferences observed here.

LIMITATIONS

Before discussing the implications of our findings, it is important to consider the limitations of our study. The extent to which our findings can be generalized to the use of discrete conjoint experiments with other patient groups is limited by several factors. First, we examined the effects of depressed mood in parents seeking mental health services for their children. Although these parents reported more symptoms of depression than 98% of the general population, they may not have shown the associated functional impairments observed in adult patients seeking services for major depressive disorders. Given evidence that even minor shifts in mood can alter choices (Caruso & Shafir, 2006), the effects of more severe depressive disorders on discrete choice conjoint tasks may well be more pronounced than those observed here.

The effects of depression may vary as a function of the attributes presented in discrete choice conjoint experiments. It is quite possible that the information processing biases associated with depressive disorders apply to health service attributes which are personally relevant and affectively loaded. Depressed informants might respond differently to tasks presenting choices between emotionally sensitive health service attributes than to tasks composed of attributes regarding consumer goods which are of less importance.

IMPLICATIONS FOR THE DESIGN OF DISCRETE CHOICE CONJOINT EXPERIMENTS

In discrete choice conjoint experiments examining health service preferences, participants may be unfamiliar with the attributes under study. Moreover, the service delivery attributes informants are asked to consider may be complex or abstract. To simplify choices and maximize informant efficiency, we used a partial profile design (Patterson & Chrzan, 2004) (G. Allenby *et al.*, 2005). The effects of depression might be more pronounced in studies using full-profile designs or surveys with choice tasks presenting a greater number of attribute levels per concept (Patterson & Chrzan, 2004).

Although the influence of depression on the preferences measured via discrete choice conjoint experiments might shift with successful treatment or symptom resolution, a number of

studies suggest that the information processing effects of depression persist after symptoms have resolved (Joormann & Gotlib, 2007). Treatments such as cognitive behavior therapy, which focus on stage 2 information processing, might yield different effects on preferences than pharmacotherapy alone (Teasdale *et al.*, 2001). This question merits further research.

The health service professionals to whom we consult are often unfamiliar with the design and interpretation of discrete choice conjoint experiments. They pose many of the same questions raised by product managers (Orme, 2006). Do patients understand the nature of choice tasks? Are the choices patients are asked to make too complex? Are patients responding randomly? How exactly do we translate choice data into shares of preference? Are these predictions valid? Although attributes with logically ordered utility values, simulations which predict health service utilization (C. E. Cunningham *et al.*, 2003), and the heuristic contribution of these methods to the service planning process (C. E. Cunningham *et al.*, 2005) provide our colleagues with an important measure of face validity, information regarding more familiar principles such as informant consistency (reliability) and the predictive validity of hold out task simulations increases the credibility of our recommendations and the probability that our results will inform service design decisions. Given the information they provide regarding service preferences, and their contribution to the credibility of our methods, we routinely include two identical hold out tasks in all discrete choice conjoint surveys conducted by our Patient-Centred Service Research Unit.

Johnson suggested that the concepts in hold out task choice sets should include attributes of moderate importance to informants (R. M. Johnson, 1997). In many health service contexts, however, there are relatively few studies that could provide a priori information about patient treatment or outcome preferences. In the present study, an experienced team of health service researchers had difficulty estimating the relative importance of the information content, process, and outcome attributes included in our survey. Our experience is consistent with previous studies suggesting that the preferences of patients and professionals are often quite different (C. E. Cunningham *et al.*, 2005; Pieterse *et al.*, 2007). Information choices which seemed illogical to the research team were preferred by a small but important segment of overwhelmed parents. Given relatively low importance scores, however, the attributes selected for inclusion in hold out tasks exerted little influence on the choices of participants. Not surprisingly, the choices of a significant percentage of informants shifted across the 30 choice tasks included in the survey. Under these circumstances, the hit rates observed in this study confirm that it is difficult to predict hold out task choices (R. M. Johnson, 1997).

In discrete choice conjoint experiments, informants often show a preference for familiar attributes (Salkeld, Ryan, & Short, 2000). The pattern of choices observed in this study suggests that inconsistent responses to identical hold out tasks reflected a shift in preferences as choice tasks prompted informants to consider the relative importance of unfamiliar health service attributes (G. Allenby *et al.*, 2005). As others have noted, hit rates for hold out task simulations increased while Mean Absolute Errors (MAE) decreased across choice tasks (G. Allenby *et al.*, 2005). In this study, for example, attributes depicting logistical features of the information exerted less influence on later choice tasks while attributes depicting the content and outcome of information became more important. Although some propose that informants who respond inconsistently to identical hold out tasks should not be included in data analysis (R. M. Johnson, 1997; Pinnell, 2005), this would have resulted in the loss of almost 50% of our sample. Evidence that preferences shift meaningfully as a function of exposure to a relatively large

number of choice tasks suggests that, in conjoint studies introducing unfamiliar product or service attributes, inconsistent responses are not simply a measure of reliability.

Finally, depression is associated with complex shifts in the information processing and decision making mechanisms which influence performance on discrete choice conjoint experiments. The information processing deficits observed in depressed patients are evident prior to the onset of depressive symptoms, may persist once symptoms have resolved, and are associated with a wider range of the psychiatric disorders that often accompany health problems (Jongen, Smulders, Ranson, Arts, & Krabbendam, 2007; Kerr *et al.*, 2005; Rinck & Becker, 2005). The relatively high prevalence of depression and related psychiatric disorders in patient populations emphasizes the importance of methods such as latent class and hierarchical Bayes that account for informant heterogeneity and can inform the development of health services that better meet the needs of these patients (G. Allenby *et al.*, 2005; Orme, 2006).

REFERENCES

- Ahmed, S. F., Blamires, C., & Smith, W. (2007). Facilitating and understanding the family's choice of injection device for growth hormone therapy by using conjoint analysis. *Archives of Disease in Childhood*.
- Allenby, G. M., Arora, N., & Gintner, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32, 152-162.
- Allenby, G., Fennell, G., Huber, J., Eagle, T., Gilbride, T., Horsky, D., *et al.* (2005). Adjusting choice models to better predict market behavior. *Marketing Letters*, 16(3-4), 197-208.
- Andrews, J. N., Swank, P. R., Foorman, B., Fletchers, J. M. (1995). Effects of educating parents about ADHD. *The ADHD Report*, 3, 12-13.
- Arora, N., Allenby, G. M., & Ginter, J. L. (1998). A hierarchical bayes model of primary and secondary demand. *Marketing Science*, 17(1), 29-44.
- Arora, N. K., & McHorney, C. A. (2000). Patient preferences for medical decision making: Who really wants to participate? *Medical Care*, 38(3), 335-341.
- Bakken, D. G. (2007). The weakest link: A cognitive approach to improving survey data quality. Sawtooth Software Conference Proceedings, Sequim, WA.
- Basen-Engquist, K., Fouladi, R. T., Cantor, S. B., Shinn, E., Sui, D., Sharman, M., *et al.* (2007). Patient assessment of tests to detect cervical cancer. *International Journal of Technology Assessment in Health Care*, 23(2), 240-247.
- Beevers, C. G. (2005). Cognitive vulnerability to depression: A dual process model. *Clinical Psychology Review*, 25(7), 975-1002.
- Beusterien, K. M., Dziekan, K., Flood, E., Harding, G., & Jordan, J. C. (2005). Understanding patient preferences for HIV medications using adaptive conjoint analysis: Feasibility assessment. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 8(4), 453-461.
- Bowskill, R., Clatworthy, J., Parham, R., Rank, T., & Horne, R. (2007). Patients' perceptions of information received about medication prescribed for bipolar disorder: Implications for informed choice. *Journal of Affective Disorders*, 100(1-3), 253-257.
- Caldon, L. J., Walters, S. J., Ratcliffe, J., & Reed, M. W. (2007). What influences clinicians' operative preferences for women with breast cancer? an application of the discrete choice experiment. *European Journal of Cancer (Oxford, England : 1990)*, 43(11), 1662-1669.
- Caruso, E. M., & Shafir, E. (2006). Now that I think about it, I'm in the mood for laughs: Decisions focused on mood. *Journal of Behavioral Decision Making*, 19(2), 155-169.
- Chi, T. C., & Hinshaw, S. P. (2002). Mother--child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology*, 30(4), 387(14).

- Chinburapa, V., Larson, L. N., Brucks, M., Draugalis, J., Bootman, J. L., & Puto, C. P. (1993). Physician prescribing decisions: The effects of situational involvement and task complexity on information acquisition and decision making. *Social Science & Medicine*, 36(11), 1473-1482.
- Corkum, P., Rimer, P., & Schachar, R. (1999). Parental knowledge of attention-deficit hyperactivity disorder and opinions of treatment options: Impact on enrollment and adherence to a 12-month treatment trial. *Canadian Journal of Psychiatry*, 44(10), 1043-1048.
- Coulter, A., Entwistle, V., & Gilbert, D. (1999). Sharing decisions with patients: Is the information good enough? *British Medical Journal*, 318(7179), 318-322.
- Cunningham, C. E., Evans, R., Buchanan, D., Kostrzewa, L., Miller, H., Tobon, J., *et al.* (2005). Modelling inpatient youth mental health service preferences using conjoint analysis. *Joint Annual Meeting of the Canadian and American Academies of Child and Adolescent Psychiatry*. Toronto, ON.
- Cunningham, C. E., Pettingill, P. & Boyle, M. H. (2007). The brief child and family phone interview version 3: Interviewer's manual.
- Cunningham, C. E., Boyle, M., Offord, D., Racine, Y., Hundert, J., Secord, M., *et al.* (2000). Tri-ministry study: Correlates of school-based parenting course utilization. *Journal of Consulting and Clinical Psychology*, 68(5), 928-933.
- Cunningham, C. E., Bremner, R., & Boyle, M. (1995). Large group community-based parenting programs for families of preschoolers at risk for disruptive behaviour disorders: Utilization, cost effectiveness, and outcome. *Journal of child psychology and psychiatry, and allied disciplines*, 36(7), 1141-1159.
- Cunningham, C. E., Deal, K., Neville, A., Rimas, H., & Lohfeld, L. (2006). Modeling the problem-based learning preferences of McMaster university undergraduate medical students using a discrete choice conjoint experiment. *Advances in Health Sciences Education: Theory and Practice*, 11(3), 245-266.
- Cunningham, C. E., Buchanan, D., & Deal, K. (2003). Modelling patient-centred children's health services using choice-based conjoint hierarchical bayes. *10th Annual Sawtooth Software Conference*, San Antonio, TX. 249-256.
- DeSarbo, W. S., Ramaswamy, V., & Cohen, S. H. (1995). Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6(2), 137-147.
- Dwight-Johnson, M., Lagomasino, I. T., Aisenberg, E., & Hay, J. (2004). Using conjoint analysis to assess depression treatment preferences among low-income latinos. *Psychiatric Services*, 55(8), 934-936.
- Dwight-Johnson, M., Sherbourne, C. D., Liao, D., & Wells, K. B. (2000). Treatment preferences among depressed primary care patients. *Journal of General Internal Medicine: Official Journal of the Society for Research and Education in Primary Care Internal Medicine*, 15(8), 527-534.

- Elgar, F. J., Mills, R. S. L., McGrath, P. J., Waschbusch, D. A., & Brownridge, D. A. (2007). Maternal and paternal depressive symptoms and child maladjustment: The mediating role of parental behavior. *Journal of Abnormal Child Psychology*, *35*(6), 943-955.
- Erickson, K., Drevets, W. C., Clark, L., Cannon, D. M., Bain, E. E., Zarate, C. A., Jr., *et al.* (2005). Mood-congruent bias in affective Go/No-go performance of unmedicated patients with major depressive disorder. *American Journal of Psychiatry*, *162*(11), 2171-2173.
- Fraenkel, L., Gulanski, B., & Wittink, D. (2007). Patient willingness to take teriparatide. *Patient Education and Counseling*, *65*(2), 237-244.
- Godolphin, W., Towle, A., & McKendry, R. (2001). Evaluation of the quality of patient information to support informed shared decision-making. *Health Expectations*, *4*(4), 235-242.
- Gotlib, I. H., Kasch, K. L., Traill, S., Joormann, J., Arnow, B. A., & Johnson, S. L. (2004). Coherence and specificity of information-processing biases in depression and social phobia. *Journal of Abnormal Psychology*, *113*(3), 386-398.
- Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology*, *113*(1), 121-135.
- Janz, N. K., Lakhani, I., Vijan, S., Hawley, S. T., Chung, L. K., & Katz, S. J. (2007). Determinants of colorectal cancer screening use, attempts, and non-use. *Preventive Medicine*, *44*(5), 452-458.
- Johnson, F. R., Ozdemir, S., Manjunath, R., Hauber, A. B., Burch, S. P., & Thompson, T. R. (2007). Factors that affect adherence to bipolar disorder treatments: A stated-preference approach. *Medical Care*, *45*(6), 545-552.
- Johnson, R. M. (1997). *Including holdout choice tasks in conjoint studies*. Sequim Washington: Sawtooth Software Research Paper Series. from <http://www.sawtoothsoftware.com/techpap.shtml>
- Johnston, C., Seipp, C., Hommersen, P., Hoza, B., & Fine, S. (2005). Treatment choices and experiences in attention deficit and hyperactivity disorder: Relations to parents' beliefs and attributions. *Child: Care, Health and Development*, *31*(6), 669-677.
- Jongen, E. M., Smulders, F. T., Ranson, S. M., Arts, B. M., & Krabbendam, L. (2007). Attentional bias and general orienting processes in bipolar disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 168-183.
- Joormann, J., & Gotlib, I. H. (2007). Selective attention to emotional faces following recovery from depression. *Journal of Abnormal Psychology*, *116*(1), 80-85.
- Katon, W., Lin, E. H., & Kroenke, K. (2007). The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry*, *29*(2), 147-155.
- Kazdin, A. E., Holland, L., & Crowley, M. (1997). Family experience of barriers to treatment and premature termination from child therapy. *Journal of Consulting and Clinical Psychology*, *65*(3), 453-463.

- Kerr, N., Scott, J., & Phillips, M. L. (2005). Patterns of attentional deficits and emotional bias in bipolar and major depressive disorder. *The British Journal of Clinical Psychology*, 44(Pt 3), 343-356.
- Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4), 545-557.
- Lane, D., Carroll, D., Ring, C., Beevers, D. G., & Lip, G. Y. (2002). The prevalence and persistence of depression and anxiety following myocardial infarction. *British Journal of Health Psychology*, 7(Pt 1), 11-21.
- Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, Vol. 15, 2, pp. 173-191.
- Leslie, L. K., Weckerly, J., Plemmons, D., Landsverk, J., & Eastman, S. (2004). Implementing the american academy of pediatrics attention-deficit/hyperactivity disorder diagnostic guidelines in primary care settings. *Pediatrics*, 114(1), 129-140.
- Lewis, S. M., Cullinane, F. M., Carlin, J. B., & Halliday, J. L. (2006). Women's and health professionals' preferences for prenatal testing for down syndrome in australia. *The Australian & New Zealand Journal of Obstetrics & Gynecology*, 46(3), 205-211.
- Lissman, T. L., & Boehnlein, J. K. (2001). A critical review of internet information about depression. *Psychiatric Services*, 52(8), 1046-1050.
- Maughan, A., Cicchetti, D., Toth, S. L., & Rogosch, F. A. (2007). Early-occurring maternal depression and maternal negativity in predicting young Children's emotion regulation and socioemotional difficulties. *Journal of Abnormal Child Psychology*, 35(5), 685-703.
- McDonald, H. P., Garg, A. X., & Haynes, R. B. (2002). Interventions to enhance patient adherence to medication prescriptions: Scientific review. *Journal of the American Medical Association*, 288(22), 2868-2879.
- McGregor, J. C., Harris, A. D., Furuno, J. P., Bradham, D. D., & Perencevich, E. N. (2007). Relative influence of antibiotic therapy attributes on physician choice in treating acute uncomplicated pyelonephritis. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, 27(4), 387-394.
- Montgomery, P., Bjornstad, G., & Dennis, J. (2006). Media-based behavioural treatments for behavioural problems in children. *Cochrane Database of Systematic Reviews (Online)*, (1)(1), CD002206.
- Moore, S. A., & Zoellner, L. A. (2007). Overgeneral autobiographical memory and traumatic events: An evaluative review. *Psychological Bulletin*, 133(3), 419-437.
- Murphy, F. C., Sahakian, B. J., Rubinsztein, J. S., Michael, A., Rogers, R. D., Robbins, T. W., et al. (1999). Emotional bias and inhibitory control processes in mania and depression. *Psychological Medicine*, 29(6), 1307-1321.
- O'Connor, A. M., Bennett, C., Stacey, D., Barry, M. J., Col, N. F., Eden, K. B., et al. (2007). Do patient decision aids meet effectiveness criteria of the international patient decision aid

- standards collaboration? A systematic review and meta-analysis. *Medical Decision Making*, 27(5), 554-574.
- Orme, B. K. (2006). Getting started with conjoint analysis: Strategies for product design and pricing research. *Madison: Research Publishers*.
- Oudhoff, J. P., Timmermans, D. R., Knol, D. L., Bijnen, A. B., & Van der Wal, G. (2007). Prioritising patients on surgical waiting lists: A conjoint analysis study on the priority judgements of patients, surgeons, occupational physicians, and general practitioners. *Social Science & Medicine*, 64(9), 1863-1875.
- Owens, P. L., Hoagwood, K., Horwitz, S. M., Leaf, P. J., Poduska, J. M., Kellam, S. G., *et al.* (2002). Barriers to children's mental health services. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41(6), 731-738.
- Patterson, M., & Chrzan, K. (2004). Partial profile discrete choice: What's the optimal number of attributes? *2003 Sawtooth Software Conference Proceedings*, San Antonio, TX.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The use of multiple strategies in judgment and choice*.
- Phillips, K. A., Johnson, F. R., & Maddala, T. (2002). Measuring what people value: A comparison of "attitude" and "preference" surveys. *Health services research*, 37(6), 1659-1679.
- Phillips, K. A., Maddala, T., & Johnson, F. R. (2002). Measuring preferences for health care interventions using conjoint analysis: An application to HIV testing. *Health Services Research*, 37(6), 1681-1705.
- Pieterse, A. H., Stiggelbout, A. M., Baas-Thijssen, M. C., van de Velde, C. J., & Marijnen, C. A. (2007). Benefit from preoperative radiotherapy in rectal cancer treatment: Disease-free patients' and oncologists' preferences. *British Journal of Cancer*, 97(6), 717-724.
- Pinnell, J. (2005). Practical suggestions for CBC studies. *2004 Sawtooth Software Conference Proceedings*, San Diego, CA. 43-47.
- Rinck, M., & Becker, E. S. (2005). A comparison of attentional biases and memory biases in women with social phobia and major depression. *Journal of Abnormal Psychology*, 114(1), 62-74.
- Rubinsztein, J. S., Michael, A., Underwood, B. R., Tempest, M., & Sahakian, B. J. (2006). Impaired cognition and decision-making in bipolar depression but no 'affective bias' evident. *Psychological Medicine*, 36(5), 629-639.
- Ruland, C. M. (1999). Decision support for patient preference-based care planning: Effects on nursing care and patient outcomes. *Journal of the American Medical Informatics Association*, 6(4), 304-312.
- Ryan, M., & Gerard, K. (2003). Using discrete choice experiments to value health care programmes: Current practice and future research reflections. *Applied Health Economics and Health Policy*, 2(1), 55-64.

- Ryan, M., Scott, D. A., Reeves, C., Bate, A., van Teijlingen, E. R., Russell, E. M., *et al.* (2001). Eliciting public preferences for healthcare: A systematic review of techniques. *Health Technology Assessment (Winchester, England)*, 5(5), 1-186.
- Salkeld, G., Ryan, M., & Short, L. (2000). The veil of experience: Do consumers prefer what they know best? *Health Economics*, 9(3), 267-270.
- Sawtooth Software Inc. The CBC latent class technical paper (version 3), 1-20. Retrieved November 16, 2007 from <http://www.sawtoothsoftware.com/education/techpap.shtml>
- Segrin, C. (2000). Social skills deficits associated with depression. *Clinical psychology review*, 20(3), 379-403.
- Spindler, H., & Pedersen, S. S. (2005). Posttraumatic stress disorder in the wake of heart disease: Prevalence, risk factors, and future research directions. *Psychosomatic Medicine*, 67(5), 715-723.
- Spoth, R., & Redmond, C. (1993). Identifying program preferences through conjoint analysis: Illustrative results from a parent sample. *American Journal of Health Promotion*, 8(2), 124-133.
- Teasdale, J. D., Scott, J., Moore, R. G., Hayhurst, H., Pope, M., & Paykel, E. S. (2001). How does cognitive therapy prevent relapse in residual depression? Evidence from a controlled trial. *Journal of Consulting and Clinical Psychology*, 69(3), 347-357.
- Vermeire, E., Hearnshaw, H., Van Royen, P., & Denekens, J. (2001). Patient adherence to treatment: Three decades of research. A comprehensive review. *Journal of Clinical Pharmacy and Therapeutics*, 26(5), 331-342.
- Watkins, P. C., Vache, K., Verney, S. P., Muller, S., & Mathews, A. (1996). Unconscious mood-congruent memory bias in depression. *Journal of Abnormal Psychology*, 105(1), 34-41.
- Wigton, R. S., Hoellerich, V. L., & Patil, K. D. (1986). How physicians use clinical information in diagnosing pulmonary embolism: An application of conjoint analysis. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 6(1), 2-11.
- Wills, C. E., & Holmes-Rovner, M. (2006). Integrating decision making and mental health interventions research: Research directions. *Clinical Psychology (New York)*, 13(1), 9-25.

DETERMINING PRODUCT LINE PRICING BY COMBINING CHOICE BASED CONJOINT AND AUTOMATED OPTIMIZATION ALGORITHMS: A CASE EXAMPLE

*MICHAEL G. MULHERN
MULHERN CONSULTING*

ABSTRACT

Within the context of maximizing revenue for a two-product line, this case study explores reliability or consistency in two ways: 1) between revenue maximizing prices indicated by the multinomial logit hierarchical Bayes model and optimization algorithms and 2) across automated search algorithms within the optimization software. In both instances, consistency was quite high. Specifically, the revenue maximizing prices indicated by the search algorithms replicated those suggested by simply viewing the shape of the price demand curve as derived from the choice based conjoint model. Further, four of the five search algorithms contained in the Sawtooth Software Advanced Simulation Module generated identical revenue maximizing prices. This case study also explored whether the pricing strategy should be altered given competitive response to the pricing changes. The recommendation was that pricing remain the same. The paper also reports the management decisions implemented based on the research and marketplace reaction to them.

INTRODUCTION

Ratings and choice based conjoint analysis have become popular tools among survey researchers to gain insight into a variety of marketing problems, most frequently product design and pricing. Most pricing studies focus on the best way to price a product or service by simulating various scenarios. In this paper, the focus is on optimizing price for a two product line using automated search algorithms. We also explore the search algorithms' consistency with the pricing implied by the choice model as well as the consistency across the search algorithms themselves.

The paper is structured in the following manner. First, the management issue, study background, and survey method are presented. Then, a set of optimization issues are raised and the analytical results supporting or refuting them are discussed. Third, the management decisions generated by the research are offered and their market impacts discussed. Lastly, general conclusions are outlined.

MANAGEMENT ISSUE

Management was about to introduce feature enhancements to a two-product line. The enhancements and the competitive set were already established so the task was to determine the set of prices that would optimize revenue. The client firm's product line consisted of a mid range and high end product. The mid range product's price range was \$6,995 to \$14,995 and the high end product's price range was \$14,995 to \$24,995. Low end products (\$1,995 to \$6,995) also

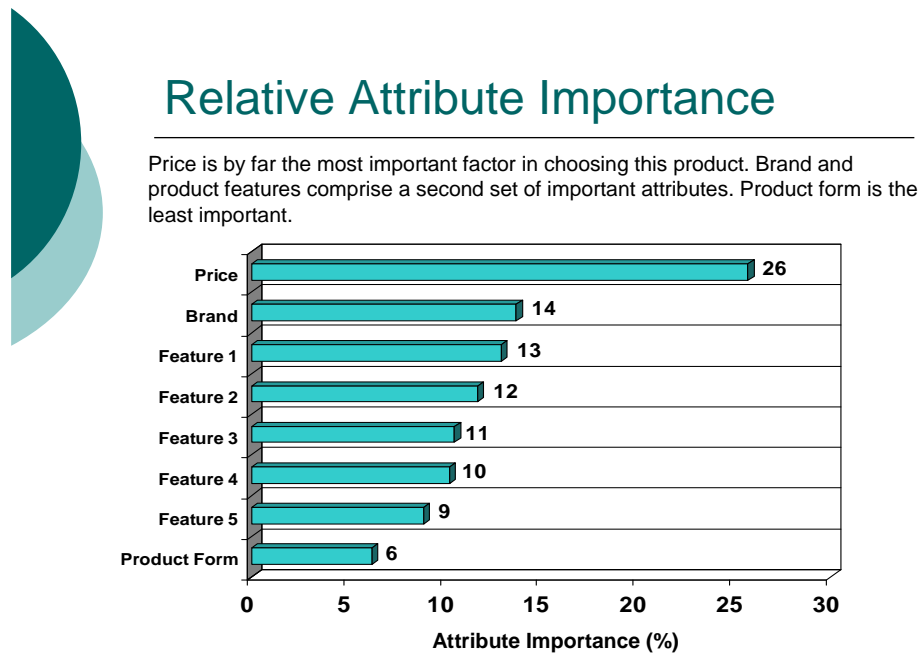
existed in this product market space. However, the client firm did not participate in this portion of the market. Revenue maximization rather than profit maximization was the objective as management was not comfortable assigning costs to each attribute level.

BACKGROUND AND METHOD

The products of interest were computer hardware and software trouble shooting systems used by engineers and technicians. The choice based conjoint design consisted of eight attributes and twenty three levels. The attributes included brand, price, product form, and five features. The price attribute encompassed the entire competitive range; from \$1,995 to \$24,995. Data from 644 respondents were collected via an online survey. The multinomial logit (MNL), Hierarchical Bayes (HB) model proved to be the most predictive to a holdout task. Mean absolute error was 5%. In terms of respondent reliability, 70% of respondents chose the same alternative in a replicated choice task.

As noted in Figure 1, price was the most important attribute in the study and product form was the least important. (Given the wide range of prices studied, it's not surprising that price was most important, as the importance is computed simply by comparing the utility of the extreme levels for each attribute included in the study.)

Figure 1

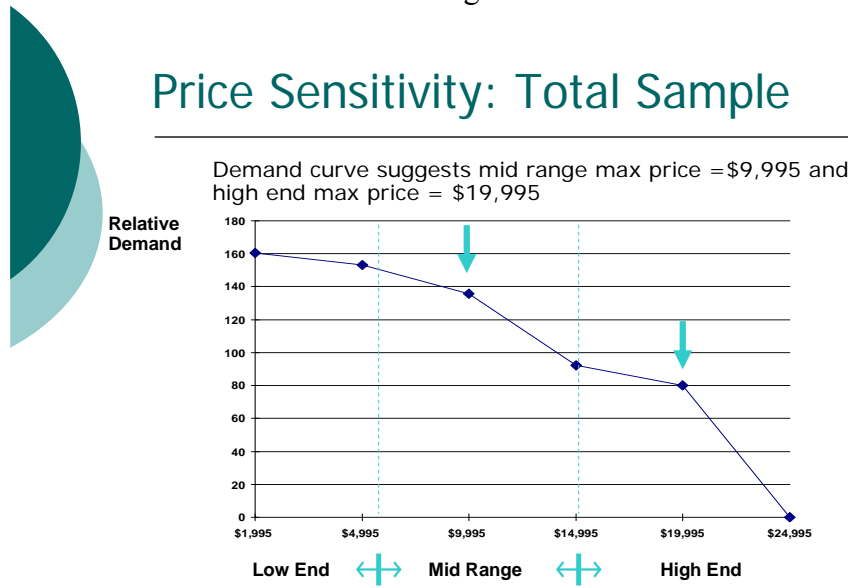


5

Figure 2 indicates that the slope of the demand curve suggests the greatest price sensitivity exists between \$9,995 and \$14,995 and over \$19,995. Consequently, the shape of the demand curve indicates that the mid range product's maximum price should be \$9,995 and the high end product's \$19,995.

Figure 2

Price Sensitivity: Total Sample



7

OPTIMIZATION QUESTIONS

The optimization questions investigated were threefold:

- Are the revenue optimizing prices consistent with the overall demand curve?
- Given the existing competitive set, do the maximum revenue optimizing prices vary by algorithm?
- Given changes in the competitive set, what are the revenue, share, and pricing implications for the client? In other words, what is the impact of competitor driven market changes?

METHOD

Sawtooth Software's Advanced Simulation Module (ASM) was used to investigate these questions. ASM consists of a set of five automated search routines that can be used to optimize a variety of functions, including revenue. Each of these algorithms was run on a base case with the most predictive choice based conjoint (i.e. MNL/HB) model. For those algorithms susceptible to local optima, the mean for five replications is reported. For the exhaustive algorithm, the global optimum is reported. Price was constrained and allowed to vary incrementally within the range of prices included in the study. That is, not only the measured levels were used in the optimizations, but the full range of price values was investigated.

RESULTS

Consistency – Across optimization algorithms and with the choice model

As noted in Table 1, the revenue maximizing price for the high end product converged at \$19,995 and was consistent across all five algorithms. At \$9,995, the mid range product price was similar for four of the five algorithms, with the gradient routine suggesting a considerably lower price. Further investigation of the five replications used to generate the gradient algorithm mean indicated that the two replicates that generated the highest revenue identified prices very close to those identified by the other algorithms (i.e. \$10,035 and \$9,612). Two of the replications found substantially lower mid range product prices, thus reducing the mean considerably. This suggests that the gradient algorithm found local optima in several of the replications.

Table 1

Algorithm	Revenue Maximizing Price for <i>High End</i> Product	Revenue Maximizing Price for <i>Mid Range</i> Product
Grid	\$19,995	\$9,995
Gradient	\$19,995	\$8,209
Stochastic	\$19,995	\$9,995
Genetic	\$19,995	\$9,995
Exhaustive	\$19,995	\$9,995

In general, the convergence among algorithms confirmed the conclusions one would make simply by the two “elbows” seen in the derived demand curves - price the high end product at \$19,995 and the mid range product at \$9,995. However, not all choice-based conjoint studies lead to apparent “elbows” in the demand function, and thus the question of optimization may not be so easily resolved by a visual inspection of the price part-worth utilities.

Impact of Competitive Marketing Responses

The next optimization issue addressed was the impact of competitive changes on the revenue maximizing prices. Using ASM’s Exhaustive method, five competitive marketing responses were investigated:

- Potential Competitive Actions
 - Price reduction by market share leader - 10, 20, 30%

- Actual Competitive Actions
 - Line extensions (improved functionality and higher prices by newer, lower share entrants)
 - Major leap
 - Major increase in functionality accompanied by a major increase in price (Low end vendor adds mid range product)
 - Minor enhancement
 - Improved functionality of one level of one attribute accompanied by a modest increase in price

Table 2 indicates that the revenue maximizing price for the high end product should remain at \$19,995. In four of the five competitive situations, the revenue maximizing price for the mid range product should remain at \$9,995. However, one anomaly occurred. When the leading competitor dropped their price by 30%, the exhaustive optimization algorithm suggested a considerably higher price of \$14,995 for the mid range product. However, if this finding were implemented, both revenue and share would suffer.

Table 2

		Base Case	Market Leader Price Decrease = 10%	Market Leader Price Decrease = 20%	Market Leader Price Decrease = 30%	Major Line Extension	Minor Line Extension
Optimal Price	Hi End	19995	19995	19995	19995	19995	19995
	Mid Range	9995	9995	9995	14995	9995	9995
		Index	Index (Relative to Base)	Index (Relative to Base)	Index (Relative to Base)	Index (Relative to Base)	Index (Relative to Base)
Max Revenue	Hi End	1.00	1.00	0.99	1.00	0.99	0.97
	Mid Range	1.00	0.99	0.97	0.89	0.98	0.97
	Total	1.00	1.00	0.99	0.98	0.99	0.97
Market Share	Hi End	1.00	1.00	0.99	1.00	0.99	0.97
	Mid Range	1.00	0.99	0.97	0.59	0.98	0.97
	Total	1.00	1.00	0.99	0.89	0.99	0.97

In an attempt to gain more insight into this anomaly, a separate profit (rather than revenue) optimization was conducted. Synthetic costs were generated with lower costs matched with lower attribute functionality and higher costs matched with higher attribute level functionality. The counterintuitive result remained.

Although one anomaly remained, the general recommendation to management was that the revenue maximizing prices should not be affected by the competitive market responses modeled.

WHAT REALLY HAPPENED – SIX MONTHS LATER

In this section, three phenomena will be discussed:

- Product/Price Introductions
- Changing Competitive Landscape
- Dialogue within the Corporate Hierarchy

PRODUCT/PRICE INTRODUCTIONS

Management introduced three products; two mid range products – one at \$9,995, another at \$12,995 and a high end product at \$19,995. The high end product exceeded its forecast while both mid range products floundered. The high end product was an easy sell and the mid range products, especially the one priced at \$12,995, were a difficult sell. The end user sales force took the path of least resistance and focused their efforts on the high end product. The distributor channel, which only carried the mid range products, essentially abandoned their sales efforts by allocating their time and resources elsewhere since the commission on these difficult to sell mid range products was not worth their effort.

Several lessons were learned from this experience:

- Ignore the optimization and simulation results at your peril
- Variables not modeled (e.g. sales force and distributor perceptions) in the conjoint exercise can have a major impact on marketplace response
- Make it easier for the sales team to “get a foot in the door” by clearly differentiating “base” from “optional” features

The future will determine if demand has shifted temporarily or permanently to the fully featured high end product.

CHANGING COMPETITIVE LANDSCAPE

The competitor that launched a minor extension was perceived as a low share, niche player with favorable word of mouth who developed quality products sold via a single channel of distribution. Shortly after the study was completed, they received an infusion of venture capital that allowed them to make acquisitions to broaden their product line, increase their promotional budget and enhance their trade show presence, as well as enter additional distribution channels.

Several lessons were learned from this experience:

- Ongoing competitive changes can be modeled
- Brand equity/value measures can be short lived

It remains to be seen if this competitor will become a major competitor across the product spectrum or continue to be a niche player.

DIALOGUE WITHIN THE CORPORATE HIERARCHY

During the planning and budgeting process, senior management exerted considerable pressure on product/marketing management to increase prices and, therefore, revenue. Product/marketing management refused, citing the simulations and optimizations. A middle ground was reached when product/marketing management offered to consider repricing product extensions (i.e. options) rather than the base product. The lesson is that our research can be used in unexpected ways.

CONCLUSIONS

From the researcher's perspective, we learned

- Automated search routines can be used to confirm or refute the recommendation suggested by a simple view of the shape of the demand curve estimated via MNL/HB
- Multiple search routines allow the analyst to test the reliability of the optimization findings. This can be beneficial if the analyst has a large data set and/or time pressures preclude using exhaustive search
- The simulator can be used in conjunction with automated search to refine pricing strategy
- Watch out for local optima with certain automated search routines

From the manager's perspective, we learned

- Convergence is encouraging (especially for high end product)
 - Between the price points suggested by the shape of the price utility function and the optimizations
 - Across the search routines within the optimization software
- In this data set, competitive actions do not impact the revenue maximizing prices
- Optimizations consider neither intangibles nor unmeasured marketing variables (e.g. breadth of distribution, size of sales force). However, the analyst may decide to explore these elements in the market simulator.

LIMITATIONS AND FUTURE RESEARCH

The most obvious limitation of this paper is that only a single database was analyzed. Further, the revenue maximizing prices may be overstated since there was no budget constraint imposed. More research is clearly needed to determine the impact of imposing a budget constraint during the optimizations.

REFERENCES

Bakken, David G. (2003), Using Genetic Algorithms in Marketing Research, Sawtooth Software Conference Proceedings, 319-330.

Sawtooth Software, Inc (2003), Advanced Simulation Module for Product Optimization v1.5. Technical Paper.

USING CONSTANT SUM QUESTIONS TO FORECAST SALES OF NEW FREQUENTLY PURCHASED PRODUCTS

*GREG ROGERS
PROCTER & GAMBLE*

INTRODUCTION

To predict purchase incidence across time, forecasters of consumer goods sales have used the Negative Binomial Distribution (NBD). For instance Ehrenberg (1972), demonstrated the effectiveness of NBD as a tool to estimate sales in various industries. The NBD is a stochastic model that combines Poisson and Gamma probability distributions in order to account for heterogeneity across households in purchase timing (Poisson) and frequency of purchase (Gamma). The NBD model is used in most commercial applications of consumer goods forecasting at companies such as BASES (a division of the AC Nielsen company).

Inputs of consumer appeal may come from various sources, the most common of which is stated purchase intent from a 5-point scale as asked in a concept test. Purchase intent must be calibrated to empirical data, as there are known differences between claimed and actual purchase behaviour. The calibration is crucial to creating an accurate forecast, but it is often very difficult to do effectively as there are known differences in scale usage across countries and/or categories. This makes it difficult to commence forecasting in a new country or category as the calibration of purchase intent to in-market trial is often weak. With a sufficient number of calibration points, this method can be effective as demonstrated by Jamieson & Bass (1989). Others have used preference share measures from conjoint data as input to an NBD-based forecast model. For example, Zufryden (1997), has done work in this area. These models have the ability to put more context on the buying decision in the consumer research, but the relatively limited commercial use of this approach to date means there are fewer validations (comparing predicted to actual sales volume) than the claimed purchase intent method. As conjoint approaches like choice based conjoint (CBC) continue to grow in popularity, there is reason to believe that the input of consumer appeal will move away from purchase intent to preference share.

CBC studies for fast moving consumer goods can cost well over \$100,000 US, with the cost depending on such design elements as number of items in the study, number of respondents in the study, level of realism in the choice task, number of retail channels modelled, etc. The custom nature of these studies also requires more time to design and analyse than many other research efforts (e.g., concept test, equity tracking, etc.). Given the current cost and complexity of choice models based on CBC, researchers are beginning to look at alternative ways to create choice models. For example, the research industry has looked for efficiencies by developing software that handles questionnaire design, creates a respondent interface for the questionnaire, and offers standard methods of analysis (i.e., suite of tools from Sawtooth Software).

An alternative approach to CBC explored in this research makes use of a constant sum question (also referred to as 'chip allocation'). Constant sum is a type of comparative rating scale in which the respondent in a survey is instructed to divide a given sum among two or more items based on some criterion (e.g., their likelihood of purchasing). This is usually expressed as

a single question of the form “...thinking of the next 10 purchases you will make of <<insert category name>>, how many will be of each of the following products”? The respondent is then presented with a list of products in order to allocate their next 10 purchases. Market researchers are interested in being able to use share of preference (SOP) from a constant sum question since it is much faster and cheaper to execute than a CBC study (especially in countries or categories without many historical concept test results). In this research, the SOP from a single constant sum question is compared to the SOP from CBC. An alternative method to estimate base trial using the Dirichlet model is suggested, and an assessment of the most appropriate way to ask the constant sum question is explored.

THE CONCEPT OF BASE TRIAL

There are many approaches of new product forecasting using SOP, but they typically make direct adjustments on the SOP for awareness and availability. These models do not build volume ‘from the bottom up’; they make macro adjustments to the SOP to create a ‘long run’ estimate of volume. Clearly, this approach to forecasting has some limitations, particularly in being able to decompose volume into trial, repeat, repeats per repeater, etc. which is very beneficial when creating marketing programmes and tracking initiatives.

A common approach to estimate the volume for a new product in its first year is to estimate the trial and repeat volume separately (Fourt and Woodlock 1960). In order to estimate trial many forecasting models use the concept of ‘base trial’ to describe the trial in an environment of 100% product distribution (anyone who wants it can find it), 100% awareness (all potential buyers are aware of product), and the absence of any marketing actions that could temporarily change consumer interest (i.e., trade promotions or sampling). The base trial is estimated using data from consumer research, such as claimed purchase intent or preference share.

Many forecasting systems use the negative binomial distribution (NBD) to account for purchase timing, and in that way estimate how base trial builds over time. The NBD approach is well suited to this and has been documented extensively for use in this situation. While NBD may be helpful, it only estimates the build over time; it does not estimate the base trial at the end of year one on its own—that is an input to the model. The NBD requires inputs of buying behaviour for the new item—penetration, purchase frequency, a constant (k), and the time when the penetration and purchase frequency will be reached. The base trial at time t can be estimated using the following method (Anscombe 1950):

$$(1) \quad BaseTrial_t = 1 - \left(1 + \frac{m_t}{k}\right)^{-k}$$

Where,

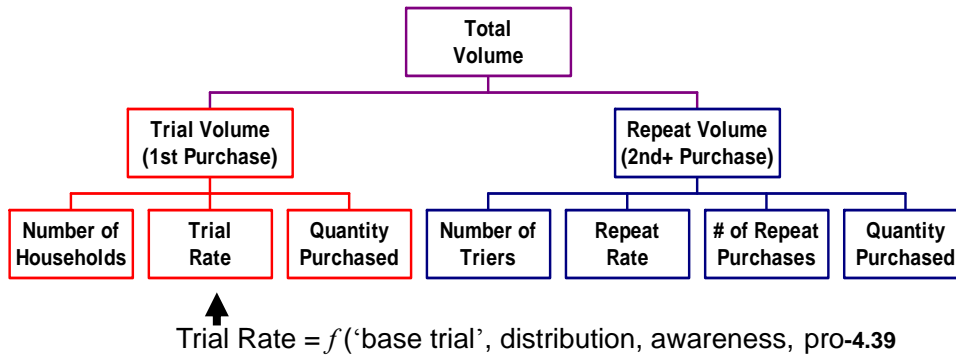
$Base\ trial_t$ = trial at time t unadjusted for awareness, distribution, promotion, seasonality, etc.

m_t = penetration/purchase frequency (at time t)

k = the shape parameter and may be solved for through iteration.

As base trial is calculated by period (often by week for fast moving consumer goods) over the course of a year, it can be adjusted according to the time dependent exogenous marketing variables like distribution, awareness (derived primarily from media plans) and trade promotions. This leads to the final estimate of trial, one of the key elements in the forecast.

Figure 1



USING THE DIRICHLET MODEL TO ESTIMATE BASE TRIAL

For all of the attractive properties of SOP, on its own it does not provide the necessary base trial estimate, and therefore cannot be used within traditional NBD based forecast models. A way of breaking down share into its purchase behaviour components is required. The Dirichlet model can do this efficiently, and requires inputs that are readily sourced: *category penetration*, *category purchase frequency*, *item share*, and the *variance of the beta distribution* (sometimes referred to as the *switching constant S* – notation from Ehrenberg). The switching constant S is a category parameter related to the degree of homogeneity of item selection—it can be estimated using household panel data, or directly from SOP data.

There are many ways to achieve a given market share. For instance, a 10% share for an item can be realised from 100% of the population buying the item 10% of the time (assuming pack sizes are equal across items and purchase frequency is equal across buyers) or 10% of the population buying the item 100% of the time—or, more likely, somewhere in between. This example demonstrates the usefulness of the Dirichlet model as it estimates the balance of penetration (often referred to as trial for a new item) and purchase frequency (a measure of loyalty) that lead to a specified share.

The switching constant S is a key input to the Dirichlet model that relates to the degree of switching among items in the category. It is inversely related to the degree of variation in individual share levels across the population as shown in (2).

$$(2) \quad \text{Variance}_{n,i} = \text{SOP}_i(1 - \text{SOP}_i) \left(\frac{1}{1 + S} \right)$$

Rearranging,

$$(3) \quad S = \frac{\text{SOP}_i(1 - \text{SOP}_i)}{\text{Variance}_{n,i}} - 1$$

Where,

SOP_i = Share of Preference

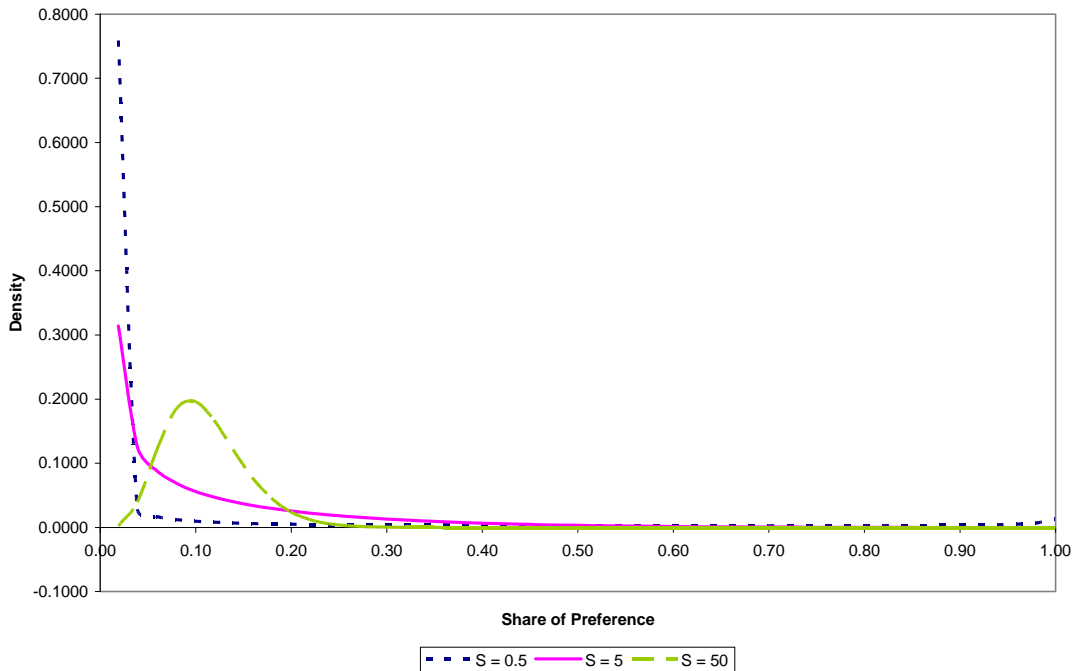
$Variance_{n,i}$ = Variance of SOP across n respondents for item i

S = Dirichlet switching parameter

S can take any positive value from zero to infinity. It is calculated for each item and then averaged across the items—this results in a single S value per category. When S is small (e.g. 0.1 to 0.5), when the item share varies a lot across the population, with numbers of individuals having share close to zero and others having share close to 100%. Conversely, when S is large (e.g. >10), the item share across the population varies much less, with most individuals having share close to the mean share. Figure 2 shows how the Dirichlet model estimates the share across the population for various values of S , and a mean share of 10%.

Figure 2

Beta Density of Preference Share for Various S Values



COMPARING SOP ACROSS RESPONDENTS (CBC VS. CONSTANT SUM)

The constant sum question may be asked in many ways and the context may vary, but the most common form of the question relates to the ‘next 10 purchases’. In the case of a new product introduction, a respondent views a concept board for the new product and immediately following that exposure is asked to indicate their next 10 purchases in the category. The respondent has 10 ‘chips’ (either figuratively or literally in some cases) and may allocate them in any fashion they want as long as they use all 10 ‘chips’—hence the name ‘constant sum’.

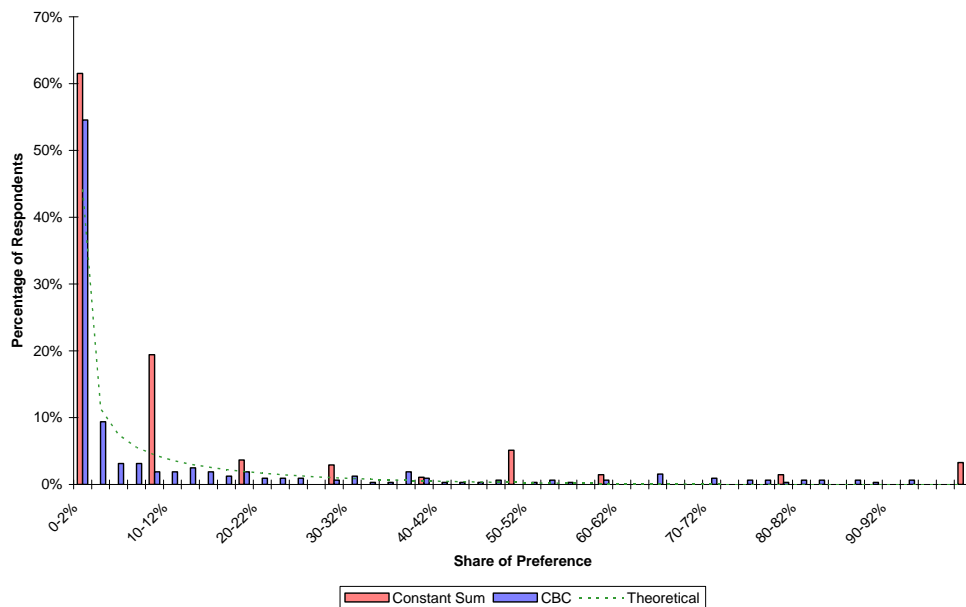
To illustrate the differences between CBC and constant sum results, the results from a new product introduction in the Laundry Detergent category in the United States are examined. Figure 3 shows the distribution of SOP across respondents for constant sum, CBC, and a theoretical benchmark. The constant sum and CBC data are plotted as a histogram and the theoretical benchmark is a continuous function. The theoretical benchmark is the plot of the Dirichlet estimate of SOP across respondents given the category parameters (penetration and purchase frequency), using the same SOP and S value from the CBC data.

Figure 3 shows the CBC frequencies of SOP across respondents reasonably matching the theoretical estimate from the Dirichlet. There is more divergence between the SOP collected through CBC vs. a simple constant sum question. Notably, the constant sum question has more respondents selecting the item in general (higher average SOP), and has more respondents selecting it exclusively than the CBC (SOP=100%). This may occur simply because it is a single question vs. the multiple choice tasks a respondent answered in the CBC studies. In the CBC study, a respondent would have to select the test item in every choice task to show a 100% share of preference, whereas in the constant sum question the respondent achieves 100% share of preference for the item by putting all of their chips against that item. This may be done on purpose, but it also could be due to the simplification of their response so that they expedite completion of the questionnaire.

Most respondents will allocate zero, 1, 5 or 10 ‘chips’ to the new product in a constant sum question. This is likely due to a simplification strategy on the part of the respondent—it would be more precise to indicate selecting an item in 3 out of their next 10 purchases, but that precision does not fit with the fuzziness of the task in their mind. Allocating the next ten purchases across an array of brands is challenging for a respondent, and so a simplification strategy that provides a ‘rough estimate’ is evoked.

Figure 3

Share of Preference Across Repondents



The share of preference for the CBC and the constant sum question shows a higher average SOP coming from the constant sum question. The Dirichlet S parameter can be estimated from the same dataset by calculating the standard deviation of shares across respondents. The constant sum study also has a lower value of the S parameter than the CBC study. This traces directly to the fact that there is greater heterogeneity, or ‘polarisation’, across respondents in the constant sum results.

The higher SOP and lower S from the constant sum studies leads to a similar estimate of base trial as that coming from the CBC studies (that have a lower SOP but higher S). This is a very interesting, and useful result. The Dirichlet model is essentially attempting to estimate the buying behaviour that would result in a given share. A lower value of S indicates relatively more heterogeneity in brand/item preference in the category. The greater the heterogeneity from the beta-binomial distribution with a relatively low S means that some people do not buy at all, yet some buy almost exclusively. That situation leads to a low penetration and high purchase frequency. Conversely, a relatively high S parameter will result in more people selecting the item (higher penetration) but they will not buy it as often on average.

Using the Dirichlet model to estimate the base trial given the respective SOP and S parameters (along with inputs of category penetration and purchase frequency), shows the two methods yield similar results. While SOP and subsequent base trial estimates from a constant sum question are higher than that from CBC, using the Dirichlet model to estimate base trial brings them much closer together than the respective SOP’s on their own. Table 1 shows a pattern emerging across studies and categories that indicates some consistency between the CBC and constant sum estimates of base trial. Although the base trial estimates are similar, it appears that constant sum derived base trial estimates are consistently greater than CBC derived base trial estimates.

Table 1

	Case 1		Case 2		Case 3		Case 4	
	CBC	CS	CBC	CS	CBC	CS	CBC	CS
Sample Size								
Share of Preference (SOP)	8.2%	11.9%	9.1%	16.1%	29%	46%	35%	48%
Calculated S Parameter	4.1	2.2	1.2	0.5	2.6	0.8	1.8	0.8
Base Trial Estimate	28.3%	32.3%	21.0%	26.2%	62.2%	66.3%	57.0%	61.1%
Base Trial Index (CS/CBC)		114		124		107		107

*CS denotes ‘constant sum’

All cases are new product introductions in the laundry detergent category in the United States, except for case 4, which is a new product introduction in the dentifrice category in the United States. In each case, the constant sum study was fielded within 3 months post the CBC study.

EFFECT OF CONTEXT ON THE CONSTANT SUM RESPONSE

The chip allocation question is a simple question, but can be asked in many different ways. Like any survey question, the context in which the question is asked can affect the response. Is there an optimal way to ask the question? Data was collected across four concepts, across three different consumer goods categories to explore the sensitivity of SOP estimates from a constant sum question when the question was asked in various ways (all data collected in the United States). Six different executions of the constant sum were explored as summarised in Table 2.

In cell 1 the constant sum question was added in the simplest manner possible to a concept test questionnaire. The respondent was shown a concept for the test product and then asked the constant sum question after the purchase intent question. The item list for allocation of the “next 10 purchases” was at the SKU level with no prices shown next to the SKU description. In all cells the item list was shown in alphabetical order. The only change from cell 1 to cell 2 is that average shelf prices were shown by each SKU in the item list in cell 2. Cell 3 was the same as cell 2 except the test concept was shown amongst a competitive set prior to asking the purchase intent and constant sum questions. Cell 4 was the same as cell 3 except a picture of the category shelf was shown in addition to the written item list. Cell 5 was the same as cell 2 except a two stage item was shown to the respondent. The respondent was first shown an item list at the brand level where they would allocate their “next 10 purchases”. If they selected the test brand, then they were then asked to allocate another “next 10 purchases” at the SKU level for that brand. This two stage item list was tested to see if simplifying the list initially helped respondents behave in a more typical buying mindset. Finally, cell 6 was the same as cell 2 except the constant sum question was asked after the purchase intent question.

Table 2

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6
Price	No prices shown	Average price	Average price	Average price	Average price	Average price
Concept	Test brand only	Test brand only	Comp. context	Comp. context	Test brand only	Test brand only
Picture/List	List	List	List	Shelf picture	List	List
Brand/SKU level	SKU	SKU	SKU	SKU	Brand>SKU	SKU
Question location	After PI	After PI	After PI	After PI	After PI	Before PI

	Base Trial					
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6
Concept 1	14.2%	13.0%	14.4%	13.9%	14.1%	12.1%
Concept 2	22.8%	21.0%	23.1%	22.4%	22.5%	19.6%
Concept 3	7.7%	11.1%	na	na	na	7.5%
Concept 4	26.8%	24.6%	na	na	na	21.0%

As the results in Table 2 show, the manner in which the constant sum question is asked has a limited impact on the results. While there is a directional difference for SOP and the consequent base trial as they are lower if the constant sum question is asked before the purchase intent question, none of the cells is significantly different from any other (at the 95% confidence level).

CONCLUSION

This research has shown the promise of using a relatively simple constant sum question for forecasting trial of new consumer products. The use of the Dirichlet model has been applied to yield an estimate of base trial, which can then be used in a typical NBD model to build a forecast of sales across time. The use of the Dirichlet and NBD models in this way is no different than others have done previously (e.g., Zufryden 1988). What is different about this research is the use of the constant sum question to generate the share of preference. The way in which the respondents answer a constant sum question leads to a greater variance in share of preference across respondents, and this may well be due to simplification strategies on the part of the respondents. Further research to understand why respondents respond they way they do could lead to improvements in how the constant sum question is asked. The examination of context effects on the response to a constant sum question showed minimal effect on the results in this research, yet this was hardly an exhaustive study of context effects—there is an opportunity for further research in this area.

This research is focused on the practical usage of choice models as it tries to address current barriers of cost and complexity of CBC models. Practitioners should not be led to believe that constant sum questions could provide the multitude of scenarios of CBC models (i.e., share of preference for a test item at varying price levels). However, it is clear that a simple constant sum question can provide the necessary data to create a sales forecast—at a fraction of the cost of CBC models, and without the burden of databases to calibrate results as with purchase intent measures.

REFERENCES

- Ehrenberg, A. S. C., (1972), *Repeat Buying—Theory and Applications*, North-Holland, New York.
- Fourt, L., Woodlock, J., (1960), “Early Prediction of Market Success for New Grocery Products,” *Journal of Marketing*, Vol. 25, pp. 31-38.
- Jamieson, L., Bass, F., (1989), “Adjusting Stated Intention Measures to Predict Trial Purchases,” *Journal of Marketing Research*, Vol. 26, Iss. 3, pp. 336-345.
- Zufryden, F. S., (1988), “Using Conjoint Analysis to Predict Trial and Repeat-Purchase Patterns of New Frequently Purchased Products,” *Decision Sciences*, Vol. 19, Iss. 1, pp. 55-71.

REPLACEMENT MODELING: A SIMPLE SOLUTION TO THE CHALLENGE OF MEASURING ADDING AND SWITCHING IN A POLYTHERAPY CHOICE ALLOCATION MODEL

LARRY GOLDBERGER
ADELPHI RESEARCH BY DESIGN

INTRODUCTION: THE PROBLEM

The choice allocation model (Louviere & Woodworth, 1983) is the preferred approach for measuring preference share in the pharmaceutical industry. Since physicians make prescription choices on a daily basis for a diverse set of patients with a given condition, this approach better mirrors the treatment choices they are likely to make than a single choice model. Viewed another way: respondents are asked to make 100 choices for a typical set of 100 patients and the results may be modeled as a standard choice model. The advantage of this approach over a single choice model is that it captures the use of niche products which might be preferred for a limited segment of the patient population.

A complicating factor in the use of this technique is the possibility that treatment choice allocations will exceed 100%. This is quite common in pharmaceutical research, as a combination of drugs, also known as polytherapy, are commonly prescribed to treat a patient suffering from a condition, such as diabetes, high blood pressure, or psychiatric disorders. In the treatment of diabetes, for example, new drugs may be either added-to or substituted-for current drugs as the patient's ability to produce insulin deteriorates over time. Since the average patient receives more than one drug, allocations now add to more than 100% of patients so that the allocation total is no longer fixed and both share and allocation total need to be solved to accurately reflect the choices that physicians make.

Of course, if the number of treatment combinations is limited, each combination can be treated as a separate choice. Often, however, the number of treatment choices is so extensive that this is not feasible. For example, with a new product and 12 drugs in the current market, there are 13 single-drug plus 78 two-drug combinations possible for a total of 91 choices.

The real problem is that, unlike simple choice situations, physicians do not simply substitute the new product for an existing product. When a new product is introduced physicians may choose not to use the new drug, to substitute the new drug for one or more of the current drugs, or simply to add the new drug to the current treatment regimen. A single respondent may make all these choices across their patient population, applying each of them to a percentage of his patients. Since the set of all possible treatment combinations is often too lengthy to measure in a survey, these choices are difficult to capture in a standard choice model as the average number of drugs per patient varies along with the percentage of the treatment choices represented by each drug.

The choices that physicians make when using a new product can affect the allocation total in different ways:

- Adding the new product to the current regimen increases the allocation total
- Substituting the new product for one or more current treatments can either reduce the allocation total or keep it constant.

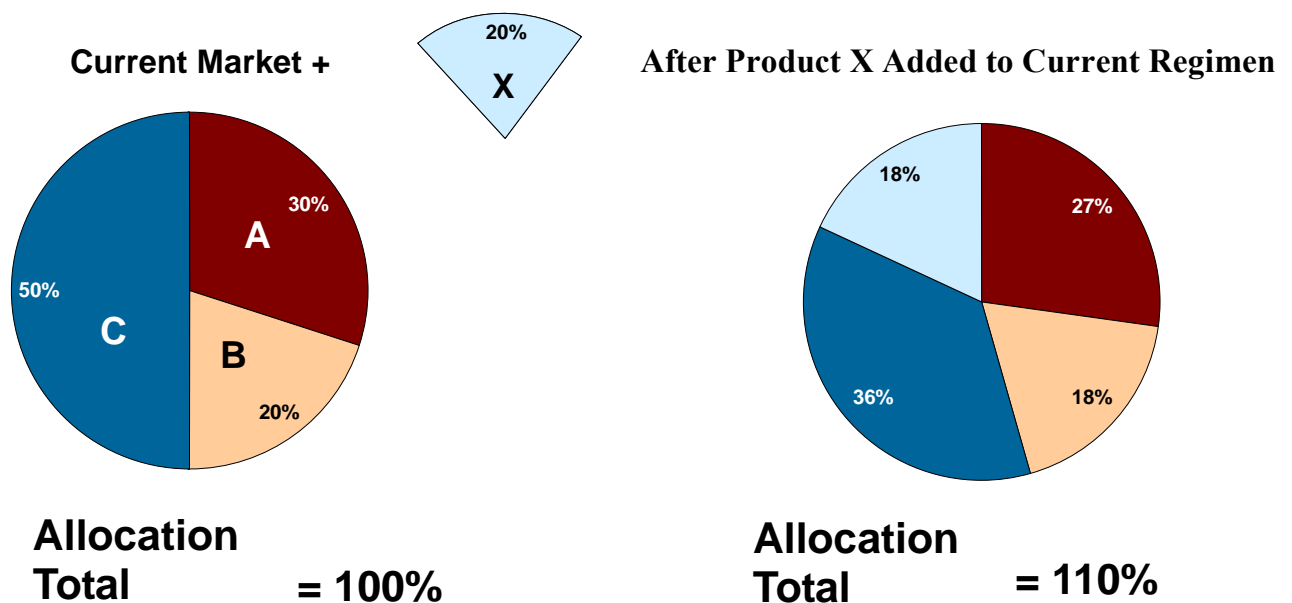
Typically, the effect falls somewhere in between addition and substitution as physicians who are likely to use the new drug have different treatment predilections, with some physicians tending to use the new drug as an add-on treatment and others tending to substitute it for an existing drug. A further complication is that the tendency to add or substitute the new drug may depend on the specific characteristics of the drug in a given scenario.

In the following example, the current market consists of three products:

- Product A, capturing 30% of the market
- Product B, with 20% of the market, and
- Product C, with 50% of the market.

Assume that Product X is introduced with a configuration that physicians say they will use in 20% of their patients. In this example, Product X is added for those patients on Products A and B, while it is always substituted for Product C for those patients on that treatment. Assuming that the use of Product X is proportional across all patients, Product X replaces Product C in 10% of the patients, and is added to products A and B in the remaining 10%. Thus the allocation total increases by 10 percentage points (when used in patients on Products A and B).

Chart 1:
Current Regimen Compared to Regimen After Product X Enters the Market



In order to produce a model that accurately captures the total scripts that will be written across all patients, both the allocation total and the share of scripts needs to be modeled. The following table shows the resulting allocation total and re-percentage share.

Table 1:
Allocation Total and Re-percentage Share

Current Scenario: 100 patients, 3 drugs, 1 drug per patient					
Drug	A	B	C	Total	
Number of Drugs	30	20	50	100	
Number of Patients	30	20	50	100	
New Scenario (same 100 patients): Product X introduced.					
6 patients on A have X added, 4 patients on B have X added 10 patients on C have X substituted, .					
Number of Patients	24	16	40	10	100
Number of drugs	24	16	40	10	110
Drug distribution summary	A	B	C	X	
Allocation volume	30	20	40	20	110
Allocation share	27%	18%	36%	18%	100%

This situation creates two problems that are not easily handled with the standard choice modeling approach:

- In addition to solving for the share of each product, it is necessary to solve the total allocation if the model is to represent the share of patients on each drug.
- Since cannibalization differs sharply for each current treatment, with some products being substituted-for and others added-to, as the new products share changes it is necessary to capture the disproportional changes in the current treatments.

CURRENT APPROACHES TO DEALING WITH THE PROBLEM

The solutions that have been used in the past include:

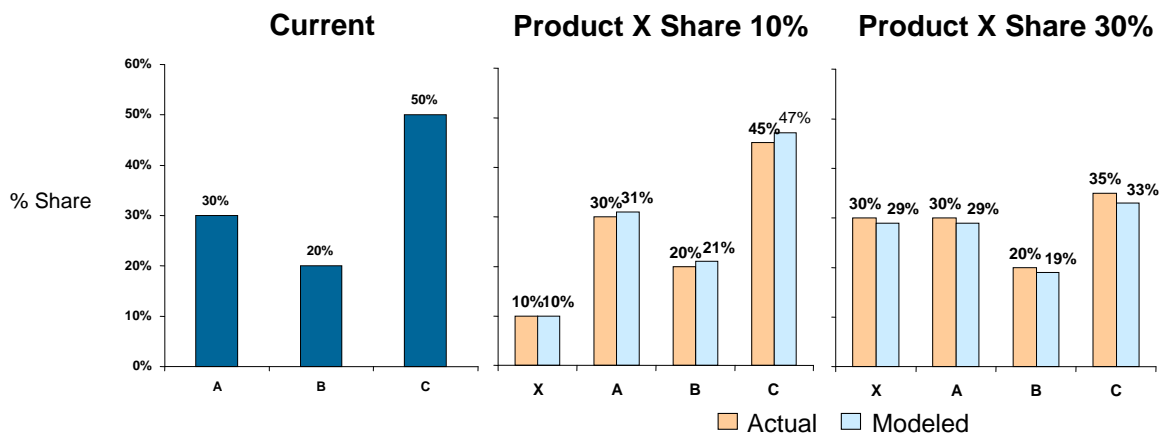
Average Allocation Total Approach

In this approach the model simply calculates the average size of the allocation total across all scenarios and applies that factor to shares predicted by the model. This approach clearly fails to distinguish variations in addition and substitution that are dependent on the characteristics of the new product, overestimating shares of all the products when substitution is more dominant and underestimating shares when addition is more dominant.

Furthermore, this approach does not deal with the problem of differential cannibalization as the shares of all the current products are simply reduced proportionately to accommodate the new product (Allenby, et. al, 2005). Although the use of HB can ameliorate this problem to some extent, it does not solve it. That, for example, the choice of no treatment for a proportion of patients with a concomitant condition that precludes treatment. The no treatment option may be reduced only slightly by the introduction of a new product, and could be evident in all respondents.

Using the previous example, where the current treatment starts out as 100% and increases by half the total of the new product's share, the following chart illustrates that applying an average allocation total of 110% overestimates the share of all the products when the share of Product X is below average, and underestimates the shares when Product X is above average.

Chart 2:
Inaccuracies Inherent in the Average Allocation Approach



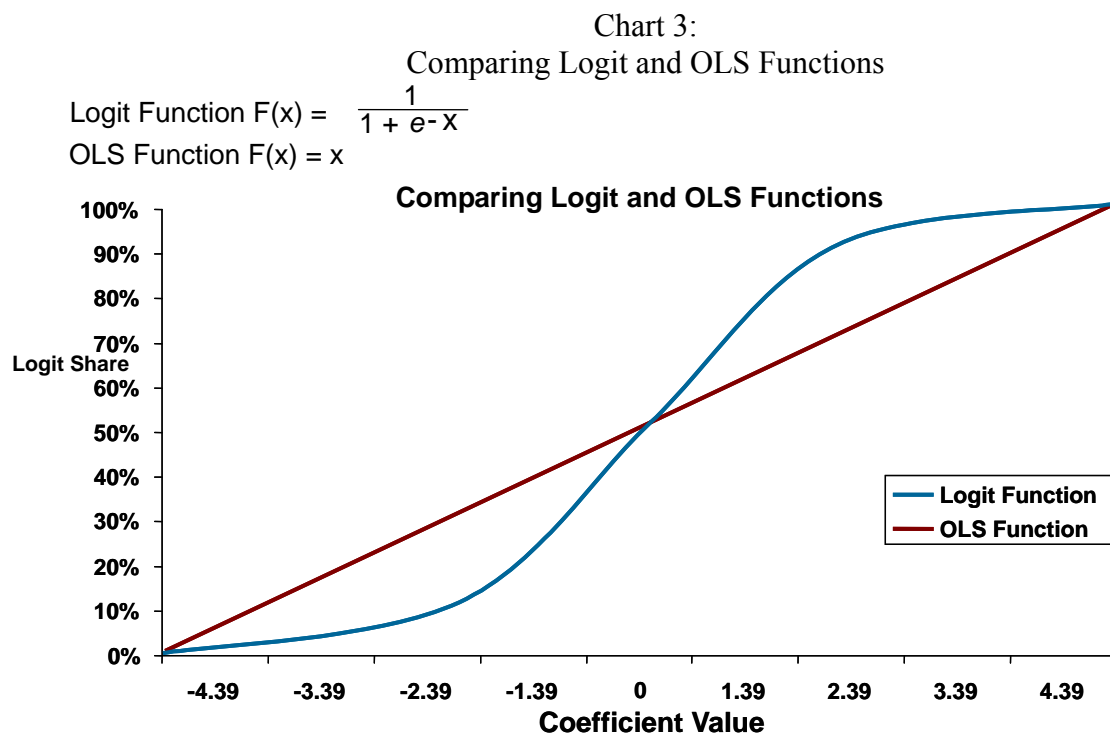
Allow for Expansion of the Allocation Total by Adding an Expansion Choice

Adding an expansion choice to the model is a slight improvement over the cruder *Average Allocation Total* approach. Here the expansion choice represents the difference between the maximum allocation total and the minimum allocation total, so that the share of this expansion choice is set to zero when the allocation total is greatest. As the shares of the new product increase, the share for the expansion slice decreases, thus allowing some expansion of the market as would be expected when a new product is added to some extent to the current therapy choices.

This allows only for a crude measure of addition and substitution as a whole, and does a poor job of measuring differential cannibalization of the current products.

Solve Separate Models for Share and Allocation Total

A different approach to the problem of measuring changes in both share and allocation total is to solve separate models: Using a standard choice model to solve for the percentage of the total represented by each choice and a simple OLS regression to model the allocation total. This approach allows for a better measure of the allocation total but fails to identify specific products that are added-to or substituted-for and can result in obvious reversals when the two models are not coordinated. That is, due to the natures of the logit and OLS functions, the share for the new product may increase at a greater or lower rate than the increase in the allocation total, depending on the starting share for the new product prior to the change in a given characteristic.



Binomial Approach Solving Separate Models for Each Product

In frustration, modelers often abandon the multinomial approach entirely and resort to solving separate binomial choice models for each product. The binomial approach appears to be the best of these solutions in capturing this partial cannibalization effect as the share of patients treated for each drug can change as the new product attributes become more or less attractive. Thus, for a substituted drug, as the value of the new drug increases, the value of the alternative drug decreases. So as the percentage of patients treated with the new drug increases, the percentage of patients treated with the alternative decreases. Conversely, a drug which is added-to will be relatively insensitive to changes in attribute levels of the new drug and will not change as readily.

Unfortunately, while the binomial approach does a fairly good job of portraying the survey data in a model, it can result in illogical results or reversals. In practice, a model must not only

describe the data as closely as possible, but also avoid reversals that undermine the credibility of the model in the eyes of our clients. Similar to that which is seen in cross-effects models, the binomial approach is particularly prone to reversals when attribute levels are changed in the new product simulator, causing alternative choices to increase or decrease far more than would be possible based on the share change in the new product.

Strictly speaking, choice models are not intended to be change models and these illogical changes may just reflect error in the estimates, but our clients do not see it this way. For example, if a change in the attribute levels of the new product results in a 5-point share increase and an alternative drug declines 6-share points, then the credibility of the model will be undermined.

The reason this may occur is due to the nature of the coefficients in a standard choice model which represent a change in the likelihood that a choice will be made, multiplying this change in likelihood by the current likelihood. Accordingly, the effect that a level change will have is dependent on the current likelihood, so when the share of the new product is low to begin with (based on the other attributes) a change in the alternative (necessarily low to begin with) will be disproportionately large relative to the change in the new product.

Chart 4:
Relationship of the Logit Function and Changes in Share

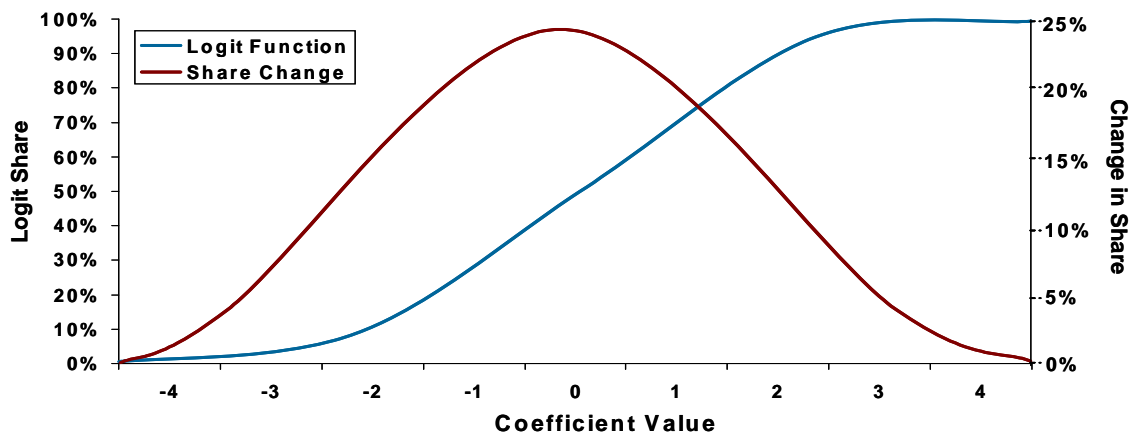


Table 2:
Relationship of Changes in Share to Changes in the Coefficient Value
with the Logit Function

Value X	Exp(X)	Change in Value	No choice Alternative EXP(0)	Sum	Share X	Absolute Point Change in Share	Percent Increase in Share	Percent increase in EXP(X)
-5	0.01		1	1.0	1%			
-4	0.02	1.00	1	1.0	2%	1%		
-3	0.05	1.00	1	1.0	5%	3%	164%	172%
-2	0.14	1.00	1	1.1	12%	7%	151%	172%
-1	0.37	1.00	1	1.4	27%	15%	126%	172%
0	1.00	1.00	1	2.0	50%	23%	86%	172%
1	2.72	1.00	1	3.7	73%	23%	46%	172%
2	7.39	1.00	1	8.4	88%	15%	20%	172%
3	20.09	1.00	1	21.1	95%	7%	8%	172%
4	54.60	1.00	1	55.6	98%	3%	3%	172%
5	148.41	1.00	1	149.4	99%	1%	1%	172%

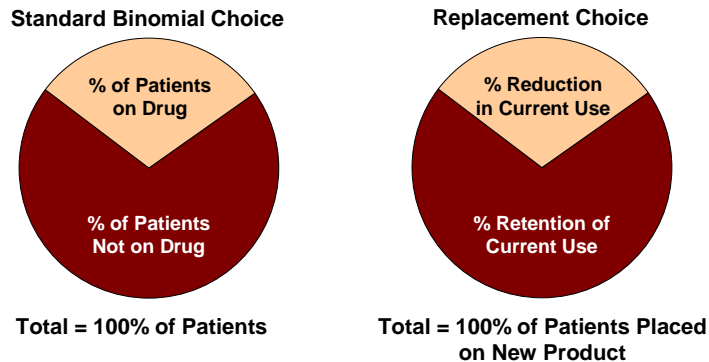
PROPOSED SOLUTION:

MODEL THE LIKELIHOOD THAT A CURRENT PRODUCT WILL BE REPLACED

The solution proposed in this paper is a simple one: model the changes in prescribing of the new drug as its characteristics change using binomial logit, while modeling the likelihood that a product will be substituted-for rather than added-to for the current products. This approach not only avoids reversals but more accurately mirrors the data in the choice allocation.

As illustrated below, the basic difference between the standard binomial and the replacement choice approaches is that the replacement approach models the likelihood that the current product will be replaced each time the new product is prescribed, rather than the likelihood that the current product will be prescribed for a given patient.

Chart 5:
Dependent Variable in the Standard Binomial and Replacement Approaches



In both approaches, the likelihood that the new product will be chosen for a patient is modeled using binomial logit:

Model new product using binomial logit

$$\text{Logit } P(X) = a + \sum B_i X_i$$

where

- $P(X)$ = Probability of prescribing Product X for a given patient
- a = Constant
- B_i = New product coefficients
- X_i = New product attribute levels
- $1 \leq P(X) \leq 0$

In estimating the shares for the current products, however, the replacement model estimates the likelihood that each new current product will be replaced each time a new product is prescribed for a given patient. This likelihood is then multiplied by the likelihood that the new product will be prescribed for a given patient and subtracted from the current share of that particular product. Since the share of the current product may not logically be negative, the formula is as follows:

$$P(C_j) = CU_j - \text{logit } P(R/X) * P(X)$$

where

- $P(C_j)$ = Probability of prescribing Current Product C_j for a given patient
- $P(R/X) = \text{Logit } P(R) = c + \sum C_{ji} X_i$ = Probability of replacing Product C_j when Product X is prescribed
- CU_j = Current share of patients receiving Current product j
- C_{ji} = Replacement coefficients
- c = Constant
- $1 \leq P(R/X) \leq 0$

such that $P(C_j) > 0$, else $P(C_j) = 0$

Thus the odds ratio is quite different for the replacement model which describes the likelihood that Product C will be replaced each time Product X is chosen.

Table 3:
Odds Ratios in the Standard Binomial and Replacement Approaches

	<u>Current</u>	<u>Scenario</u>	<u>Binomial</u> <u>Odds</u>	<u>Replacement</u> <u>Odds</u>
Product X	0%	10%	10/90	10/90
Product A	20%	20%	20/80	0/10
Product B	30%	30%	30/70	0/10
Product C	50%	45%	45/55	5/5

On the one hand, the replacement model easily handles current products that are added-to (untouched alternatives), since the likelihood that that particular current product will be replaced is zero, and the new share simply equals current share. On the other hand, cannibalized alternatives are replaced at a modeled replacement rate, contingent on the share change in new product. Since the replacement ratio ranges from 0 to 1, current product estimates range between current share and zero share, so the projected use can never exceed current share. Changes in the share of the cannibalized products are dependent on changes in the share of the new product, so the changes in projected use are compatible with changes in the share of the new product.

CASE STUDY: COMPARISON OF REPLACEMENT AND STANDARD CHOICE MODELING

To demonstrate the ability of replacement modeling to accurately model the impact of a new product in a polytherapy market, two survey data sets were modeled comparing the standard binomial approach and the replacement modeling approach. The two data sets were taken from a single study in which the estimated use of a new product was measured for two distinct patient types.

- n = 200 respondents
- Choice set = New product and 11 alternative choices
- New product design = 3 two-level and 2 three-level attributes
- Respondent task = Each respondent allocated choices for each patient type (≥ 100) for
 - Current patients
 - Future patients: 7 scenarios out of 12
- Two of the seven scenarios were designated as holdout tasks and used to measure the success of the two approaches to modeling the data based on the five remaining scenarios.

The marginal data for the two patient types follows. For patient type one, the average number of drugs prescribed for each patient is approximately 1.4; the average for the second patient type is approximately 1.3. In both cases the sum of the replacement rate exceeds one, indicating that the new product replaces just over one current product each time the new product is prescribed for a patient.

In the modeling set of scenarios for patient type one, for example, the average number of scripts written for the new product is 27%, with a replacement rate of 1.10, so that the average number of scripts is reduced by just under 0.3 ($0.27 * 0.10 = 0.27$). This is reflected in the reduction of the average number of scripts per patient from 1.44 to 1.41.

Table 4:
Marginal Data for Patient Type One

	Patient Type 1				
	Share			Replacement Rate	
	Current	Holdout	Model	Holdout	Model
Product X	0%	25%	27%		
Product 2	26%	23%	22%	16%	15%
Product 3	41%	36%	36%	20%	19%
Product 4	6%	5%	4%	4%	4%
Product 5	5%	4%	4%	3%	4%
Product 6	6%	4%	4%	6%	6%
Product 7	1%	1%	1%	1%	1%
Product 8	13%	10%	9%	12%	13%
Product 9	10%	8%	8%	9%	10%
Product 10	19%	15%	14%	16%	18%
Product 11	9%	7%	6%	9%	9%
Product 12	10%	7%	7%	10%	10%
Sum	144%	143%	141%	106%	110%

Table 5:
Marginal Data for Patient Type Two

	Patient Type 2				
	Share			Replacement Rate	
	Current	Holdout	Model	Holdout	Model
Product X	0%	9%	11%		
Product 2	21%	19%	19%	16%	14%
Product 3	20%	19%	19%	10%	11%
Product 4	4%	3%	3%	3%	4%
Product 5	43%	41%	40%	25%	29%
Product 6	4%	3%	3%	3%	5%
Product 7	1%	1%	1%	1%	2%
Product 8	7%	6%	6%	11%	11%
Product 9	6%	5%	5%	9%	11%
Product 10	11%	9%	10%	19%	15%
Product 11	7%	6%	6%	12%	11%
Product 12	8%	7%	6%	12%	13%
Sum	131%	129%	128%	121%	126%

ANALYSIS

The data was analyzed using latent class analysis for both approaches. To evaluate the models, the mean square error (MSE) was calculated by squaring the difference between the predicted and the actual shares for each holdout scenario for each individual.

To further evaluate the incidence of reversals, every potential scenario was evaluated by testing every possible combination of attribute levels and assessing the percentage of times that the predicted share exceeded the current share for any current product. Since the replacement model precludes such reversals (the predicted must be less than or equal to the current share) this time consuming analysis was only carried out for patient type one as a demonstration of the problems inherent in the standard binomial approach.

RESULTS

The chart which follows compares the actual holdout shares results with predicted shares from the standard binomial and the replacement approaches. As indicated by the standard error bars, both approaches accurately capture the hold results on an aggregate basis.

Chart 7:
Patient Type One Aggregate Results

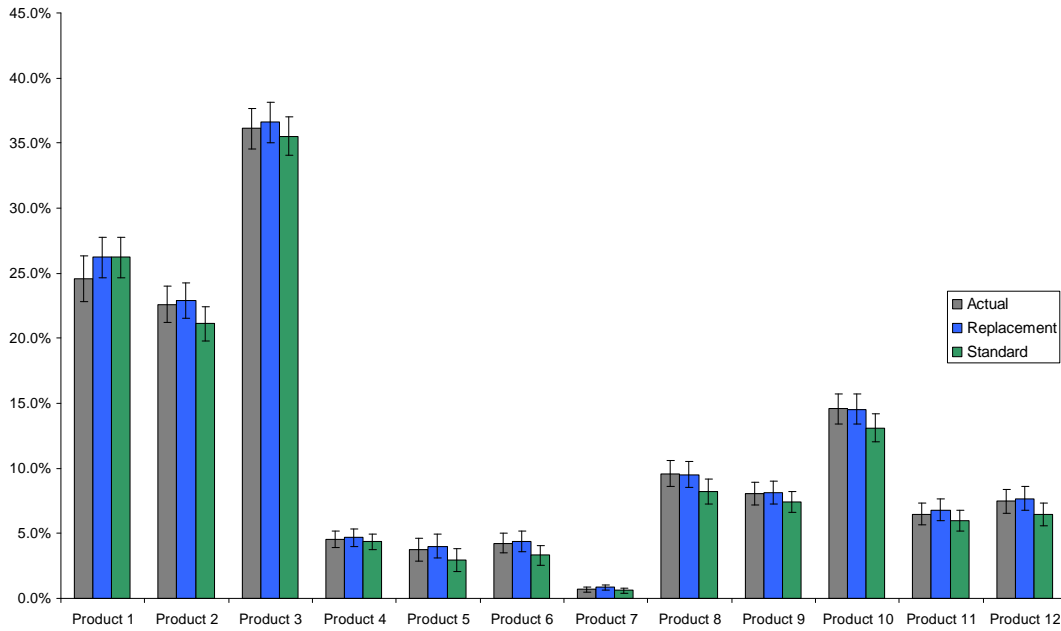
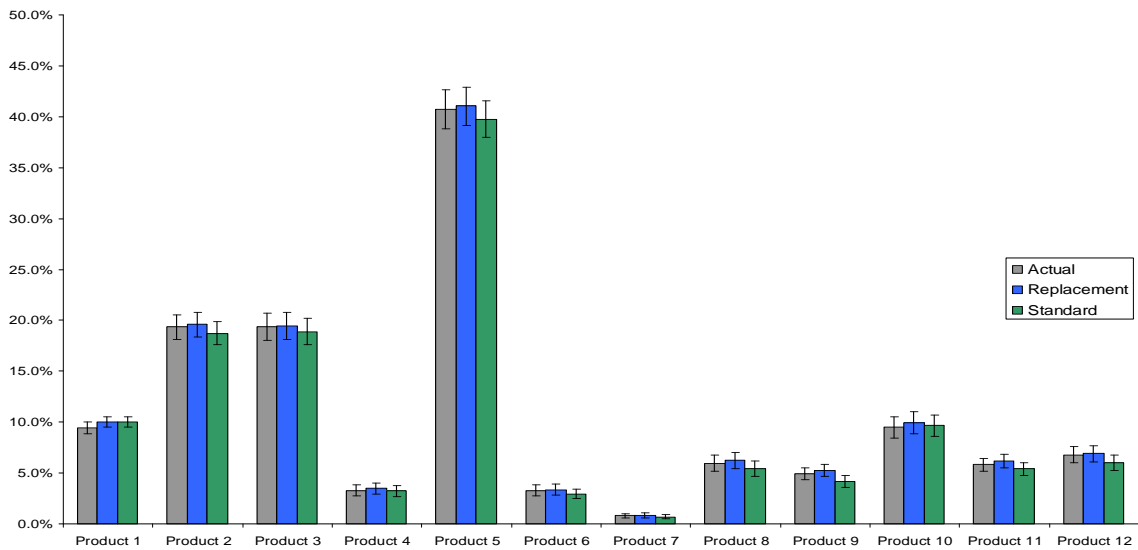


Chart 8:
Patient Type 2 Aggregate Results



The charts which follow show the average MSEs for each product for both approaches. The replacement approach consistently has lower error rates across the current products. Analysis of variance (repeated measures with holdout tasks nested within individuals) was carried out to assess the statistical significance of these differences. While the impact of the approach does not reach significance ($P < .05$) for patient type 1, the replacement approach is clearly significant for patient type 2.

Chart 9:
MSE Patient Type One

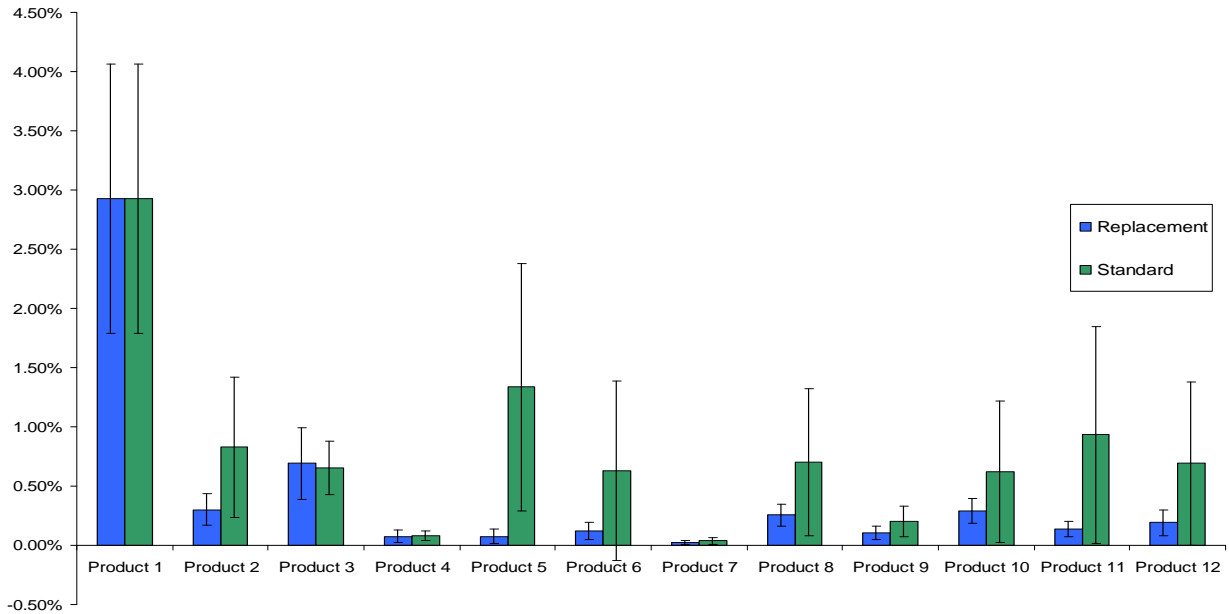


Chart 10:
MSE Patient Type 2

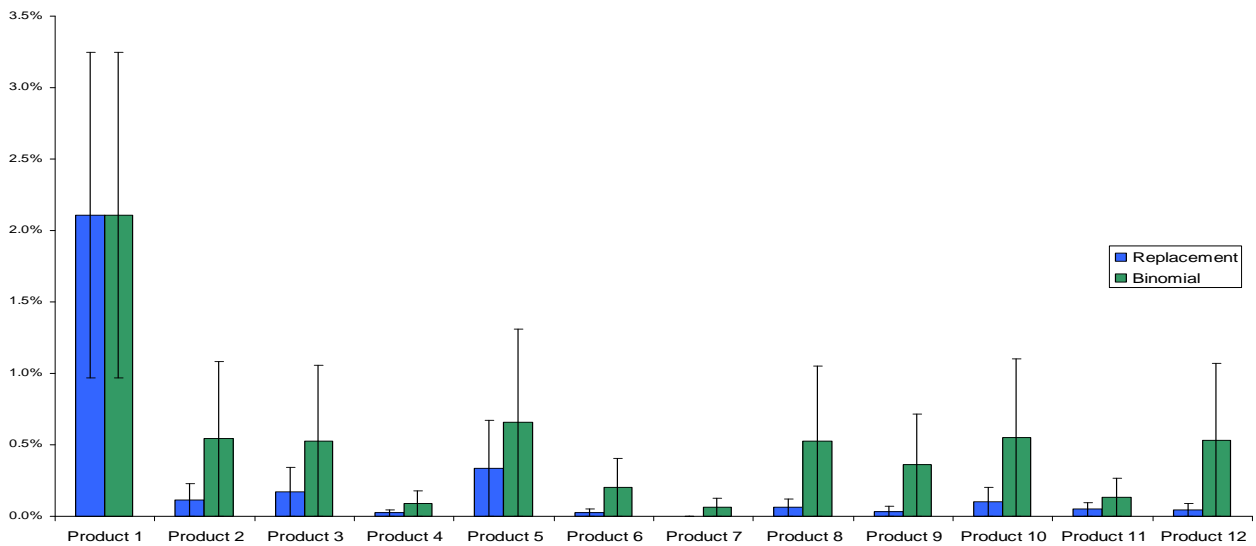


Table 6:
Patient Type One ANOVA

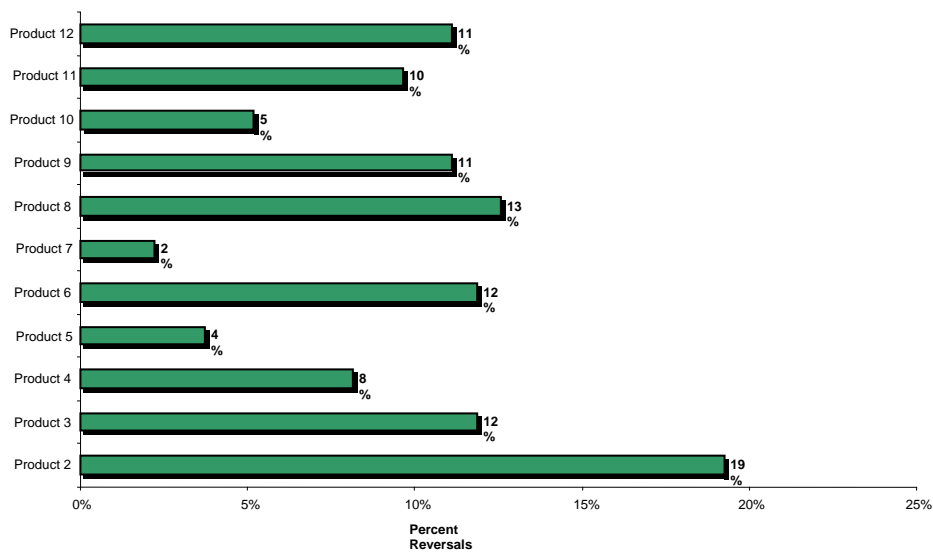
Source	DF	Anova SS	Mean Square	F Value	Pr >
Approach	1	243.57734	243.57734	3.51	0.062
Product	10	283.07138	28.30714	3.32	0.0003
Approach*Product	10	206.53686	20.65369	2.66	0.0031
Approach*holdout(subject)	268	18580.28479	69.32942		
Product*holdout(subject)	2700	22998.41541	8.51793		
Type*Prod*holdout(subject)	2680	20776.27998	7.75234		

Table 7:
Patient Type Two ANOVA

Source	DF	Anova SS	Mean Square	F Value	Pr >
Approach	1	151.45775	151.45775	6.99	<.01
Product	10	137.71414	13.77141	2.09	0.0222
Approach*Product	10	44.71739	4.47174	0.82	0.61
Approach*holdout(subject)	319	6912.43906	21.66909		
Product*holdout(subject)	3190	21028.04783	6.59186		
Type*Prod*holdout(subject)	3190	17410.63905	5.45788		

The reversal analysis carried out for patient type one shows the percentage of reversals that are found across all possible configurations of the new product. The percent of reversals range from 2% to 19%, depending upon which current product is analyzed.

Chart 11:
Reversal Analysis Patient Type One



These results demonstrate that the replacement approach is both more accurate in capturing the data and less prone to reversals.

DISCUSSION

When measuring the impact of a new product on a current market when the average number of choices exceeds one, the replacement approach captures the data more accurately than current approaches. Moreover, the changes in shares of current products that are observed when characteristics of the new product are changed are more intuitive as the changes in the current products are proportionate to the share changes of the new product. Products that are not replaced by the new product retain their current shares, while products that are heavily cannibalized are replaced at a greater than average rate.

Often, a replacement constant is sufficient to capture this effect, as a simple ratio, expressing the rate at which a current product is replaced will result in a share reduction in the current product proportionate to the increase in the new product. For example, the replacement ratios for single entity drugs that are combined into a new drug can be expected to have higher replacement ratios than other current drugs. As the attractiveness and thus the share of the new combination drug increases with changes in its characteristics, then the shares of the single-entity drugs that make up this combination will decrease proportionally, resulting in a more intuitively sensible model.

A second advantage of this approach is that it gives the modeler more control over the model so that characteristics of the new product, which might be expected to influence the replacement rate, can be included for a given product only in the expected direction. For example, an indication for monotherapy should increase the replacement ratio of most current products and can be constrained as such. These constraints can be specific to the current product being modeled. For instance, efficacy in treating a concomitant condition can be expected to increase the replacement ratio only in alternatives that primarily treat this condition.

While the value of replacement modeling is most evident in markets where the average number of choices exceeds the number of patients or occasions, it may also be useful when the allocation matches the number of patients or occasions. Current attempts to measure differential cannibalization in standard choice models often result in reversals as the changes in the shares of the current products are often disproportionate to the changes in the new product. The replacement approach does not suffer from this deficiency.

CONCLUSION

Replacement modeling offers a simple solution to the problem of measuring cannibalization in a polytherapy market. This approach can be used with any choice technique, including HB, latent class, or even aggregate logit. By modeling cannibalization directly, rather than viewing it as a by-product of the choice model, the resulting model is more accurate and less prone to counter-intuitive reversals.

REFERENCES

- Allenby, G., J. Brazell, T. Gilbride, and T. Otter (2005). "Avoiding IIA Meltdown: Choice Modeling with Many Alternatives," in *Proceedings of the Sawtooth Conference, 2004*, 209-218.
- Louviere, J. J. and G. Woodworth (1983). "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20: 350-367.

DATA FUSION TO SUPPORT PRODUCT DEVELOPMENT IN THE SUBSCRIBER SERVICE BUSINESS

FRANK BERKERS, GERARD LOOSSCHILDER
SKIM GROUP
MARY ANNE CRONK
PHILIPS LIFELINE SYSTEMS

ABSTRACT

This paper shows the relevance of data fusion by describing an actual example taken from the subscriber service business. Two major challenges in this industry are to acquire subscribers and retain them as long as possible. To be able to retain them, it is important to foresee that (possible) deactivation from the service may be imminent sufficiently long before that deactivation actually happens, to give the business sufficient time for a proper response.

In this paper we show how to do it in two (interrelated) steps. The first we call the *actionability* step. In this step we look for typical reasons for deactivation: why and where our clients went after deactivation. This information is used to influence those who may be about to deactivate an offer to stay on the service instead. The selection of potential offers is determined by concept development research. Here, we append custom concept testing data to the subscriber base to determine which service offer resonates best with which type of client and with which potential reason for deactivation.

The second is the *predictability* step: by looking at the service call pattern in the administrative database, we try to predict if deactivation is imminent. We look for “early warnings”; patterns that tend to precede deactivation.

We track our success rate of offer acceptance, and although actual results are still due, we have high hopes and good indication that our approach offers great potential for postponing deactivation, and thereby, for improving subscriber lifetime value.

INTRODUCTION

Data fusion is not new. For example, a list compiled by Robert Soong and dated on October 6, 2006 contains 1303 references to papers carrying “data fusion” or associated subjects in their titles.¹ Yet, we have not come across many papers describing the use of data fusion in market research to support product development. This is surprising because data fusion offers benefits that make it very applicable in this context.

With Data Fusion we mean the merging of datasets where the original datasets were collected separately, usually for different purposes. A key benefit of data fusion is its synergistic effect. With this we mean that by combining datasets, we can learn more than the original datasets had to offer individually. Data fusion offers at least two synergistic opportunities: (1) deeper

¹ The oldest referenced paper carrying “data fusion” in the title was dated in 1986. The term “data fusion” may have been coined around then, as the list of references goes back much further.

diagnostics, where data in one set are used to diagnose and explain effects in the other set, and (2) forecasts, where we use data from one set to predict future behaviors in the other set.

In this paper we show how we apply data fusion to support product development in the subscriber service business. Two main topics in the subscriber service business² are (1) to attract new clients (acquisition) and (2) to keep them as long as possible (retention) to optimize the total of subscribers' lifetime value and recover the investment in acquisition. We are particularly focusing on predicting the moment the subscriber may deactivate her subscription. Preventing deactivation is an important way to drive subscriber lifetime value.

The structure of this paper is as follows. After this introduction, we continue with a review of the benefits of data fusion in the subscriber service business. We demonstrate that the use of data fusion is driven by need as much as opportunity. Next, we introduce Philips Lifeline, the subscriber business in which we implemented data fusion. In the following sections we describe how we applied data fusion to predict the moment of deactivation, and if deactivation is imminent, how we tried to preempt that by making the right service offer. The "right service offer" was identified by collecting concept test data and appending them to the subscriber data base. Once that is done, we review, one last time, how data fusion helped to predict and prevent deactivation, to finish with some practical recommendations.

About data fusion in the subscriber service business

We suggest that data fusion is relevant to the subscriber service business because of a need as well as an opportunity. A key business challenge in this industry is to drive business growth through acquisition (by servicing more clients) and retention (keeping clients for longer). Examples of these segments are cable, telecom, insurances, public utilities, security services and financial services. In Europe, this has become even more challenging now that the industries are deregulated and have become more transparent in their pricing. It has become easier for consumers to switch, and as a result, they are less loyal, irrespective of their satisfaction levels. By the end of the contract date, it has become more common to switch to the supplier with the best offer: the lowest cost or best value for the money. Industries have to constantly focus on the acquisition of new clients, and costs are high. Once a client is being served, it is necessary to keep her as long as possible to optimize life-time value and generate a positive return on investment.

Another reason for promoting data fusion is because of opportunity. Data are available for "free". Because of the abundance of data, the key added value market researchers and analysts can provide is turning it into information, knowledge, and ultimately, wisdom. Data fusion may help here, because by combining data from different sources, one may gain more insight than each individual source may offer.

We promote combining of data from various sources, both internal (already available in the company) and external (from outside). As will be shown in this paper, an obvious candidate for data fusion is subscriber data, available in the service provider's administrative database. One may say that this is the backbone for our data fusion exercises. The main activities are to

² In this business, customers sign up and receive the service in exchange for a monthly fee. Philips Lifeline is an example of a subscription service, offering a Personal Emergency Response Service to elderly people living alone. Subscribers can choose to wear a device around their neck or on the wrist. If they fall, or wish to talk for another reason, they push the button and are relayed to a call center dispatcher, who offers a listening ear and defines a proper follow up, sending help if needed.

discover and detect patterns in the service usage data, and to couple those with subscriber background characteristics already known to the service company.

The “picture” of our clients becomes even clearer if external data are appended to better explain variations in service usage patterns. The current administrative dataset “just” describes the initiation, consumption, modification, and deactivation of the service as a series of transactions or events. They do not tell us why the transactions were made or events have happened, and they do not tell us what kind of client she actually is either. Hence, we would like to go back in to enquire and answer the question “Why?”. This question can be answered by appending data from various, external sources, consisting of either available external data, or of the results of custom market research. For example, this can be syndicated data from dedicated data warehouses and CRM companies, (e.g., Acxiom or Experian), or information of other companies and their administrative databases, such as credit card or insurance companies, or banks. Alternatively, we can also append custom data, such as the results of a concept test or conjoint analysis study into the consumer’s acceptance of service options and upgrades. These data are readily available and are easy to append to the subscriber base, and we have seen some phenomenal results:

Acquisition

Subscriber profile information is integrated with the company CRM system. Marketing could use this information in a variety of ways, the first of which was to identify the type of consumer that feels most attracted to the service. They also identified new populations by purchasing marketing databases that matched the ideal profile.

Retention

A group that forecasts utilization of services could identify subscribers who are more likely candidates for accepting certain service upgrades, or ones who had higher probability of participating in an incentive/rewards program versus another type of program.

Revenue forecasting became more accurate as additional subscriber information became manageable and part of the profile, instead of another factor introducing uncertainty. Results of service upgrade programs and other retention programs can be integrated back into the system, further strengthening the accuracy of the model.

Data Fusion can help to combine and convert various datasets into actionable information to drive and support the development of new business initiatives. However, data fusion as we describe it is just an analytic activity, whereas in fact, it is a lot more. Data fusion is just part of a mindset change that works its way all through the organization, putting the client at the center of attention. Data fusion may help to better understand who the client is, what she does, and why she does it. It may also become a good reason for organizing the business differently, having it revolve around “acquisition” and “retention” programs. This way, the business may become more proactive, focusing on what it wants to accomplish, and allowing for a “just-in-time response”, for example, preempting an undesired deactivation.

BUSINESS CASE: PERSONAL EMERGENCY RESPONSE SERVICE

Philips Lifeline Systems is the leading provider of personal emergency response services (PERS) in the United States and Canada. Lifeline's mission is to help people live independently, longer. Lifeline supports almost 700,000 subscribers (as of August 2007). During its 30+ year history, the company has helped more than 6 million seniors remain safer at home.

Using a personal help button (PHB) a subscriber is able to directly contact trained Lifeline response associates 24 hours a day from around their home. A Lifeline monitor assesses the individual's need and quickly sends the appropriate support. Support may be from a designated neighbor or relative for situations like helping with medications or it can be from emergency services such as fire and ambulance.

The average subscriber is 82 years old, female, with some mobility problems. They often have one or more chronic conditions such as diabetes, arthritis or high blood pressure. Many have some cognitive impairment from stroke, Alzheimer's, or other deterioration.



Figure 1:
Personal Emergency Response process

The subscriber usually comes to Lifeline after a life altering event such as a fall. A healthcare provider, social service organization, referral network (such as a case worker, discharge planner, geriatric nurse) often recommends their patient obtain a Lifeline subscription so they may resume living in their home.

Coming via the healthcare channel means their information is protected under HIPAA; this combined with the potential cognitive issues, makes traditional market research less effective for assessing new products and services.

The biggest revenue challenge comes internally from its own subscribers. The average length of subscription is 24 months. The time to recover the cost of subscriber acquisition is limited as the subscriber base is virtually completely refreshed every third year. Thus, extending subscription time has an immediate positive effect on profitability. One month of prolongation means 4% more revenue and more profit.

Although the average is 24 months, subscribers are not homogenous in their tenures. At this fragile stage of their lives, some may be just a month away from hospice while some remain on the service for several years. With such huge variations, determining a method to increase the average length of subscription, and better control and/or delay service deactivation becomes a complex exercise.

Yet, although the subscribers are unique in their combination of health and attitudes, the business process model follows the lifecycle of Acquisition, Retention and Deactivation. The same business questions to be addressed include: can we delay -- make a subscriber stay for longer? Can we control – i.e., keep targeted subscribers?

The focus of the data fusion will be on deactivation as there is more ability to accurately identify intervention situations than through acquisition and retention activities as discussed below.

Acquisition

This is an incident driven market, and acquisition occurs after an incident such as a hip fracture, stroke, or other event that disabled a potential subscriber enough to make them housebound, yet not enough to force them into a nursing home or other continuous care facility. Predicting the point of acquisition is thus difficult.

Retention

Retention within this model is difficult to predict and control for a variety of reasons. Although the likelihood of switching to a competitor is low, the likelihood they might move to hospice or a nursing home due to a slight change in health status is high. This population is extremely frail and often declining. The individual is also likely to be taking almost a dozen medications. One additional health issue or incident increases the likelihood of a catastrophic result almost exponentially. One of the objectives of effective treatment is to simply keep from adding more health issues--as opposed to being able to eliminate all health issues entirely as could be done in a younger more virile individual. The ability to predict retention involves multiple variables unrelated to traditional measures of retention.

Deactivation

As with retention, deactivation can come from a variety of sources, both voluntary (moving to other in home care) to involuntary (death or health condition that prevents continued occupation of their home).

The service provider needs to develop a better control over deactivation. The objective is to better control each of the factors, focusing on attracting the subscribers that we want and retaining the subscribers that we wish to keep, for longer. Before identifying when and why the subscriber can be expected to deactivate, there are still some items to consider.

Actionability

From a revenue standpoint we must make the right offer to the right person; otherwise there is no interest. Data fusion affords us the ability to piece together attitudes, health, living arrangements and demographic abilities for us to identify services that may help delaying and controlling deactivation.

Predictability

Revenue will be generated only when the offer is made at the right moment; in this event driven model the revenue can only be had if there is a lag between the offer and the original point of deactivation. Data fusion allowed us to detect patterns that are a precursor of imminent deactivation.

APPLICATION OF DATA FUSION TO THE SUBSCRIBER SERVICE BUSINESS

Research challenge

The business challenge at hand is to prolong subscriptions of the customers who are or may be about to deactivate their subscription. Again, prolonging the deactivation period with only one month roughly equals roughly 4% of extra revenue (at likely even less cost). This challenge has two basic business requirements: (1) Actionability. People about to deactivate have unmet needs. It is the challenge to address these needs by the right proposition. This is the activity of new concept development and acceptability testing. (2) Predictability. There is a reason, a course of events, why these people have unmet needs. The challenge is finding the right moment to intervene. We are looking for an indicator pattern that precedes deactivation, but allows for intervention.

Practical challenge

In our case, we are confronted with the following practical challenge: New concept acceptability can be tested with current or future candidate deactivators, not with those who already deactivated. We cannot find indicators of deactivation in the administrative dataset of those who did not deactivate yet.

So we will have to combine our current subscriber's preferences for newly developed concepts and patterns found in our past subscriber's behavior and apply them to our current subscribers.

Actionability

As part of the deactivation procedure we recorded reasons for deactivation of over 60,000 former clients. Note that this data is not strictly needed to support our primary process, so it is custom data appended to our administrative dataset.

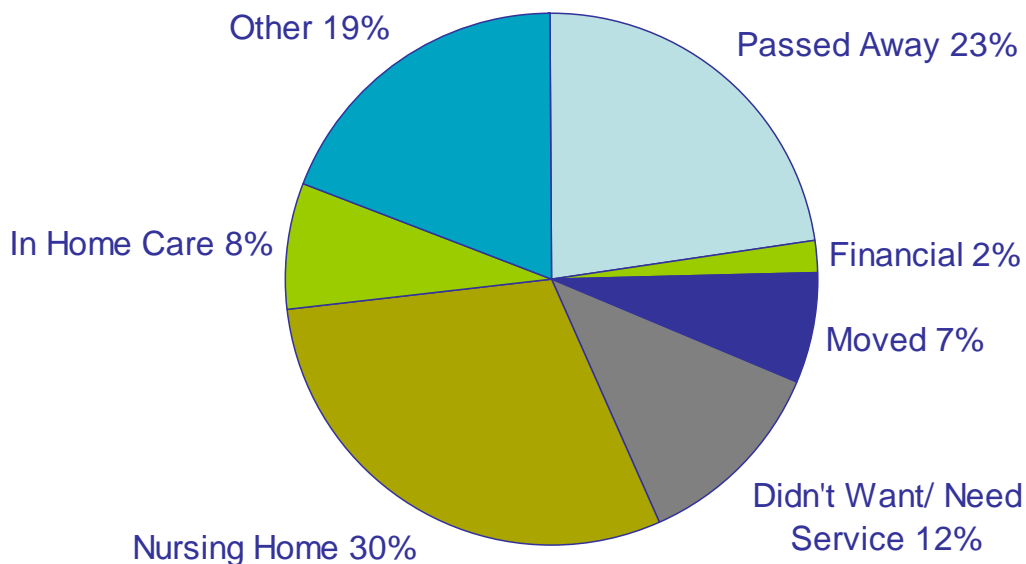


Figure 2:
Reasons for Deactivation

Adding this piece of information in itself was interesting and it did give us a hunch of what we could have done to fulfill their needs, but it seemed too broad to be actionable.

Next we added some more custom data, such as net worth and zip-coded psychographic data (Forrester, Personix) and segmentation schemes such as the Moschis segmentation (external syndicated data). Not a true surprise, but this revealed that, for example, clients with higher net worth are much more likely to move to in home care, whereas clients with low net worth tend to stay at home until they qualify for the nursing home. Psychographic variables showed us strong correlations with “Didn’t want/need service”, as some types of personality are more likely to deny any need for help. Another (not so surprising) example of what it showed is that people in the larger houses tend to move and therefore deactivate.³ Other contextual information sources used in this case are CDC, Census and NHIS.

Now we know so much about the context of those who deactivated, we must create propositions that meet the changed needs of those about to deactivate to keep them from deactivating. In this step we used several types of mostly qualitative techniques, such as focus groups, laddering (means-end-chain) to generate new or revised products based on their profiles⁴.

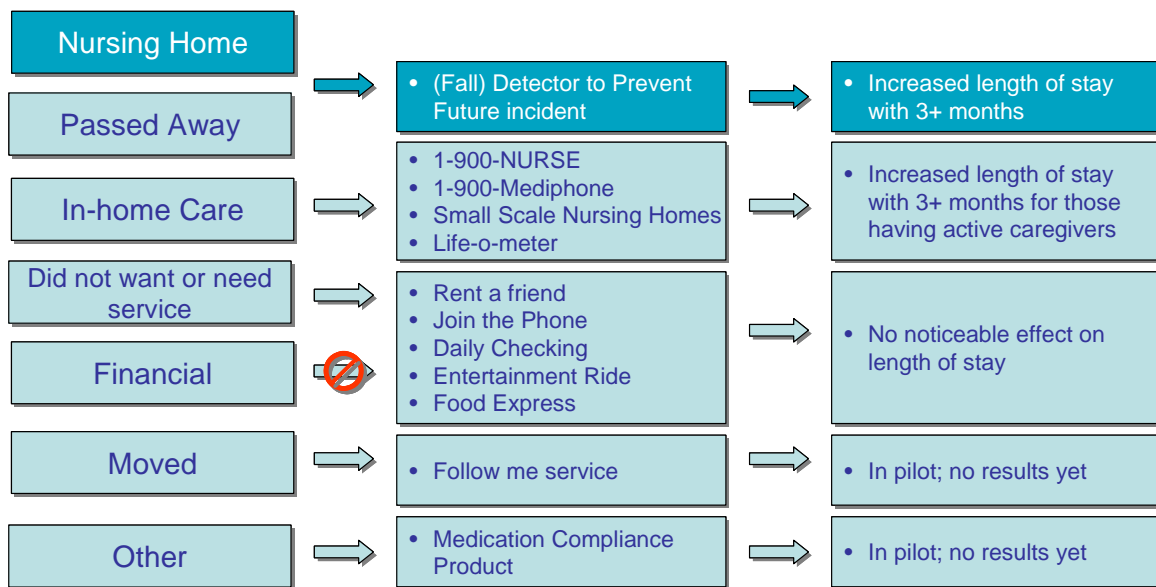


Figure 3:
Addressing Reasons for Deactivation with New Service Concepts

³ At this point we have to acknowledge that the process of recording the reason for deactivation is not at all error free, .

⁴ We found some facts about “the fall”:

1 out of 4 elderly who sustain a hip fracture will die within 6 months

1 of 2 80+ will fall this year. Those that have fallen are 2-3 times more likely to fall again

1 out of 4 elderly that fall suffer moderate so severe injuries that require ER services

In general “the fall” is seen as a signal that the last phase of physical life has started. Some of our clients will fiercely avoid admitting they have done so as if that prevents them from entering that phase. This makes it sometimes hard to classify calls well. We have found that “fallers” with a certain profile are likely to leave for a nursing home. Given the above, it should be clear that a product that helps preventing a fall and detects a fall addresses present needs.

Our administrative dataset of calls emphasized that all concepts should address social-medical needs. Most of these concepts are tested in pilots, and results so far vary from “only” improved client satisfaction (mostly social activities) to increased length of stay of over three months (fall detector).

We built “pictures” of the subscribers (who deactivated) by appending custom data (e.g. reasons for deactivation, net worth) to administrative data (client data, calls registration) and also appending external syndicated data to get insights. (Who is the subscriber? What may she need? How does she behave? What does she like and when?)

These “pictures” (or strongly descriptive profiles) served us in preparing a number of new services that can be offered to address one (or a few) reasons for deactivation of clients matching a specific profile.

Predictability

While in the actionability step the offers are prepared, we must also determine when to offer it. It is the challenge now to see when it is a good moment to expose our offers to the client. In most cases this is when we have learned that something has changed – this can be an emergency call, but it can just as well be a silence for too long.

In our administrative dataset of the deactivated subscribers we have records of when our clients called and a classification of the nature of that call as well as the chosen follow-up response (if any). So a selection of records for one subscriber could look like:

<u>date</u>	<u>time</u>	<u>client</u>	<u>call type</u>
31-Jan	19:56	310788344	social
4-Feb	1:58	310788344	social
4-Feb	3:53	310788344	social
5-Feb	21:57	310788344	fall
7-Feb	8:44	310788344	social
21-Feb	9:02	310788344	social
21-Feb	10:12	310788344	social
23-Feb	7:53	310788344	social
24-Feb	19:10	310788344	panic / false alarm
3-Mar	10:52	310788344	social
4-Mar	18:38	310788344	panic / false alarm

Figure 4:
Example of administrative dataset

Now, could we have done something extra after we receive the panic call at March 4, something that could have prevented the client from deactivating? The presumption is that we can, but first we’d have to find if we can signal something at this event – if we can predict an imminent deactivation.

In order to arrive at that we transformed the dataset, this list of events, into something more meaningful, because (we assume) it is not a single call that will precede deactivation. We summarized the calling pattern of our client’s behavior of the last months by classifying the calls in that period⁵. Note that, in line with the assumption that it’s not a single call, there has to be some time lag between the pattern to be detected and the deactivations to be ‘planned’ in order for us to be able to address the changing need and situation of the client.

⁵ Examples of call classifications are: Injury, Breathing problem, Chest pain, Dizziness, Fell.

At first we got mixed signals here, because one typical pattern (e.g. ‘intense calling’) would precede deactivation for one client, whereas the same pattern for another did not tell us a thing! After some juggling around with data we finally composed the “picture” of (1) *change* in pattern, (2) medical status and history of the client and (3) custom and psychographic data and that gave us a fairly good guess if something was about to happen. Thus by appending contextual information we are able to interpret behavioral patterns as “early warnings.”

The output of the predictability step is the definition of behavioral patterns that in conjunction with the client’s profile “early warns” an imminent deactivation.

Combining actionability and predictability at runtime

Now that we have the definition of “early warnings” (based on past subscribers’ behavior) it is still a challenge to do something about the needs of our current subscribers when they are in a similar position. A change in pattern can appear every time the phone rings, but also when it does not ring (when the number of calls drops). This implies we have to check for, or calculate, patterns regularly. Subsequently we have to ‘flag’ clients for whom, based on their profile and our warning definitions, their current recorded state indicates a risk of deactivation. This procedure is referred to as the *scoring* (e.g. using logistic regression or data distillery type of processes) of the clients in our subscriber base.

By matching profiles and reason for deactivation we also know what to offer. For example when the ‘repeated near fall’ pattern occurs and the person has a profile with medical conditions at entering the emergency response service, lower net worth and is on medication – chances are: she’ll be leaving for the nursing home – we will offer the (Fall) Detector device and service to give a little extra confidence and security, to prevent a future incident. (And consequently we’d expect an extra 3 months of subscription.)

The actual offering is executed in various ways through normal channels of contact; such as communications with the person or their caregiver..

Note that this event driven ‘machine’ is programmed to offer proposition ‘a’ to anyone matching profile ‘z’ on scoring of pattern ‘k’.

CONCLUSION

Data fusion is a conceptual derivative from sensor fusion, in which images are composed by combining, for instance, radar, sonar and infrared information. This is exactly the approach we applied: We learned patterns that lead to deactivation from past subscribers. We enhanced these by appending custom and external data based profiles. Based on these profiles, and needs inferred from the patterns, combined with concept development research techniques, we prepared new services.

We applied the data fusion runtime in an event driven system that continuously checks for patterns to occur and derives if that implies an “early warning”. If it does, then the appropriate offer is offered.

At least in one respect data fusion was indirectly successful. The process of composing a “picture” of the client also turned the business from a process driven one into a client context focused organization. Now that this market moved from a reimbursed model to out of pocket

payment the need to address individual needs is nothing less than a requirement for market presence.

The inspection of patterns combined with client contextual information and external data gave a solid basis for conducting concept development. So data fusion did deliver deeper diagnostics.

The effects of our efforts are monitored in a pilot study and as mentioned before, results so far are promising, but we have to note that there are a few peculiarities in this procedure.

The first is that the scoring is a far from exact process – it indicates that it is somewhat more likely that a deactivation is about to happen. Furthermore, we address the situational context of our client. It is not determined whether it is the timing or the offer or both that make a difference. We feel that the recognition of a change does have a positive effect on the relation with our clientele.

The second is that by launching new services (and perhaps also just by the passing of time) a new type of client may join our services. We did a lot of investigation on this particular dataset of past subscribers. This exercise may in due course have become obsolete, or to put it mildly, must be rerun every now and then.

REFERENCES AND DATA SOURCES

Moschis (1) - “Gerontographics: Life-Stage Segmentation for Marketing Strategy Development” - George P. Moschis, 1985.

Moschis (2) - "Gerontographics", Journal of Consumer Marketing, Vol. 10 No.3, pp.43-53.

Moschis, G.P., 1993.

Soong - “Data Fusion Bibliography” - Robert Soong, 1996
(www.zonalatina.com/datafusion.doc).

Census –U.S. Bureau of the Census and the Administration on Aging, Department of Health, Education, and Welfare (www.census.gov).

CDC –Centers for Disease Control and Prevention (www.cdc.gov).

Forrester –Forrester Research (www.forrester.com).

NHIS – National Health Interview Survey/National Nursing Home Survey (www.cdc.gov).

Personicx – Demographic segmentation data (www.acxiom.com).

MULTIPLE IMPUTATION AS A BENCHMARK FOR COMPARISON WITHIN MODELS OF CUSTOMER SATISFACTION

*JORGE ALEJANDRO
KURT A. PFLUGHOEFT
MARKET PROBE*

ABSTRACT

Missing values are a reality that market researchers have to face on a regular basis. Several popular techniques to handle missing values are reviewed and compared against multiple imputation. A simulation study is used to compare the accuracy of the statistical estimates of a regression model under each technique.

INTRODUCTION

Multivariate analysis of satisfaction survey data is often plagued by missing values. Although statistical theory and simulation studies have shown that there are preferred imputation methods to handle missing values, market researchers have relied on less favorable approaches. Critical data scrubbing steps may not be properly addressed since researchers are under considerable time pressure to produce results. Consequently, biased statistical results may go unnoticed by researchers and clients.

Since the cost of data collection usually far outweighs the cost of data analysis, it is important not to waste information contained in the data to produce accurate estimates (Harrell, 2001). There are numerous ways to handle missing data including non-imputation methods such as list-wise deletion as well as single imputation methods such as mean substitution. However, many missing value techniques fail to account for the uncertainty in missing data; i.e. how confident are we in ignoring or substituting values? Advanced methods such as multiple imputation (MI) may provide more accurate estimates of not only point estimates but also their standard errors (Allison, 2002).

In this research, the authors briefly review popular missing value methods and utilize simulation methods to examine bias in deriving statistical estimates for regression models of customer satisfaction. Both non-imputation methods (i.e. list-wise deletion, pair-wise deletion and the missing-indicator method) and single imputation methods (i.e. mean substitution, conditional mean substitution and expectation maximization (EM)) are compared against multiple imputation. The comparison is carried out by inducing missing values on a complete data set to show the impact of imputation in terms of bias from the “known” values. The missing values are induced in such a way that is consistent with how they occurred in the original data.

2) TYPES OF MISSINGNESS

The pattern of missingness within a data set may fall into any one of three categories (Rubin 1976): Missing Completely At Random (MCAR), Missing At Random (MAR) and

Non-Ignorable Missingness. MCAR situations imply that there is no pattern in the data set and that the data truly are missing in a random fashion. MCAR situations are assumed to be rare unless they were created as part of the experimental design. Although many missing value techniques perform well under MCAR, there are some techniques that rarely produce the appropriate statistical inferences even under this least restrictive situation

MAR implies that there is a slight pattern to the missingness but this situation is usually correctable, however, practitioners and academicians often confuse MAR as MCAR. Just because there is an “AT RANDOM” component in MAR does not indicate that there is no pattern of missingness. MAR situations are thought to be a more commonly occurring type of missingness, but, there is no definitive test to show that your missingness actually falls in this category.

The last category is Non-Ignorable missingness and as the name implies there is usually no statistical correction for such situations. To remedy this problem, the researcher would need to 1) have a thorough understanding of your data, 2) know the processes that govern missingness, and 3) be able to address the missingness in a way that leads to unbiased and precise results. The previous steps really apply to all types of missingness. Analysis of data with non-ignorable missingness can jeopardize all the results of a study.

This research will focus on MAR situations since MCAR is relatively easy to correct and Non-Ignorable is nearly impossible to correct. Following is a simple example of a MAR situation but the reader can find a precise definition of MAR in Little & Rubin, 2002. For example, we observe that high income customers are less likely to give you an evaluation of bank tellers than lower income customers. If this is the only pattern and that the missingness of this variable within high income customers is at random, you can correct this situation as long as you know their income levels. If, for example, income level was also asked on the survey and both income and teller evaluation are missing, this is a non-ignorable situation.

3) MISSING VALUE TECHNIQUES

Before addressing the design of our simulation study, let’s briefly examine a few popular missing value techniques. These techniques can be roughly classified as either imputation techniques or non-imputation techniques. Imputation techniques attempt to substitute a “reasonable” value (otherwise known as an imputed value) for the one that is missing. Once all the missing values have been replaced with imputed values, complete cases analysis can be conducted. Non-imputation techniques do not attempt to directly substitute a value for the one that is missing. An example of a non-imputation approach is the popular list-wise deletion approach which is often the default method for handling missingness.

4) NON-IMPUTATION APPROACHES

The following non-imputation techniques are examined: list-wise deletion, pair-wise deletion and the missing indicator method (MIM). List-wise deletion deletes all records where at least one of the analysis variables is missing on that record. This technique works well in a variety of settings but as missingness becomes more pronounced, an unacceptable number of records may be deleted. Even if the per item rates of missingness are low, list-

wise deletion still tends to discard an unacceptable high proportion of subjects; leading to smaller sample sizes and larger standard errors (Shafer & Olsen, 1998).

Pair-wise deletion works by estimating the variance/covariance matrix using only the pairs of variables in each record that actually exist; thereby using all of the available data. However, examining multivariate datasets in a pair-wise manner can be problematic for estimating statistics. First, the sample size differs depending upon what pair of variables is examined. Second, problems such as invalid correlations can arise. (i.e. exceeding the range of -1 and 1). There are some corrections that lead to consistent estimates for pair-wise deletion however no major statistical package has implemented them so far (Allison, 2002). Consequently, pair-wise deletion is often avoided and decried. However, there are many other missing value techniques which should really receive the same amount of scrutiny.

The missing indicator method was a popular method and taught through the 1990's at universities. The technique is used in conjunction with regression analysis where a dummy variable is created for each predictor in the data set. When the predictor is missing, the value "1" is assigned otherwise the value "0" is assigned. After adding the dummy variables, the claim was that you could use the entire data set with the dummy variables in a regression context. The regression analysis would give beta coefficients for both the real and dummy predictors, however, the beta coefficients for dummy variables could be ignored. Why? Those beta coefficients represent the effects of the missingness in isolation and consequently the researcher may discard them. It sounds like an elegant approach but again anytime someone tells you to ignore half of the results from a multivariate analysis, you need to question things. Fortunately, in 1996, Jones proved that the MIM is even biased in MCAR settings, the easiest situation to correct for.

5) IMPUTATION APPROACHES

There are a variety of imputation methods which can be employed when data are missing: versions of the following techniques are investigated in this study: simple mean substitution, conditional mean substitution, expectation-maximization (EM), and multiple imputation.

With simple mean substitution, the overall mean for the variable can be substituted for each value that is missing for that variable. This process maintains the original mean but adversely impacts many other statistical estimates. Variances and correlations may be dampened and test statistics are often overestimated as their standard errors are underestimated (Donders *et al.*, 2006). Despite the known problems with overall mean substitution, it appears to be heavily used in market research as clients rarely address such issues.

Another approach for imputation is to substitute a conditional mean. For example, a regression analysis could be used to determine the expected value of a predictor given a set of values for the remaining predictors. Such approaches may work well but they can exaggerate results once complete cases have been created from the imputed values. For example, if x_1 is used to predict x_2 , as the amount of missingness increases for x_1 the relationship between x_1 and x_2 will become more deterministic. Furthermore, when multiple predictor values are missing within a record, the process requires multiple steps or hybrid techniques. In SPSS, regression imputation is used by estimating several regression equations for each pattern of missing values (von Hippel, 2004).

Another popular “imputation” approach can be achieved by the EM algorithm. EM is not specific to missing values and it can be used to achieve statistical results such as regression without the need for an imputed data set, per se. The reason EM is included in this category is that we utilized SPSS’ implementation of the EM algorithm where an imputed EM data set can be requested for further analysis. The EM algorithm utilizes maximum likelihood estimates but the details of this approach are beyond the scope of this paper – see Dempster *et al.*, 1977.

The last technique that we will examine is multiple imputation. The general argument against all other imputation approaches is that they fail to account for the inherent uncertainty associated with the imputed value. In other words, how confident are you that your point estimate represents the “correct” value? This information cannot be gleaned from single imputation approaches as the subsequent analysis cannot magically determine if a particular value represents an original one or imputed one. MI addresses this inherent uncertainty by creating multiple imputed data sets. In each data set, the missing value will most likely have a different value. Consequently, it is claimed that the standard errors associated with the statistical estimates are more representative as they account for this inherent uncertainty.

Using MI is more complex than the other techniques as now there are many copies of the original data set with imputed values. Not only must you manage these data sets but also you need applications that can use these data sets to come up with one set of statistical estimates. The multiple imputation process is shown in Figure 1. Each data set represents an imputed version and each result represents analysis on one imputed data set. Eventually, all results must be combined to provide one set of statistical estimates. Although the MI approach can be more complex, this approach can give researchers greater confidence in their results since it produces estimates with optimal properties: consistent and asymptotically efficient (Allison, 2002).

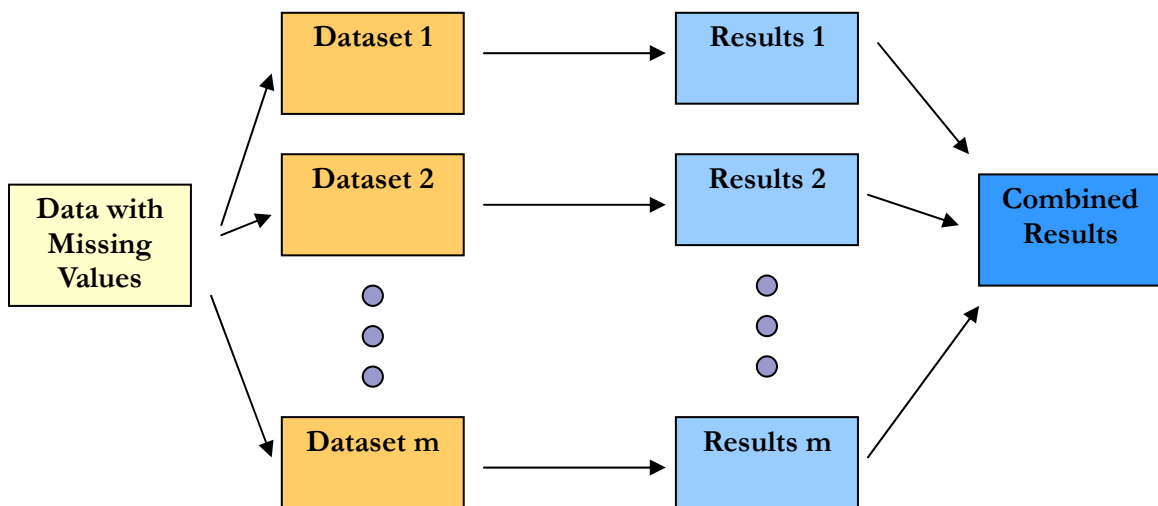


Figure 1.
Multiple Imputation Process.

In order to combine the results from analyzing the m complete datasets after using multiple imputation, available procedures like MIANALYZE in SAS, allow the analyst to combine the point and variance estimates for a given parameter.

If Q is the parameter of interest (such as beta coefficients) and $\hat{Q}_i; i=1,2, \dots, m$, are the m point estimates from each of the imputed data sets, a combined point estimate is defined as the average of those point estimates:

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$$

The estimated variance associated with \bar{Q} is a combined variance of the within-imputation and between-imputation variance (Rubin, 1987). Defining \hat{U}_i as the estimated variance from each imputed data set, the total variance is computed as follows:

Within-imputation variance:

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i$$

Between-imputation variance:

$$B = \left(\frac{1}{m-1} \right) \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$$

Total variance:

$$\text{var}(\bar{Q}) = \bar{U} + \left(1 + \frac{1}{m} \right) B$$

6) EXPERIMENTAL DESIGN

A common type of analysis in customer satisfaction studies is to predict overall satisfaction (OSAT) using customer satisfaction scores with various touchpoints such as teller, personal banker, web and drive-thru. Such analysis is consistent with Fishbein's Multi-Attribute Attitudinal (MAA) (Fishbein, 1967). The MAA model indicates that a customer's satisfaction can be derived/ranked using the performance and importance scores of the salient attributes, in this case, the touchpoint ratings. To estimate this relation, a regression model is created where OSAT is the response variable and the touchpoint satisfaction variables are the predictors.

Before missingness is introduced, the values of the beta coefficients and their standard errors can be estimated and saved for comparison. For future reference, these values will be referred to as the "known" values. Missingness can be artificially induced in a MAR fashion

to see if the missing value techniques in conjunction with regression analysis can recover these “known” values. Since a particular instance of missingness could adversely impact one particular technique, we repeated the experiment 100 times at four levels of missingness: Mild, Moderate, Major and Severe.

The “original” data set was constructed using a perturbed version of a customer bank satisfaction data with 904 complete cases. Bank satisfaction data tends to be heavily left-skewed and some researchers have suggested that banks may have hit a satisfaction ceiling (American Banker, 2006). Figure 2 graphically illustrates the concentration of high ratings for the OSAT variable; the predictor variables tend to be skewed as well. All touchpoint satisfaction ratings were measure on a 10 point Likert scale where 1 was labeled “Very Dissatisfied” and 10 was “Very Satisfied.”

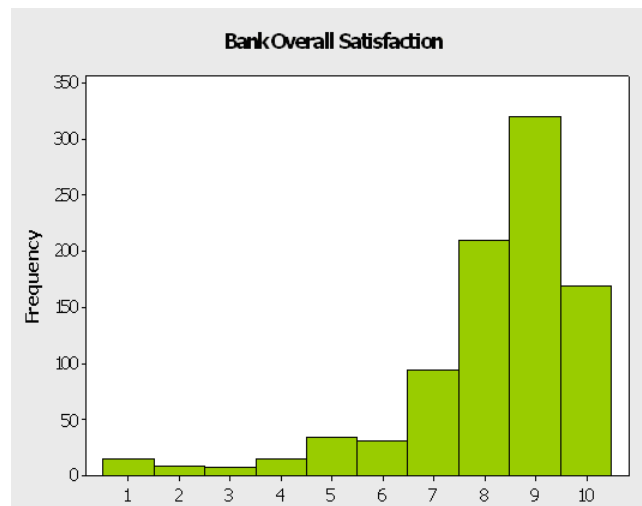


Figure 2
OSAT ratings (10-point Likert scale)

In table 1, the correlations between the predictors and OSAT are shown. Correlations of predictors with the response variable ranged from 0.67 (Personal Banker - PB) to 0.36 (Web). There are also positive correlations between the predictors themselves with the highest being between Teller and Drive-thru (0.60). For this analysis, no attempt was made to correct potential issues related to collinearity or deviations from multivariate normality.

	Osat	PB	Tell	Drive	Web
Osat	1.00				
PB	.67	1.00			
Teller	.60	.55	1.00		
Drive	.50	.39	.60	1.00	
Web	.36	.29	.30	.26	1.00

Table 1.
Correlations between predictors and OSAT

Using the “complete dataset” of 904 records, missingness was induced in such a way that the number of missing fields/records ranged from “mild” to “severe” as shown in Table 2. For the severe case, 29.1% of the fields had missing values and this affected 61% of the records. Thus only 39% of the records were unaffected by the missingness. A total of 400 datasets were generated with MAR records (100 per level). MAR data was created by disproportionably removing up to three channel scores when the scores for the personal banker were lower. Note: the personal banker and the OSAT variables are not eligible for inducing missingness.

	Field	Record
Mild	9.2%	16.1%
Moderate	14.2%	24.8%
Major	21.4%	37.6%
Severe	29.1%	61.0%

Table 2.
Levels of missingness in simulated data.

The design of an experiment which starts out with a complete data set and then artificially induces missingness to determine the impact of missing value techniques is similar to other studies in different settings (van der Heijden *et al.*, 2006).

7) EXPERIMENTAL METHOD

Much of this experimentation was conducted in R, since each of the 400 data sets need to be processed by the seven missing value techniques and then regression models estimated. For each of the 2800 runs, the regression coefficients and their standard errors were saved.

All missing value techniques were implemented in R except regression-based imputation, EM and MI. For the first two methods we used SPSS' MVA module; for MI we used SAS. The rationale for using other packages is that those options might be commonly implemented in market research shops through the use of SPSS and SAS.

For EM and MI, additional computer time was needed to either deal with convergence issues or to process the multiple data sets. For MI, ten imputed data sets were created for each of the 400 data sets defined in the experiment. Consequently, MI utilized a total of 4000 files to estimate the 100 regression results for each level of missingness.

8) EXPERIMENTAL RESULTS

Since the same general pattern of results can be seen across the different levels of missingness, we decided to show only those associated with the most extreme level: Severe. The readers can refer to the Sawtooth Conference PowerPoint Presentation for all levels of missingness as well as additional analysis such as: MCAR, Mice, Hmisc; etc.

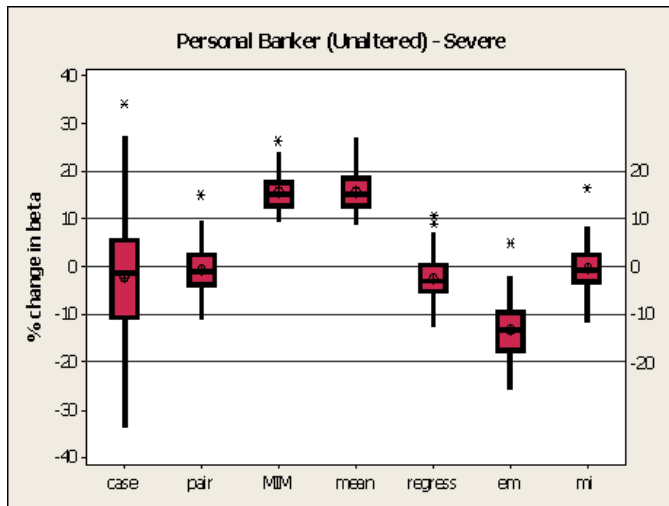


Figure 3. Percent change in beta – PB.

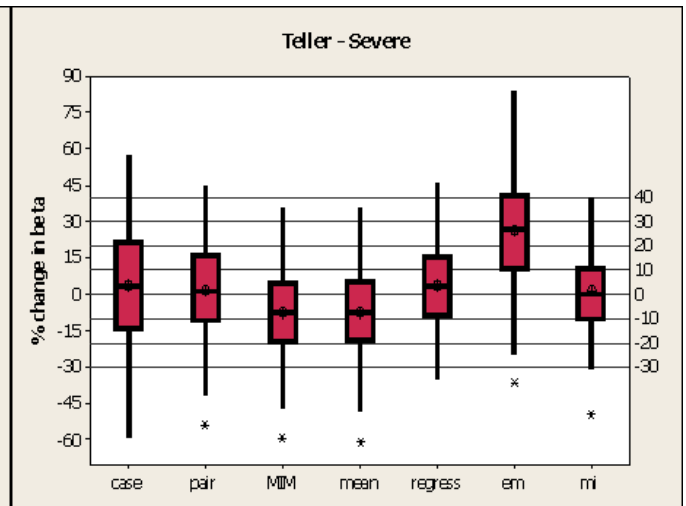


Figure 4. Percent change in beta – Teller.

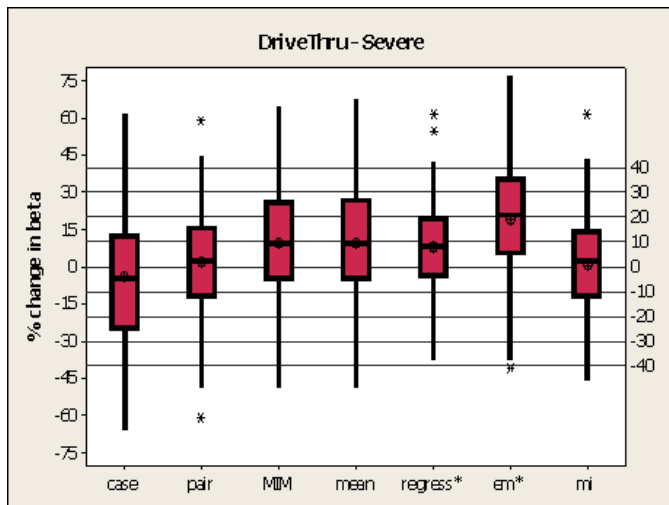


Figure 5. Percent change in beta – Drive-thru.

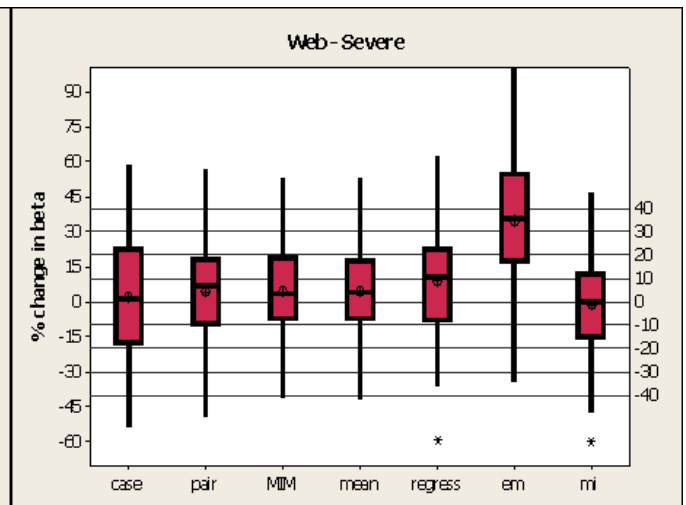


Figure 6. Percent change in beta – Web.

Figures 3 thru 6 illustrate how much the beta coefficients for each predictor were affected by each of the missing value techniques. The box plot is calculated by determining how much the beta coefficient from the imputed data set varied from the “known” value of beta. Under each missing value technique, the regression coefficients were calculated for the 100 versions of Severe missingness. Next, the percentage change in those coefficients from the “known” values of beta were calculated.

In Figure 3, the change in the beta coefficient for Personal Banker (PB) is shown. The box plot shows that for list-wise deletion that at least for one regression run, the beta coefficient was underestimated by about 35% and likewise there was another run in which beta was overestimated by 35%. The average change in percentage, as represented by the circle, is slightly above zero as well as the median, represented by the bar within the box. The box plot for list-wise deletion indicates that bias is not substantial but the precision of that estimate was varied.

Also, the PB plot indicates that there are bias issues with the MIM, overall mean substitution, EM and regression based imputation. The first two methods are not surprising but the latter two was cause for concern as better accuracy was expected. After investigating the issue a bit more, it turned out that the implementation of regression-based imputation and the creation of an EM imputed dataset are flawed in the MVA module for SPSS (von Hippel, 2004).

It is interesting to note that there were bias issues associated with the regression coefficients for the personal banker, as the personal banker had no induced missingness. However, the personal banker is obviously impacted by the missingness associated with the other predictors. In general, you will see a similar pattern of biasness with the other predictor variables. The MI approach produced box plots that were similar to that of pair-wise deletion. Both of these methods, produced a smaller range of coefficients than list-wise deletion.

The last issue to be examined is the standard error associated with the beta coefficient. This issue is important but it requires more simulation to estimate its “appropriateness.” In general, we want our estimates to be efficient but we also want them to represent the inherent uncertainty associated with either the missing value or the imputed value.

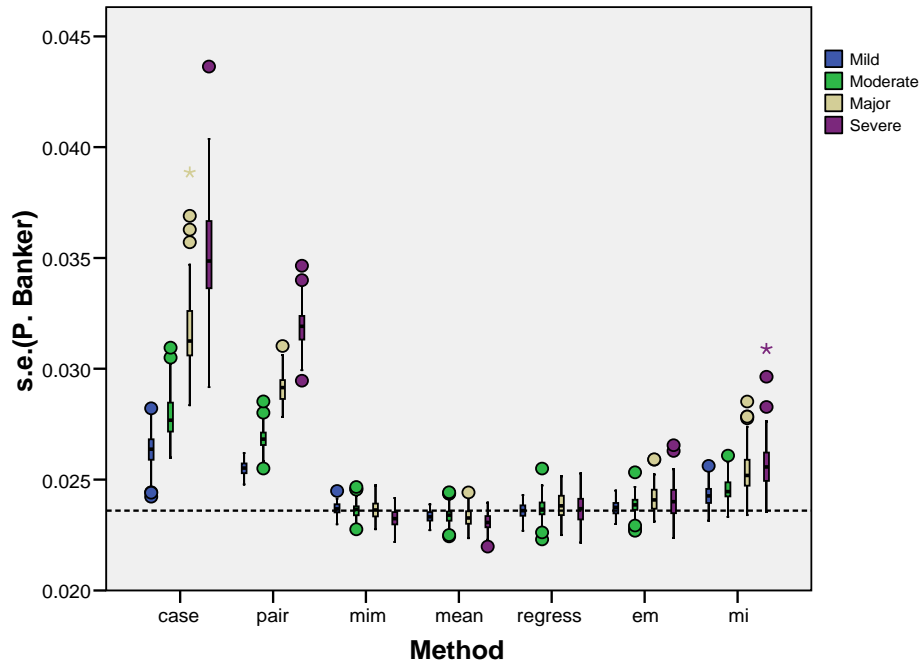


Figure 7.
Standard errors for PB.

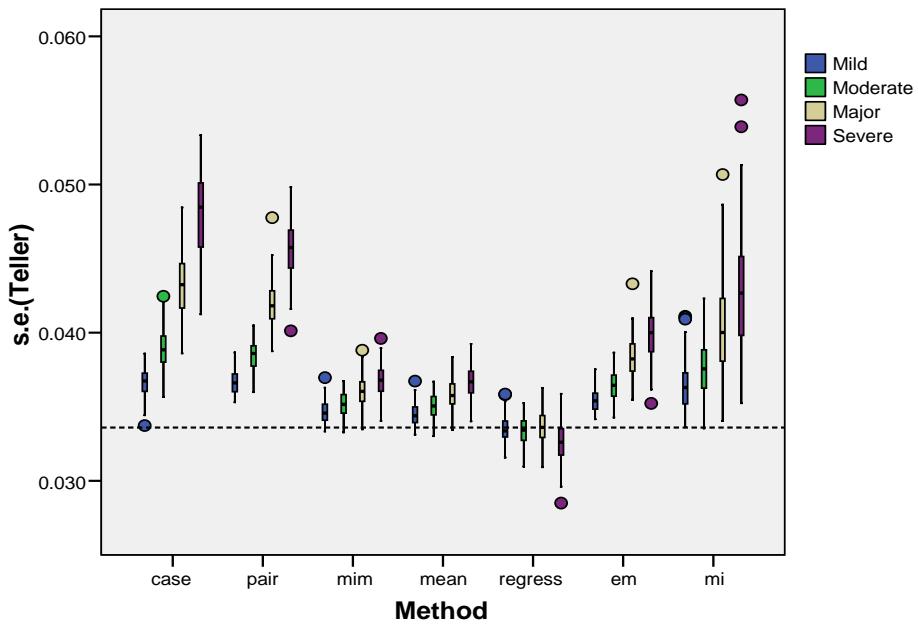


Figure 8.
Standard errors for Teller.

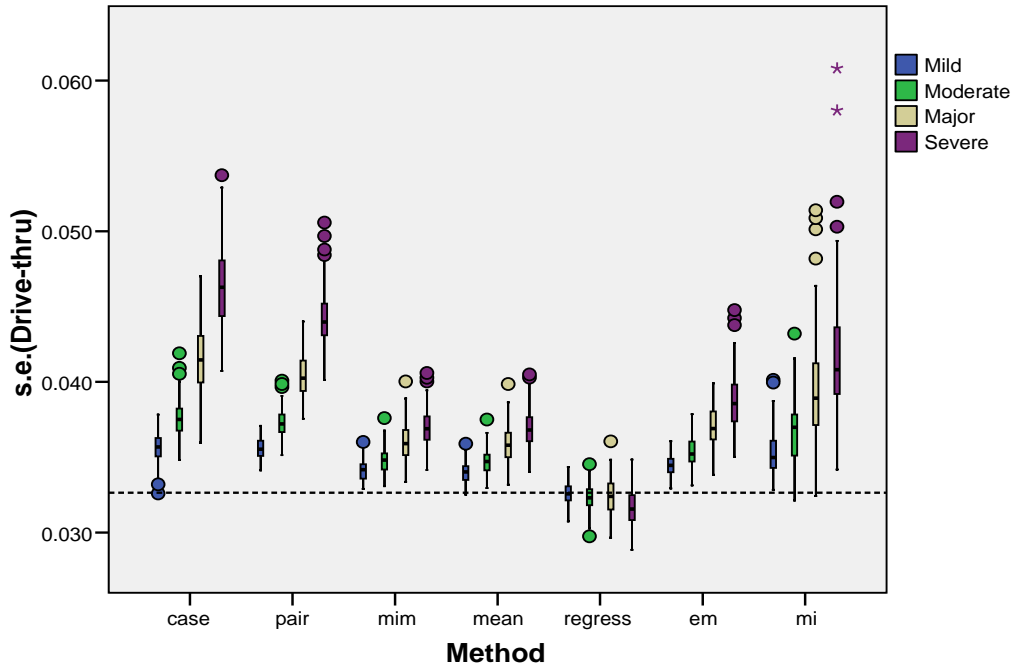


Figure 9.
Standard errors for Drive-thru.

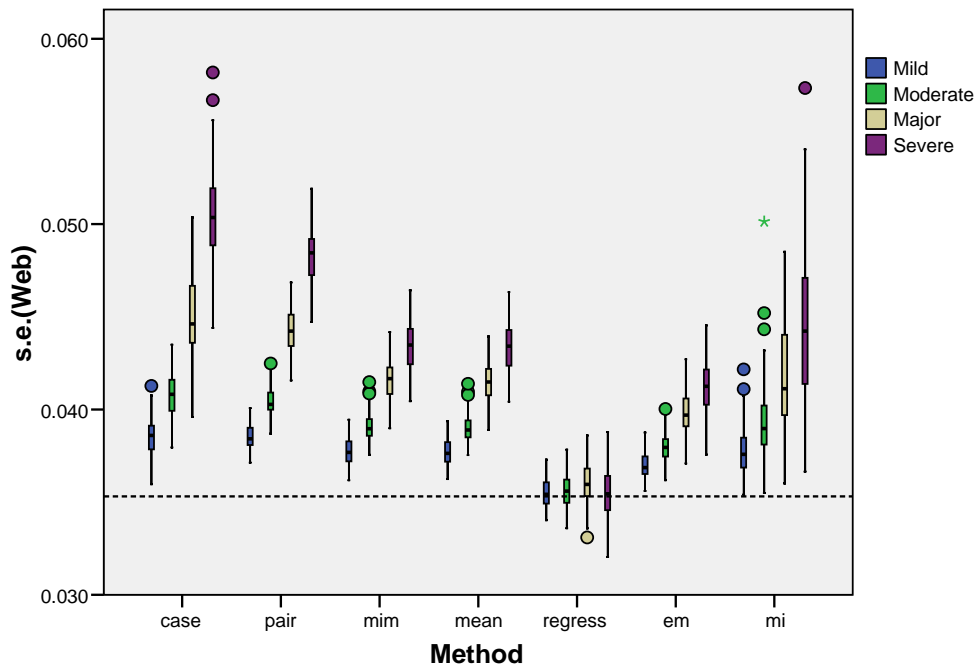


Figure 10.
Standard errors for Web.

The standard errors for each of the predictors are shown in Figures 7-10. The horizontal dashed line represents the “known” value; we would expect the standard errors to be at least above this line for predictors with missing values. That is not true for the regression-based imputation. For non-PB predictors, list-wise deletion and MI are resulting in the largest standard errors.

9) CONCLUSION

Missing values are often prevalent in many data sets. Care must be taken to understand the processes that govern missingness within your study. If it can be determined that the missingness is either MCAR or MAR, then an appropriate missing value technique can be chosen; proxy variables must exist for MAR. Clearly, there are some missing value techniques to almost always avoid: overall mean substitution, the missing indicator method, and those that are implemented in SPSS’ MVA module. Multiple imputation shows much promise and this method is increasingly being implemented in many packages including R, SAS and Amelia.

REFERENCES

- Allison, P. D. (2002). *Missing Data*, Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage.
- American Banker (2006). Banks at Satisfaction Ceiling?, February 21.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39, pp. 1-38.
- Donders, A.R., van der Heijden, G., Stijnen, T. and Moons, K. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, pp.1087-1091.
- Fishbein M. (1967). *Attitudes and Prediction of Behavior*, Readings in Attitude Theory and Measurement, New York: John Wiley and Sons, pp. 477-492.
- Harrell, F. A. (2001). *Regression Modeling Strategies*. New York: Springer-Verlag.
- Jones, M. P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*, 91 (433), pp. 222-230.
- Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New Jersey: John Wiley and Sons.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New Jersey: John Wiley and Sons.
- Shafer, J., & Olsen, M. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33 (4), pp. 545-571.
- van der Heijden, G., Donders, A.R., Stijnen, T. & Moons, K. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59, pp. 1102-1109.
- von Hippel, P.T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician*, 58 (2) pp. 161-164.

MAKING MAXDIFF MORE INFORMATIVE: STATISTICAL DATA FUSION BY WAY OF LATENT VARIABLE MODELING

LYND BACON

YOUgov/POLIMETRIX, INC.

PETER LENK

STEPHEN ROSS BUSINESS SCHOOL AT THE UNIVERSITY OF MICHIGAN

KATYA SERYAKOVA

KNOWLEDGE NETWORKS, INC.

ELLEN VECCIA

KNOWLEDGE NETWORKS, INC.

ABSTRACT

A major limitation of MaxDiff scaling or any discrete-choice conjoint methods such as choice-based conjoint (CBC) is the loss of a common origin across subjects. In these models, subjects' preferences are measured relative to a base option, which eliminates a common origin for making between-subjects comparisons. This assumption allows the ranking of options within a subject, but invalidates sorting subjects by the intensity of their preferences. We propose augmenting discrete-choice data with ratings data in order to recover the common origin. We fuse the two sources of information with a joint model that contains common parameters for the discrete-choice task and ratings scales. In particular, the partworths in the MaxDiff task enter the model for the ratings data, and the identification constraints are placed on the ratings model instead of the MaxDiff model. We demonstrate that the proposed method extends the range of applications for MaxDiff and CBC.

Acknowledgements: We thank John Wurst for his helpful and insightful comments on this article and our 2007 Sawtooth Software Conference presentation. We also acknowledge the support of Sawtooth Software.

1. INTRODUCTION

MaxDiff (AKA best/worst analysis) is a special kind of scaling methodology that has been steadily gaining popularity in recent years. Initially described by Finn & Louviere (1992), it exploits the ability of subjects to pick out extreme cases – the most and least preferred options – from sets of alternatives. The sets are based on an experimental plan allowing functions of alternatives' partworths to be estimated in the same manner as partworths are estimated using traditional or choice-based conjoint data. The procedure balances the effort required by the subject in the elicitation task with the amount of information provided by the task. The most informative, non-metric measurement method requires subjects to fully rank the options in each choice set. However, for more than around five options, full rankings are notoriously unreliable and taxing. At the other extreme, pick-best is the easiest for subjects, but the least informative for estimation. Best/worst effectively doubles the sample information over pick-best, while not requiring much additional effort from subjects. The majority of MaxDiff applications to date have been with atomic options,

such as selecting the most and least important attribute; however, MaxDiff can also be applied to composite options, such as product descriptions that consist of multiple attributes.

Since its introduction, this technique has been extended or enhanced in various ways, making it even more useful. The application of MCMC methods for estimating the parameters of Hierarchical Bayes models provides more accurate and informative score estimates at the individual level than the analysis method originally proposed. The development of algorithmic procedures for designing experiments has made it possible to employ experimental plans for MaxDiff choice sets that are good compromises between the number of choice sets and the precision of partworth estimates. Recent theoretical work has examined whether MaxDiff data are consistent with Random Utility Theory (Marley & Louviere, 2005). MaxDiff tasks have been shown to be a desirable means of generating results useful for market segmentation (Cohen, 2003; Cohen & Niera, 2003) as compared to other ways of collecting survey data.

A limitation of MaxDiff scaling, and choice-based conjoint (CBC) methods in general, is subjects' partworths are not measured on a common scale. Each survey-taker's scale scores are relative to an arbitrary origin, but that origin may or may not be the same for different survey-takers on the scale that is presumed to underlie their choices. The task does not capture sufficient information for estimating scores on a scale with the same origin for all survey-takers. Operationally, the preference for a base option or one of the intercepts for each subject is arbitrarily assumed to be zero in order to identify the model. This constraint shifts each subject's scale to zero at the base option, and the other partworths are measured relative to the base option. So, for example, if the scale is derived from "Most" and "Least" importance selection, one person's scale value of 2.5 for Option A means that this subject rates the option 2.5 above the base option. Another person's value of 2.5 for Option A only reflects the same relative distance from the base option, and does not imply that the two subjects would view Option A with the same absolute importance.

The loss of a common origin does not impede the application of MaxDiff to marketing applications where only within-subjects preferences are required. A common application is to use the estimated partworths from MaxDiff or discrete-choice as input to market share simulators. Since market share simulators are only concerned with the relative ranking of options within each subject, the loss of inter-subject comparability does not impact the derived market shares. However, the loss of a common origin does affect MaxDiff's usefulness in applications that require inter-subject comparisons, such as segmentation and targeting. Because the partworths are not measured on a common scale, it is not possible to sort subjects based on their preferences. Two subjects may give the same preference for Option A relative to the base option, but it is impossible to infer which subject prefers Options A the most.

Böckenholt (2004) considered this issue in the context of paired comparison choices, and proposed three methods of recovering a common origin. The researcher could a priori specify the common origins. Böckenholt gave the example of choosing among gambles with monetary winnings where the base option is the subject's current wealth. Using the subjects' actual current wealth, one is able to compare utilities for the gambles across subjects. This approach has limited application because it can only be used when the researcher is willing to make very strong assumptions about the origin.

His second approach is very creative and novel: it uses comparisons among bundles of options to recover the common origin. A hypothetical task would be to choose between an 80 GB video iPod™ or a 30 GB Zune™ digital media player bundled with a MP3 portable stereo player. A major, critical assumption of this approach is that the utilities are additive: the utility of the bundle is the sum of the utilities of its components. To see how bundles can resolve the origin, assume that U_1 , U_2 , and U_3 are the latent utilities for the iPod™, the Zune™, and the portable stereo player, respectively, and that the subject evaluates three choice tasks:

1. Choose between the iPod™ and the Zune™.
2. Choose between the iPod™ and the stereo player
3. Choose between the iPod™ and the bundle.

Task 1 provides information about $U_1 - U_2$; Task 2 provides information about $U_1 - U_3$; and Task 3 provides information about $U_1 - U_2 - U_3$, assuming additivity. Then the contrast Task 1 + Task 2 – Task 3 provides a pure estimate of U_1 . Without Task 3, one could only estimate the relative utilities $U_1 - U_2$ and $U_1 - U_3$.

This approach can be implemented with Sawtooth Software’s CBC/HB package if one is able to make an additional assumption about the error terms of the random utilities. The appropriate random utility model (RUM) would be the following:

$Y_1 = U_1 + \varepsilon_1$ is the random utility for the iPod™.

$Y_2 = U_2 + \varepsilon_2$ is the random utility for the Zune™.

$Y_3 = U_3 + \varepsilon_3$ is the random utility for the portable stereo player.

$Y_4 = U_2 + U_3 + \varepsilon_4$ is the random utility for the bundle.

Then the important assumption behind CBC/HB is that the error terms $\{\varepsilon_j\}$ for options in a choice set are a random sample from an extreme value distribution, which leads to the standard, logistic probabilities. A more natural assumption is to write the utility of the bundle as

$$Y_4 = U_2 + U_3 + \varepsilon_2 + \varepsilon_3,$$

which assumes that both utilities are additive in both their deterministic and random components. Conceptually, it may be a stretch to assume that the deterministic components U_2 and U_3 are additive in (4) but not the random components, as in (5). Nevertheless, the above specification (1)-(4) may provide a fruitful and practical method for identifying the origin with standard software. It’s worth noting that, aside from the strong assumption of additivity required, applications of this bundling approach are limited to those in which the alternatives being scaled can actually be combined, i.e. that aren’t mutually exclusive for some reason. Preferences for a single vacation destination, type of first job, case color for an MP3 player, or flavor of ice cream are examples of alternatives that can’t be bundled in a way that is likely to make sense to research subjects or users.

Böckenholt’s third proposal is to augment the MaxDiff or discrete-choice data with information collected on a continuous scale. For instance, combining importance ratings on

a 5 point Likert scale with the MaxDiff task. In general, we believe that this approach is superior to comparing bundles of options because it avoids the additive utility assumption for the bundle. If the bundles are not additive, then the procedure produces systematic bias in the estimates of the absolute utilities. On the other hand, fusing MaxDiff with ratings has its own challenges. First, to use ratings and discrete-choice scales to identify the origin, the model for the ratings must include the partworths used in the MaxDiff or CBC task. Second, the model for ratings data needs to accommodate well-known scale usage bias (Rossi, Gilula, and Allenby 2001) or else the imputed origin may only reflect scale usage. Finally, a large body of literature in psychometrics documents failure of procedure invariance in the elicitation of preferences (c.f. Slovic 1995). For example, Lichtenstein and Slovic (1971) demonstrated preference reversals for pairs of gambles depending when subjects are asked to price the gamble than choose the most preferred gamble. Grether and Plott (1979) systematically investigated a number of potential “rational” explanations for preference reversals and concluded that preference reversals arise from different psychological processes for the different tasks, despite their concerted attempts to prove the contrary. Consequently, any model that attempts to fuse ratings and discrete choice needs to be sufficiently flexible that the psychological processes do not distort the partworths in the discrete choice task. Here, we are making the value judgment that the discrete-choice task provides better external validity than ratings because customers in the market place choose products and do not rate them.

It is worth noting that a method that does fuse preference or importance responses obtained using different elicitation methods may provide more stable, and more generalizable results compared to any single elicitation technique as it is in a sense integrating over tasks that may each create its own method-specific bias. Also, when you use just a single task, you don’t have the opportunity to observe failure of procedural invariance. That doesn’t mean that it wouldn’t occur, of course.

The rest of the paper presents the model for fusing ratings and discrete-choice data in order to recover a common origin and demonstrates its utility in targeting subjects with two examples. In the next section, we review the underlying random utility models (RUM) for discrete-choice and MaxDiff. These models assume that the observed choices are driven by unobserved utilities for the options in a choice task. We then extend the basic RUM to fuse ratings and discrete-choice data in both the logit and probit specifications in Section 3. Section 4 reports the findings from a simulation study and demonstrates using the common origin for targeting subjects, and Section 5 concludes the paper.

2. RANDOM UTILITY MODELS FOR DISCRETE CHOICE AND MAXDIFF

2.1 Random Utility Essentials and the Loss of the Common Origin

Since McFadden’s (1974) seminal work on economic choice, random utility models (RUM) have provided the foundation for discrete-choice experiments. The concept is very simple: subject i has a latent, random utility Y_{ik} for option k . This utility is called “latent” because the researcher does not observe it directly. Instead, he or she can only observe its consequences, namely choices, best/worst, or rankings. The random utility is decomposed into a deterministic component U_{ik} and random component ε_{ik} : $Y_{ik}=U_{ik} + \varepsilon_{ik}$. If the random

components have an extreme-value or Gumbel distribution¹ McFadden (1974) showed that the choice probabilities are logistic functions of the partworths, which is the underlying assumption of Sawtooth Software's CBC/HB software. If the random components are normally distributed, then the choice probabilities are probit functions (Aitchison and Bennet 1970). The type of task determines the likelihood function that links the observed data U_{ijk} by integrating over the random component. This likelihood function varies for pick-best, full rankings, and MaxDiff due to different processes for using the random utilities to generate the response.

If the task is pick-best, then each subject is presented with J different choice tasks where each choice task has K options. Now we have three subscripts, i for subject, j for choice task, and k for option within choice task, and $Y_{ijk} = U_{ijk} + \varepsilon_{ijk}$ is the corresponding random utility. The behavioral assumption that makes pick-best work is that the subject chooses that option that corresponds to the maximum latent utility:

$$\text{Pick option } s \text{ if } Y_{ijs} \geq Y_{ijk} \text{ for all options } k \text{ in the choice set.} \quad (1)$$

The choice probabilities as functions of U_{ijk} are then computed using this inequality by integrating over the random terms.

The condition in (1) relates the observed choice to the preference structure, and focuses our attention on the loss of the common origin. The inequality in (1) holds for any linear rescaling of the latent utilities: nothing changes in the preference structure if subject i uses $Y_{ijk}^* = aY_{ijk} + b$ for arbitrary constants a and b where a is a positive. The researcher (or software package) uniquely identifies the latent utility by imposing constraints on the latent utilities. Most commonly, the scale parameter for the extreme value distribution (in logit models) or one of the variances for normal distribution (in probit models) is fixed to one. This constraint forces a to be one in Y_{ijk}^* . To eliminate arbitrary scale origins, one of the U_{ijk} is set to 0, say $U_{ij1} = 0$, which is the same as measuring the utilities of the other options with respect to option 1. If the options are composite options, $U_{ijk} = x_{ijk}'\beta_i$, where x_{ijk} is a vector describing attribute levels and β_i is a vector of partworths, then one of the intercepts is set to 0. Therein lies the crux of the problem: to estimate preferences uniquely from discrete choice data, the researcher loses the common scaling and inter-subject comparability.

To complete our review of random utility models, if the random terms are a random sample from right-skewed, extreme value (a.k.a. Gumbel) distributions with scale parameter 1, the choice probabilities are a logistic function of the $\{U_{ijk}\}$:

$$P_{ij}(s = \text{Most Preferred}) \equiv P_{ij}(s) = \frac{\exp(U_{ijs})}{\sum_{u=1}^K \exp(U_{iju})} \text{ for } s = 1, \dots, K. \quad (2)$$

Again, it is evident that one could add an arbitrary constant to each U_{ijk} without altering the choice probability unless an identifying constraint is enforced.

¹ The cumulative distribution function is $F(\varepsilon) = \exp[-\exp(-\varepsilon)]$ for the right-skewed, extreme value distribution.

2.2 Random Utility Models for MaxDiff

MaxDiff, originally proposed by Louviere and Woodworth (1990) and published by Finn and Louviere (1992) applies the basic RUM to best/worst responses. Their model assumes that subjects evaluate the difference in utility for every pair of options and selects the difference with maximal utility. The random utility for each ordered pair (s,t) in choice task j is:

$$Y_{ij,st} = U_{ijs} - U_{ijt} + \varepsilon_{ij,st} \text{ for } s, t = 1, \dots, K \text{ and } s \neq t.$$

Note that $Y_{ij,st}$ is not equal to $Y_{ij,ts}$. This MaxDiff model assumes that the option with the maximal differences is selected:

Option s is best and option t is worst if $Y_{ij,st} > Y_{ij,uv}$ for all other pairs (u,v).

Assuming extreme value distributions for the random terms, the MaxDiff probabilities for subject i and choice task j are:

$$\begin{aligned} P_{ij}(s = \text{Most Preferred}, t = \text{Least Preferred}) &\equiv P_{ij}(s, t) \\ &= \frac{\exp(U_{ijs} - U_{ijt})}{\sum_{u=1}^K \sum_{v=1: v \neq u}^K \exp(U_{iju} - U_{ijv})} \text{ for } s \neq t \end{aligned} \quad (3)$$

Finn and Louviere (1992) and Flynn, Louviere, Peters, and Coast (2007) apply MaxDiff to atomic options for aggregate attitudes for food safety and quality of life. Of course, MaxDiff could also be used with composite options, in which case the choice probabilities are:

$$\begin{aligned} P_{ij}(s = \text{Most Preferred}, t = \text{Least Preferred}) &\equiv P_{ij}(s, t) \\ &= \frac{\exp\left([x_{ijs} - x_{ijt}] \beta_i\right)}{\sum_{u=1}^K \sum_{v=1: v \neq u}^K \exp\left([x_{iju} - x_{ijv}] \beta_i\right)} \text{ for } s \neq t \end{aligned} \quad (3')$$

Once again, the MaxDiff model is identified by assuming that one of the utilities (or intercepts for composite products) is zero.

Although the choice probabilities in (3) and (3') seem complex, this formation for MaxDiff can easily be estimated in Sawtooth Software's CBC/HB package if the number of options K in a choice task are not too large. For example, suppose that a brand study is performed with 5 brands and nominal price. To identify the model, the partworth for brand 5 is assumed to be zero. Each choice task consists of three options (K=3). In Figure 1, the choice set uses brands B1, B3, and B4, with prices \$5, \$4, and \$7, respectively. The left-side of Figure 1 gives the design matrix for this choice task where the first 4 columns identify the brand (Brand 5 is the base brand), and the last column is price. The three rows represent the three options in the choice set. The right-side of Figure 1 gives the corresponding .cho file to implement MaxDiff in CBC/HB. At the top "6 1" means that there are 6 possible choices, corresponding to the different ordered pairs of B1, B2, and B3, and 1 choice is made. Taking

all possible pairwise differences of the rows of the matrix on the left-hand-side, where we added “B1-B3” etc. to indicate which brands are in the pairwise differences, forms the cho matrix on the right-hand-side. In this example, we assumed that the subject picked B3 as the best and B1 as the worst, which corresponds to the difference “B3-B1” on the right-hand-side. Consequently, the “3 99” at the bottom indicates that row 3 was selected, and 99 is a stop-code.

Figure 1:
An example of a Sawtooth Software cho matrix for the original formulation of MaxDiff where B3 is the best option and B1 is the worst option.

$$\begin{array}{l}
 \text{B1} \rightarrow \\
 \text{B3} \rightarrow \\
 \text{B4} \rightarrow
 \end{array}
 \begin{bmatrix}
 1 & 0 & 0 & 0 & 5 \\
 0 & 0 & 1 & 0 & 4 \\
 0 & 0 & 0 & 1 & 7
 \end{bmatrix}
 \Rightarrow
 \begin{array}{l}
 \begin{bmatrix}
 1 & 0 & -1 & 0 & 5-4 \\
 1 & 0 & 0 & -1 & 5-7 \\
 -1 & 0 & 1 & 0 & 4-5 \\
 0 & 0 & 1 & -1 & 4-7 \\
 -1 & 0 & 0 & 1 & 7-5 \\
 0 & 0 & -1 & 1 & 7-4
 \end{bmatrix}
 \begin{array}{l}
 \leftarrow \text{B1 - B3} \\
 \leftarrow \text{B1 - B4} \\
 \leftarrow \text{B3 - B1} \\
 \leftarrow \text{B3 - B4} \\
 \leftarrow \text{B4 - B1} \\
 \leftarrow \text{B4 - B3}
 \end{array}
 \end{array}
 \begin{array}{l}
 6 \ 1 \\
 3 \ 99
 \end{array}$$

A mathematically equivalent expression for the MaxDiff choice probabilities in Equations (3) or (3') is:

$$P_{ij}(s, t) \propto P_{ij}(s)Q_{ij}(t) \text{ for } s, t = 1, \dots, K \text{ and } s \neq t. \quad (4)$$

where $P_{ij}(s)$ is the probability of selecting the best from Equation (2) and

$$Q_{ij}(t) = \frac{\exp(-x'_{ijt}\beta_i)}{\sum_{k=1}^K \exp(-x'_{ijk}\beta_i)} \text{ for } t = 1, \dots, K. \quad (5)$$

Note, in particular, the proportionality sign in Equation (4).

A Sawtooth Software (2005) technical report derives the probabilities $Q_{ij}(t)$ in Equation (5) from RUM for the least preferred option where

$$Y_{ijk}^* = x'_{ijk}\beta_i + \varepsilon_{ijk}^* \text{ for } k = 1, \dots, K$$

and the random terms $\{\varepsilon_{ijk}^*\}$ are a random sample from a *left-skewed*, extreme value distribution with scale parameter one². Then option t is least preferred if $Y_{ijt}^* < Y_{ijk}^*$ for all $k \neq t$, and the choice probability is given in Equation (4). Cohen (2003) and Cohen and Orme (2004) use a simple method for approximating the MaxDiff probabilities in Equation (3) by relaxing the constraint that the most and least preferred options have to be different:

² The cumulative distribution function is $F(\varepsilon) = \exp[-\exp(\varepsilon)]$ for the left-skewed, extreme value distribution.

$$P_{ij}(s,t) = P_{ij}(s)Q_{ij}(t) \text{ for } s, t = 1, \dots, K. \quad (5')$$

The proportionality sign in Equation (4) has been replaced by an equality sign in Equation (5'). In Equation (5'), the best and worst options are mutually independent, while they are dependent in Equation (4). Behaviorally, this formulation is different from the original MaxDiff, which assumed that the subject selected the maximum pairwise difference. This specification, which was also used by Finn & Louviere (1992) for aggregate data, corresponds to a two-stage behavioral model where the subject evaluates all of the options using right-skewed extreme value distributions and selects the best. Then he or she reevaluates all of the options with left-skewed extreme value distributions and selects the worst.

The Sawtooth Software (2005) technical report shows how to estimate the model used by Cohen and Orme using Sawtooth Software's CBC/HB and by treating the single best/worst responses as two responses in standard CBC. The best response is coded as usual with design matrix X_{ij} for subject i and choice set j . The worst response is coded with design matrix $-X_{ij}$. Figure 2 gives the cho matrix for the best/worst data in Figure 1. The top "3 1" indicates that there are 3 options in the choice set, and one option was selected. The first matrix is the standard design matrix for choice-based conjoint. The "2 99" indicates that the second row (Brand 3) was selected as "best," and "99" is a stop-code. The second "3 1" indicates, as before, three options in the choice task with one choice only. The second matrix is the coding for the "worst" choice task, which is merely the negative of the top matrix. The "1 99" indicates that the first option was selected as worst, and "99" is the stop code.

Figure 2:
The Sawtooth Software cho matrix for the Cohen/Orme best/worst model specification using the data in Figure 1.

$$\begin{array}{c}
 3 \ 1 \\
 \begin{bmatrix} 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 7 \end{bmatrix} \\
 2 \ 99 \\
 3 \ 1 \\
 \begin{bmatrix} -1 & 0 & 0 & 0 & -5 \\ 0 & 0 & -1 & 0 & -4 \\ 0 & 0 & 0 & -1 & -7 \end{bmatrix} \\
 1 \ 99
 \end{array}$$

We propose an alternative formulation for best/worst data that follows directly from the basic RUM in Section 2.1. If s is the best alternative and t is the worst alternative, then

$$Y_{ijs} > Y_{ijk} > Y_{ijt} \text{ for } k \neq s \text{ or } t, \text{ and } k = 1, \dots, K. \quad (6)$$

Here, we assume that the random terms are either a random sample from a right-skewed distribution (standard logit model) or a normal distribution (probit model). The behavioral assumption is that the subject evaluates the latent utility for each option and selects the options with the maximum and minimum utilities. The reader may wonder why other authors have not thought of (6). The sad news is that the specification in (6) does not lead to a tidy expression for the choice probabilities. However, a close inspection of Equation (6) reveals that best/worst responses correspond to partially ranked data. Our approach to the problem is to impute the missing ranks and use the exploding logit model of Beggs, Cardell and Hausman (1981) and Chapman and Staelin (1982).

In particular, for subject i and choice task j , let R_{ij1}, \dots, R_{ijK} be the ranks of the options where R_{ij1} is the index of the least preferred option and R_{ijK} is the index of the most preferred options. If all of the options in the choice task were ranked ordered by the subject, then $Y_{ijR1} < Y_{ijR2} < \dots < Y_{ijRK}$. The exploding logit model for fully ranked data is:

$$P(R_{ij1}, \dots, R_{ijK}) = \prod_{s=2}^K \frac{\exp(U_{ijR_s})}{\sum_{k=1}^s \exp(U_{ijR_k})} \text{ for } R_{ij1} < \dots < R_{ijK} \quad (7)$$

The marginal distribution for the best and worst options can be obtained from Equation (7) by summing over the intervening ranks. Since these models are estimated by using MCMC, an alternative to direct computation is to impute the missing ranks in the MCMC algorithm by the following procedure. Given all of the parameter estimates, generate the latent utilities Y_{ijk} for k not equal to R_{ij1} or R_{ijK} :

$$Y_{ijk} = U_{ijk} - \ln[-\ln(u)] \text{ where } u \text{ is uniform on } [0,1] \text{ for } k \neq R_{ij1} \text{ or } R_{ijK}.$$

Then the missing ranks are based on these imputed latent variables. For the probit formulation, one merely adds an additional constraint: the lower inequality in Equation (6), to the standard MCMC algorithm of Albert and Chib (1993) and McCulloch and Rossi (1994).

2.3 Simulation Study Comparing MaxDiff Methods

At this point, a natural question is if these three procedures have any practical differences. Their underlying behavioral models start with RUM. However, the processes of using the random utilities to select the best/worst alternatives are qualitatively different, and the resulting likelihood functions are different. We use a short simulation study to see if there are quantitative differences in the results. The study simulates 100 subjects who evaluate 26 to 36 choice tasks. Each choice task has 4, full profile options. Each profile corresponds to a brand (4 brands). The profiles also include price X_1 (continuous scale) and a 0/1 dummy X_2 for advertising. The individual partworths $\{\beta_i\}$ are related to subject-level demographics – $\ln(\text{income})$ and household size – through a multivariate regression model. Table 1 compares the true values of $\{\beta_i\}$ to their estimates using the original MaxDiff (Equations 3 and 3'), the Finn & Louviere / Cohen & Orme specification "LR Skew" (Equation 5), and the rank imputed exploding logit "RIMEX" (Equations 6 and 7). Based on this simulation's result there is nothing that distinguishes one approach over the other. Figure

3 provides a graphical display of Table 1 by plotting the true and estimated partworths for the three procedures.

Table 1.

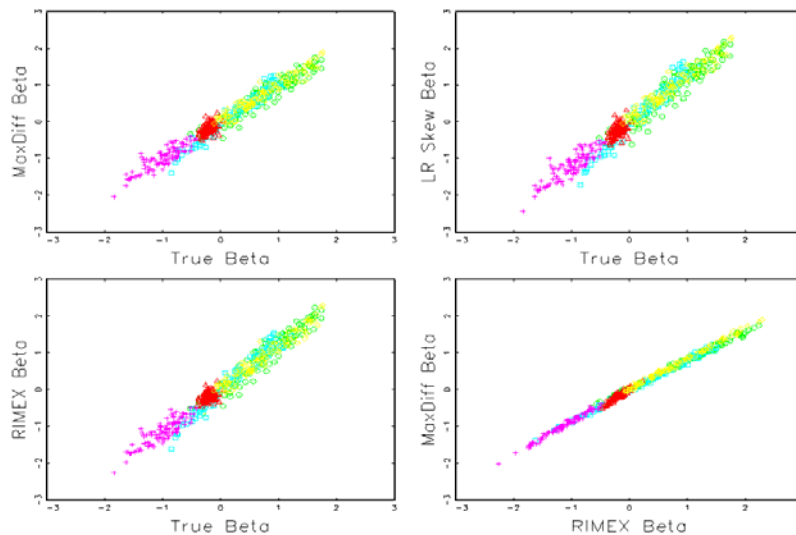
Simulation results for comparing the estimated partworths of the original MaxDiff, procedure of Cohen/ Orme (“LR Skew”), and the rank imputed exploding logit (“RIMEX”).

Correlation	Brand 1	Brand 2	Brand 3	X1	X2
True vs RIMEX	0.965	0.979	0.386	0.876	0.967
True vs MaxDiff	0.955	0.978	0.433	0.866	0.966
True vs LR Skew	0.952	0.978	0.409	0.865	0.967
RIMEX vs MaxDiff	0.997	0.997	0.937	0.989	0.997
RIMEX vs LR Skew	0.997	0.997	0.944	0.991	0.997
MaxDiff vs LR Skew	0.999	1.000	0.994	0.997	1.000
RMSE	Brand 1	Brand 2	Brand 3	X1	X2
True vs RIMEX	0.284	0.282	0.119	0.224	0.219
True vs MaxDiff	0.182	0.172	0.125	0.161	0.129
True vs LR Skew	0.294	0.323	0.153	0.272	0.249
RIMEX vs MaxDiff	0.262	0.172	0.049	0.164	0.172
RIMEX vs LR Skew	0.065	0.089	0.075	0.073	0.062
MaxDiff vs LR Skew	0.244	0.189	0.066	0.210	0.204

RMSE is root mean squared error.

Figure 3.

True and estimated partworths from the simulation study.



We also ran other simulations that varied the number of subjects and number of choice tasks per subject and obtained similar conclusions that the estimates from the three approaches did not systematically differ, despite our concerted effort to show the contrary. Our results are not bad news for practitioners, as they suggest that current MaxDiff modeling practices provide reasonably good results, *ceteris paribus*.

3. FUSING MAXDIFF WITH RATINGS TO IDENTIFY THE ORIGIN

This section augments the best/worst or discrete-choice data with ratings data to identify the origin for measuring partworths. In describing a model to fuse discrete-choice and ratings data, we assume that the partworths from the discrete-choice model are the focal set of parameters. Moreover, we do not want the psychological processes used for ratings to systematically distort the preference structure from the choice task. We assume that the discrete-choice task has more external validity than the ratings task. Our main purpose of fusing the choices and ratings is to identify the common scale for inter-subject comparisons. Other objectives, such as better estimates, are secondary concerns and not investigated here. However, we need a model that uses the ratings data to recover the origin without contaminating the discrete-choice partworths. A more detailed description of the model is given in Lenk and Bacon (2007), where it is noted that the data fused with choice data need not be ratings data. They could be behavioral measures based on purchase histories or web site visits, for example. Our model specification is related to that in Vriens, Oppewal, and Wedel (1998) who fused a ratings conjoint task with importance ratings to obtain better partworth estimates.

We use a threshold model (Aitchison and Silvery 1957 and Rossi, Gilula, and Allenby 2001) to relate the ratings data to a latent variable, and then we specify a joint model for the latent variables used in the ratings and discrete-choice tasks. The threshold model converts the ordinal responses on a rating scale to a continuous scale. It assumes that the observed ordinal responses arise due to the continuous latent value falling in regions determined by ordered cutpoints C_1 to C_H for an H point scale. The model for the latent variables is:

$$W_{im} = \varphi_i + \alpha_m + v'_{im} \Psi \beta_i + \xi_{im} \quad (8)$$

where

1. φ_i is a random effect for subject i . The random effects are random sample from a normal distribution with mean 0 and standard deviation τ . These parameters help ameliorate scale usage effects.
2. α_m is a parameter for item m that adjusts the ratings model for compatibility effects.³
3. v_{im} is the design vector for item m and is observed. For atomic options, e.g. importance ratings, it is a vector of zeros and a single one.
4. Ψ is a square, diagonal matrix with zeros on the off-diagonals and positive entries on the diagonals. Ψ adjusts the partworths for prominence effects⁴ when going from discrete-choice to ratings tasks. Its elements allow making inferences about the appropriateness of fusing the choice and ratings data.
5. β_i are the partworths from the discrete-choice or MaxDiff task.
6. ξ_{im} is a random term, which is either right-skewed extreme value for the logit model or normally distributed for the probit model. The scale parameters (logit) or error variances (probit) depend on the item m .

³ Compatibility effects occur when a stimulus is more compatible with a particular elicitation task. See Tversky, Sattath, and Slovic (1988).

⁴ Different attributes become more or less prominent depending on the task. See Nowlis and Simonson (1997).

The probabilities for the item responses are:

$$P_{im}(1) = F\left[\frac{C_1 - (\varphi_i + \alpha_m + v'_{im} \Psi \beta_i)}{\zeta_m}\right]$$

$$P_{im}(h) = F\left[\frac{C_h - (\varphi_i + \alpha_m + v'_{im} \Psi \beta_i)}{\zeta_m}\right] - F\left[\frac{C_{h-1} - (\varphi_i + \alpha_m + v'_{im} \Psi \beta_i)}{\zeta_m}\right].$$

for $h = 2, \dots, H-1$

$$P_{im}(H) = 1 - F\left[\frac{C_H - (\varphi_i + \alpha_m + v'_{im} \Psi \beta_i)}{\zeta_m}\right]$$

where $C_1 < \dots < C_H$ are the cutpoints. In the logit model, F is the cumulative distribution function for the right-skewed, extreme value distribution in footnote 4, and ζ_m is the scale parameter for item m . In the probit model, F is the cumulative distribution function of the standard normal distribution, and ζ_m is the error standard deviation for item m . This model for the ratings includes the partworths from the choice task in an indirect fashion to accommodate potential psychological biases when going from discrete-choice tasks to rating tasks. When these biases are absent, then α_m is zero and the Ψ is the identity matrix.

The RUM for the discrete-choice or Max/Diff task is:

$$Y_{ijk} = x'_{ijk} \beta_i + \varepsilon_{ijk}$$

$$\beta_i = \Theta' z_i + \delta_i$$

where

1. Y_{ijk} is subject i 's latent utility for profile k in choice task j
2. x_{ijk} is the design vector for profile k .
3. ε_{ijk} is the random term either from a right-skewed, extreme value distribution (logit model) or multivariate normal distribution (probit model).
4. z_i is a subject-level covariate; Θ is a matrix of regression coefficients, and δ_i is a multivariate normal error term with mean 0 and covariance Λ .

We use standard priors for the parameters (See Lenk and Bacon 2007). For pick best data, we assume the constraint (1). For best/worst data, we impose the constraint in Equation (6). The model could be easily modified for the original MaxDiff formulation or the Cohen/Orme approach.

The key to recovering the common origin is moving the standard identification constraints from the choice task to the ratings model. We do not constrain the partworths $\{\beta_i\}$, and we do not assume that the scale parameter (logit) or a variance (probit) for the random term $\{\varepsilon_{ijk}\}$ is one. Instead, we assume that one of the $\{\alpha_m\}$ is zero, and one of the diagonal elements of Ψ is one.

4. APPLICATIONS

Lenk and Bacon (2007) present simulation results that indicate the model in Section 4 is identified, and the estimated parameters are close to their true values. Here, we present two applications that illustrate the practical benefits retaining a common origin for subjects' utility scales in order to compare subjects on their preference structures.

4.1 Educational Goals

A study on the importance of eight educational goals had 1470 subjects both rate the importance for each goal, and then perform a best/worst task consisting of 8 choice sets with three goals per choice set. Without ratings data, we identify the model by assuming that the utility for Goal 8 is zero. We fitted the model with just the best/worst data and with both the best/worst and ratings data. Also, we fitted both logit and probit models, which gave similar results. We will report the results for the probit model. We will not report all of the estimates of the parameter in Section 4, and we will only focus on the estimated partworths.

Figure 4 plots the estimated partworths for the two models: the x-axis is for the standard MaxDiff model without ratings and the y-axis is our proposed model. The graph shows that with a common origin, subjects are distinguished by the absolute utility evaluations. We emphasize this point by segmenting the subjects into three tiers. The top tier consists of subjects with three or more partworths that are greater than 6, and the bottom tier consists of subjects with three or more partworths less than 1.5. There are approximately 20% of the subjects in each of the top and bottom tiers. The top tier represents subjects who are highly concerned with education, while the bottom tier is not. Figure 5 presents boxplots of the partworths for the three tiers with ratings and without ratings. The boxplots show the partworths increasing through the tiers when they are estimates with the ratings, but they are flat when estimated without ratings.

Figure 4.
Estimated partworths for educational goals with and without ratings.

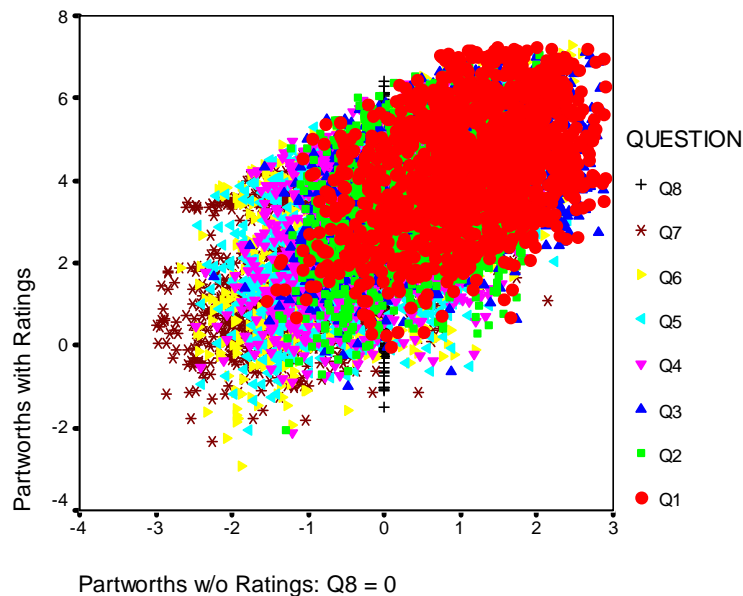
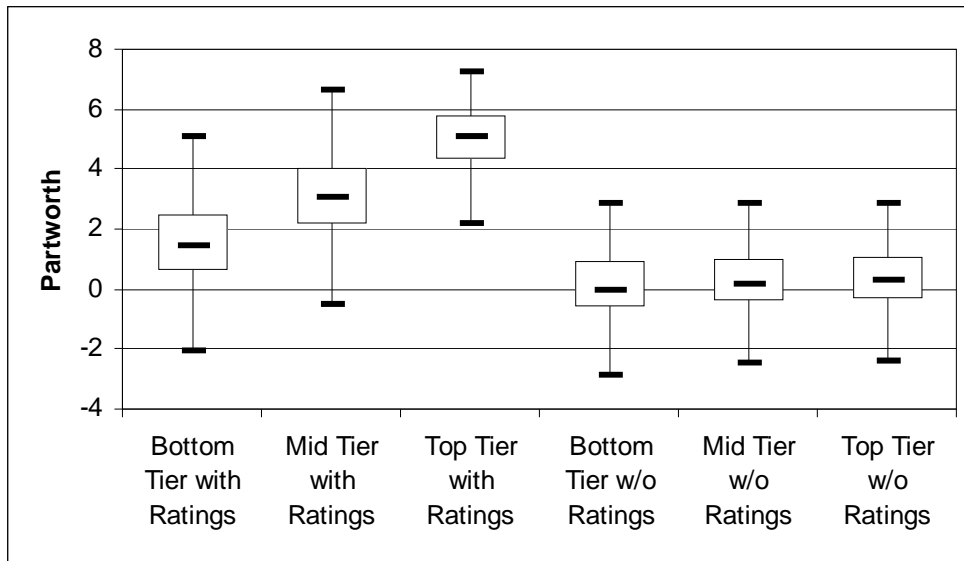


Figure 5
 Boxplots of partworths segmented by tiers with and without ratings.



4.2 PC Design

The second application uses data from a conjoint study of personal computers. The participants were 210 MBA students at a major university. These subjects rated 20 profiles on a 0 to 10 Likert scale for the likelihood of purchase, and for each profile, they also indicated Buy/No Buy. The data are taken from Lenk *et al.* (1996). We use a threshold model for the discrete choice task where subject i selects “Buy” for profile j if his or her random utility $Y_{ij} > \beta_{iT}$. The threshold can be interpreted as the utility of the “outside good.” For example, if the subject already owns a PC, then β_{iT} can be interpreted as the utility for that PC. In standard binary choice models without ratings, the utility of the outside good is assumed to be 0. In our fusion model, we are able to separately estimate both the utility of the outside good and the intercept, which is the utility of the “average” PC in the study since we use effects coding for the binary attributes.

Figure 6 displays boxplots of the partworth heterogeneity for the models with and without the importance rating data. With importance rating (right-hand panel), the partworth for the outside good (the threshold) and the “average PC” (the intercept) are separately estimated, while they are confounded for the model without ratings. Figure 7 sorts subjects according to their utility of the outside good and superimposed the price partworth. The correlation between the partworths for price and the outside good is -0.376 . The plot shows that the subjects with smaller utilities of outside goods also tend to be less price sensitive, and conversely. This implies that subjects who own a fairly good PC require an outstanding deal to buy a new PC, while subjects with low utility for the outside good are willing to pay more.

Figure 6.
 PC design study: boxplots of the heterogeneity in the estimated partworths .
 The estimates without ratings is the left-hand panel, and the model with ratings is the right-hand panel.

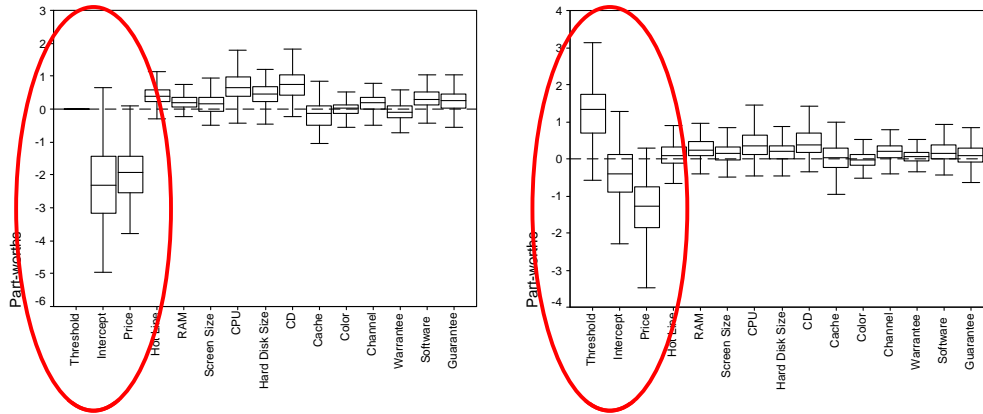
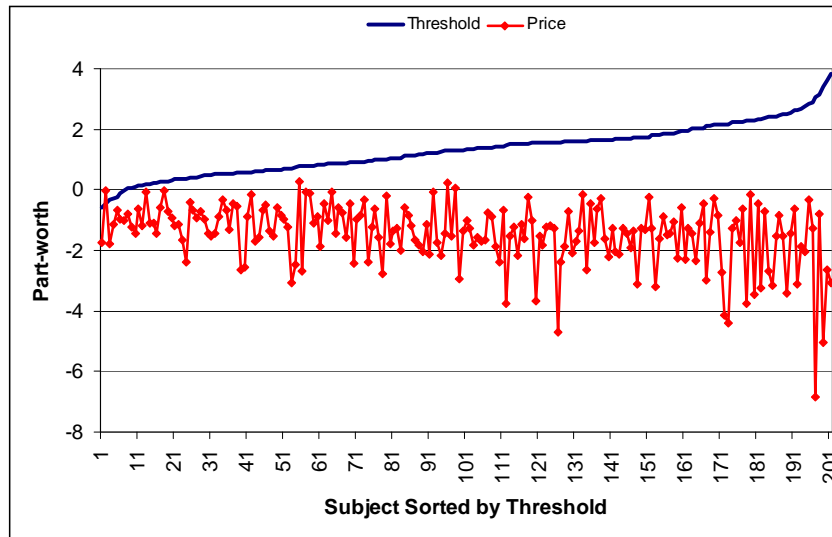


Figure 7.
 PC design study: subjects are sorted by their utilities for the outside good.



5. CONCLUSION

The purpose of this paper is twofold: to make marketing researchers aware of an overlooked limitation of choice-based models, the loss of a common origin, and to provide a remedy by augmenting choice data with ratings data. Without a common origin inter-subject comparisons are not meaningful. Applications such as market share simulations, which only rely on the relative partworths within each subject, are still valid without a common origin. However, applications such as segmentation and targeting that rely on comparisons of preference structures among subjects are misleading unless partworths are measured on a common scale. We propose augmenting choice-based data with auxiliary data, which in the present case are ratings data, to recover a common scale. Our fusion model accommodates

psychological biases inherent in ratings, as compared to choices, and scale usage biases for ratings. In two datasets we demonstrate that recovering a common origin increases the utility of choice-based experiments.

REFERENCES

- Albert J H and Chib S (1993). "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88 (422), Jun, 669-679.
- Aitchison J and Bennet JA (1970). "Polychotomous quantal response by maximum indicant," *Biometrika*, 57, 253-262.
- Aitchison J and Silvery SD (1957). "The generalization of probit analysis to the case of multiple responses," *Biometrika*, 44, June, 131-150.
- Beggs S, Cardell S and Hausman J (1981). "Assessing the potential demand for electric cars," *Journal of Econometrics*, 17, 1-20.
- Böchenholt, U. (2004). "Comparative judgments as an alternative to ratings: Identifying the scale origin," *Psychological Methods*, 9 (4), 453-465.
- Chapman, R G and Staelin R (1982). "Exploiting rank ordered choice set data within the stochastic utility model," *Journal of Marketing Research*, 19 (3), Aug, 288-301.
- Cohen, Steve (2003). "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," *2003 Sawtooth Software Conference Proceedings*, Sequim, WA.
- Cohen, S. & Orme, B. (2004). "What's your preference?" *Marketing Research Magazine*, Summer, 32-37.
- Cohen, S. & Neira, L. (2003). "Measuring preference for product benefits across countries: Overcoming scale usage bias with maximum difference scaling." Paper presented at the Latin American conference of the European society for opinion and marketing research, Punta del Este, Uruguay (pp. 1-22). Reprinted in *Excellence in International Research: 2004*. ESOMAR, Amsterdam, Netherlands, 2004.
- Finn A and Louviere J J (1992). "Determining the appropriate response to evidence of public concern: The case of food safety," *Journal of Public Policy & Marketing*, 11 (1), 12-25.
- Flynn TN, Louviere J J, Peters T J, and Coast J (2007). "Best-worst scaling: What it can do for health care research and how to do it," *Journal of Health Economics*, 26, 171-189.
- Grether, D. M. and Plott, C. (1979). "Economic theory of choice and the preference reversal phenomena," *American Economic Review*, 69, 623-638.
- Lenk, P and Bacon, L (2007). "Estimating Common Utility Origins and Scales in Discrete-Choice Conjoint with Auxiliary Data," working paper, The University of Michigan.
- Lenk, P., DeSarbo, W., Green, P., and Young, M. (1996). "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173-191.

- Lichtenstein, S. and Slovic, P. (1971). "Reversal of preferences between bids and choices in gambling decisions," *Journal of Experimental Psychology*, 89, 46-55.
- Louviere J and Woodworth G (1990). "Best-Worst scaling: A model for largest differences judgements," working paper, Faculty of Business, University of Alberta.
- Marley, A. A. J. and Louvier, J. J. (2005). "Some probabilistic models of best, worst, and best-worst choices," *Journal of Mathematical Psychology*, 49, 464-480.
- McCulloch, R and Rossi P E (1994). "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64 (1-2), Sept-Oct, 207-240.
- McFadden D (1974). "Conditional logit analysis of quantitative choice behavior," *Frontiers in Econometrics*, ed by P Zarembka, Academic Press, New York, 105-142.
- Nowlis, S. M. and Simonson, I. (1997). "Attribute-task compatibility as a determinant of consumer preference reversals," *Journal of Marketing Research*, 34 (2), May, 205-218.
- Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach," *Journal of the American Statistical Association*, 96 (453), Mar, 20-31.
- Slovic, P (1995). "The Construction of Preference," *American Psychologist*, 50 (5), May, 364-371.
- Sawtooth Software (2005). *The MaxDiff/Web System Technical Paper*, Sawtooth Software, Inc.
- Tversky, A., Sattath, S., and Slovic, P. (1988). "Contingent weighting in judgment and choice," *Psychological Review*, 95 (3), July, 371-384.
- Vriens, M., Oppewal, H., and Wedel, M. (1998). "Ratings based versus choice-based latent class conjoint models - An empirical comparison," *Journal of the Market Research Society*, 40 (3), July, 237-248.

ENDOGENEITY BIAS – FACT OR FICTION?

QING LIU

UNIVERSITY OF WISCONSIN

THOMAS OTTER

GOETHE UNIVERSITY

GREG ALLENBY

OHIO STATE UNIVERSITY

Adaptive designs can lead to an endogeneity bias in parameter estimates. We re-examine this issue in light of the likelihood principle, and show that once the data are collected, this endogeneity is ignorable for likelihood-based estimation. The likelihood principle is implicit to Bayesian analysis, and discussion is offered for dealing with endogeneity in marketing.

1. INTRODUCTION

Regression models are at the core of conjoint analysis, where customer evaluations (y) of product offerings are related to product features (X) using coding schemes that reflect the presence or absence of specific attribute-levels. The evaluations can be in the form of ratings, in which case least squares estimation is used to produce part-worth (β) estimates; or in the forms of rankings and choices, where more sophisticated methods (e.g., maximum likelihood, Bayesian methods) are used for estimation. In all these models and methods of analysis, the product characterizations reflected in the matrix X are assumed to be uninformative about the part-worths, β , and are collectively referred to as "independent variables" – i.e., variables that are determined independently.

The characterization of X as "independently determined" makes sense. In conjoint analysis, we learn about the part-worths (β) by the responses provided by the respondent to the product descriptions. We don't learn about the part-worths from the product descriptions themselves. While the selection of specific product descriptions for evaluation or choice can affect *how much* we learn, and is the subject area known as statistical experimental design, *what* we learn in conjoint analysis comes from how respondents react to product offerings presented to them.

But, what happens when the choice of which products to display is driven by what the analyst has learned about the part-worths? In this case, while the analyst is not the originator of part-worth information, he or she uses it to set X to obtain informative responses. An example is Sawtooth Software's Adaptive Conjoint Analysis (ACA), which employs an algorithm that selects the next set of stimuli based on previous answers provided by the respondent. When this occurs, the product configurations shown to the respondent are not independently determined – they are a function of the part-worths (β) and determined from within the system of study. In other words, the product features are no longer *independent* variables.

Hauser and Toubia (HT) (2005) recently illustrated that part-worth estimates from Sawtooth Software's ACA package are prone to something called endogeneity bias originating from the relationship between the conjoint design matrix, X , and respondent part-

worths. Bias is a property that reflects the average performance of an estimator, like least squares, over multiple studies. In this paper we re-examine the bias reported by HT in light of a statistical principle known as the "likelihood principle." This principle was originally proposed by the eminent statistician R.A. Fisher (1922) as a means of conducting statistical analysis, and it is a basic tenant of modern Bayesian analysis. We show that, according to the likelihood principle, the endogeneity created by adaptive designs is ignorable – that is, it does not affect the analysis of one's data.

So, does endogeneity bias matter or doesn't it? The answer to this question depends on your point of view about statistics and the likelihood principle – it depends on whether you are a Bayesian or not (e.g., a classical, or frequentist, statistician). This situation may not be comforting to readers who see Bayesian/classical debates as irrelevant to practical analysis. In this case, being a Bayesian makes a difference.

In this paper we introduce the likelihood principle and discuss its role in statistical analysis. We then examine the endogeneity bias identified by HT, and show that the endogeneity is not of concern for likelihood-based estimation. Readers are encouraged to consult Liu, Otter and Allenby (2007) for a more detailed discussion of this topic.

2. THE LIKELIHOOD PRINCIPLE AND STATISTICAL ANALYSIS

The likelihood principle states that we learn about model parameters, such as part-worths, from data through something called the likelihood function. The likelihood function represents the statistical model that is assumed to generate the data. In ratings-based conjoint analysis, a regression model is assumed:

$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + \varepsilon_t \quad (1)$$

where y_t is a respondent's evaluation of product profile "t", the x_t values are values of attributes and their levels that describe the product, and the β values are the model parameters, the conjoint part-worths, that an analyst is interested in estimating. The term on the far right side of equation (1) is statistical error that is usually assumed to be normally distributed:

$$\varepsilon_t \sim Normal(0, \sigma^2) \quad (2)$$

The error term is needed to explain why responses from a series of product profiles (t) are not perfectly consistent. Equation (2) says that the product evaluations can deviate from what would be expected if β were known. That is, respondents are not expected to be perfectly consistent in their evaluations. The amount of inconsistency exhibited in their responses is described by the parameter σ^2 that reflects the variance of departures from perfect consistency. Respondents providing "noisy" responses are associated with large values of σ^2 , and respondents with stable and predictable responses are associated with small values of σ^2 .

Equations (1) and (2) form the likelihood function for ratings-based conjoint analysis. They provide a description of how we believe the data arise, and provide the necessary structure for estimating and interpreting the part-worths (β). According to equation (1), the

data are generated from a process where product attributes and their levels (x_t) are weighted by the part-worths (β), added together, and then added to the normally distributed error term. Since the error term has a mean of zero, the product ratings provided by the respondent are, on average, reflective of a respondent's true evaluation. The actual response differs from this true evaluation for any of many possible reasons, ranging from not paying sufficient attention to the product description, to errors in evaluation, to failing to perfectly remember and know their true part-worth, which are all left unspecified. The normal distribution is used to characterize these unobserved factors that make their way into actual responses.

The interpretation of part-worth estimates in conjoint analysis also relies on the likelihood function. Equation (1) has each product attribute and attribute-levels combining to form the expected overall evaluation. Since all of the attributes are related to the same outcome, y , they are related to each other. The value of conjoint analysis is that it facilitates measurement of the value of one attribute in terms of the others. If one attribute is price and another is some measure of performance, conjoint analysis provides a method of measuring the monetary value of performance levels by identifying the combinations of these attributes that leave the overall expected evaluation (y) fixed. In the economic literature, these combinations produce "indifference" curves. The likelihood function therefore provides a rich description of the data generation process and its structure that is integral to decision-making.

Statistical Analysis

The likelihood function represents a statistical model of how an analyst views the data generating process. The likelihood maps out a procedure for how to go about simulating data according to the model: given x_t and β , an analyst would first calculate the first set of terms on the right side of equation (1), and then generate a random draw from a normal distribution with mean zero and variance σ^2 . These would be added together to produce one response, y_t . Thus, equations (1) and (2) provide a detailed set of instructions for moving from β and σ^2 to y , given x .

Statistical analysis reverses this process by moving from y to β and σ^2 , given x . The analyst observes responses (y) to the product descriptions (x) and seeks to learn about the model parameters β and σ^2 . According to the likelihood principle, the likelihood function is the device by which this learning takes place. Equations (1) and (2) can be combined to form an expression for the statistical distribution of the observed data:

$$y_t \sim Normal(\beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt}, \sigma^2) \quad (3)$$

which indicates that the response to product description t is distributed normal with mean equal to $\beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt}$ and variance σ^2 . The likelihood (ℓ) is defined as the probability of the observed data given the model parameters (β , σ^2) and product descriptions (X). Assuming there are T independent observations, we have:

$$\ell(\beta, \sigma^2 | y) = \pi(y | \beta, \sigma^2, X) = \prod_{t=1}^T Normal(y_t | \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt}, \sigma^2) \quad (4)$$

where $\pi(\cdot)$ is used to denote a probability density, and the vertical bar "|" denotes conditional probability. For specified parameter values (β, σ^2) and product descriptions (X), the likelihood provides the probability of the observed data (y). According to equation (4), the likelihood function is actually the product of T functions, each expressing the likelihood of a specific response, y_t .

A simple approach to learning about the model parameters is to search for parameters that maximize the likelihood of the observed data. This approach, known as maximum likelihood estimation, engages in a search for the values of parameters that maximize equation (4):

$$\arg \max_{\beta, \sigma^2} \ell(\beta, \sigma^2 | y) \quad (5)$$

where "arg max" means the arguments (β, σ^2) that yield the maximum value. Some values of β and σ^2 are implausible and lead to small values of the likelihood function. Other values, however, provide a good fit to the data and yield larger values of the likelihood. Maximum likelihood estimates are defined as those parameter values that provide the best fit to the data.

A more formal approach to learning about model parameters involves the use of Bayes theorem. According to Bayes theorem, the posterior probability of the model parameters given the data (y, X) is a function of the likelihood and the prior distribution of the model parameters:

$$\begin{aligned} \pi(\beta, \sigma^2 | y, X) &= \frac{\pi(y | \beta, \sigma^2, X) \pi(\beta, \sigma^2)}{\pi(y | X)} \\ &\propto \ell(\beta, \sigma^2 | y, X) \pi(\beta, \sigma^2) \\ &\propto \text{"likelihood"} \times \text{"prior"} \end{aligned} \quad (6)$$

Bayesian analysis adheres to the likelihood principle because all the learning that takes place in an analysis is channeled through the likelihood function. The likelihood is used to update prior knowledge about the parameters to arrive at the posterior distribution.

3. TO BE OR NOT TO BE BAYES

Many alternatives are available for estimating conjoint part-worths, and not all estimation methods require assumptions about the statistical distributions of errors. Ordinary least squares is an example of a moment-based estimator that does not require the assumption of normally distributed error terms. Part-worth estimates can be obtained using the principle of least squares, and it is only when conducting hypothesis tests and evaluating test statistics (e.g., t-statistics) that distributional assumptions are needed.

In general, non-Bayesian estimates require additional assumptions for making statistical statements about the estimated part-worths. Maximum likelihood estimates that are based on equation (5), for example, assume that the likelihood function itself is locally quadratic in the region of the maximum so that estimates can be considered normally distributed. In addition, non-Bayesian estimators employ the concept of a sampling experiment to calibrate the

uncertainty of estimates. In a sampling experiment, statistical procedures such as a maximum likelihood estimator are hypothetically applied to many imaginary datasets (generated from the same parameters), and various properties such as the accuracy of the estimator are studied. These properties include bias, variance and the mean squared error of the estimated values from the true, hypothetical value. Simulation experiments are often used to measure these values across imaginary datasets.

Bayesian methods are different. Bayesian methods do not require the use of sampling experiments, and do not employ any additional assumptions for characterizing part-worth estimates. Equation (6) shows that Bayesian analysis is based on the posterior distribution, which conditions on the observed data y and the design matrix X and is derived using Bayes theorem to move from the likelihood to the posterior. The likelihood is a statement of the data generating process given the model parameters. The posterior, which is proportional to the likelihood times the prior, uses the observed data to derive probabilistic statements about the model parameters, including point-estimates similar to those obtained from maximum likelihood estimates, and confidence intervals. Moreover, Bayes estimators can be shown to have excellent sampling properties in the spirit of simulated datasets (see Liu, Otter and Allenby 2007, Rossi, Allenby and McCulloch 2005), even though their derivation completely ignores the sampling perspective.

The important point for our discussion is this: a key difference between Bayesian and non-Bayesian analysis is whether or not analysis conditions on the available data. Bayesian analysis conditions on the observed data, while non-Bayesian analysis employs the concept of a sampling experiment. We believe that sampling experiments are useful means to measure performance of an estimator and other procedures when data are not available. But, once data are collected, analysis should be likelihood-based and condition on the data.

4. FACT OR FICTION?

We now turn to the central question of this essay – does endogeneity in adaptive designs such as Sawtooth Software's ACA induce bias and is it of concern for parameter estimation? The answer is "yes" and "no." Yes, it is present. No, it is not of concern as it does not change the likelihood function of the data.

As discussed earlier, bias is a non-Bayesian concept that measures the average deviation of an estimate from its true value in a sampling experiment across imaginary datasets. In a regression model of the type used in all conjoint analyses, it can be shown that a requirement for unbiased estimates is that all the error terms $\{\varepsilon_t\}$ in equation (1) are independent of all the product descriptions $\{x_t\}$. This assumption is violated in Sawtooth Software's ACA procedure where a respondent's answers to previous questions are used to informatively construct the next question. When this occurs, the future product description (e.g., x_t) is influenced by past responses ($y_{<t}$), which by equation (1) is also linked to the past error terms ($\varepsilon_{<t}$). This results in a bias that is also present in any time series model that relates current outcomes to past outcomes.

To illustrate the bias in a simple conjoint application used by Liu, Otter and Allenby (2007), consider the model is $y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$, $t=1, 2, 3$ and $\varepsilon_t \sim \text{Normal}(0, 25)$. The

value of x for the first observation is $x'_1 = (x_{11}, x_{21}) = (1,0)$, the value of x for the second observation is $x'_2=(0,1)$ and:

$$x'_3 = \begin{cases} (1,-1) & \text{if } y_1 y_2 > 0 \\ (1,1) & \text{if } y_1 y_2 \leq 0 \end{cases} \quad (7)$$

The design rule in equation (7) is meant to mimic the "utility balance" criterion used by the ACA software. The true values of the regression coefficients are $\beta_1=1$ and $\beta_2=2$. Figure 1 illustrates the biasing effect of endogeneity using 1000 replicates of samples, each consisting of 1000 homogenous respondents. The figure shows that the regression coefficients exhibit positive bias, with $E[\hat{\beta}_1] = 1.198$ and $E[\hat{\beta}_2] = 2.406$. Each triangle character in the figure represents the mean of individual-level OLS estimates computed from one sample of $j = 1, \dots, 1000$ homogenous respondents (i.e., $\sum_{j=1}^{1000} \hat{\beta}_j / 1000$ where $\hat{\beta}_j$ is estimated with three observations $\{y_{jt}, x_{jt}\}$, $t=1,2,3$).

Figure 1
Expected Value of OLS Estimate across 1000 Replications

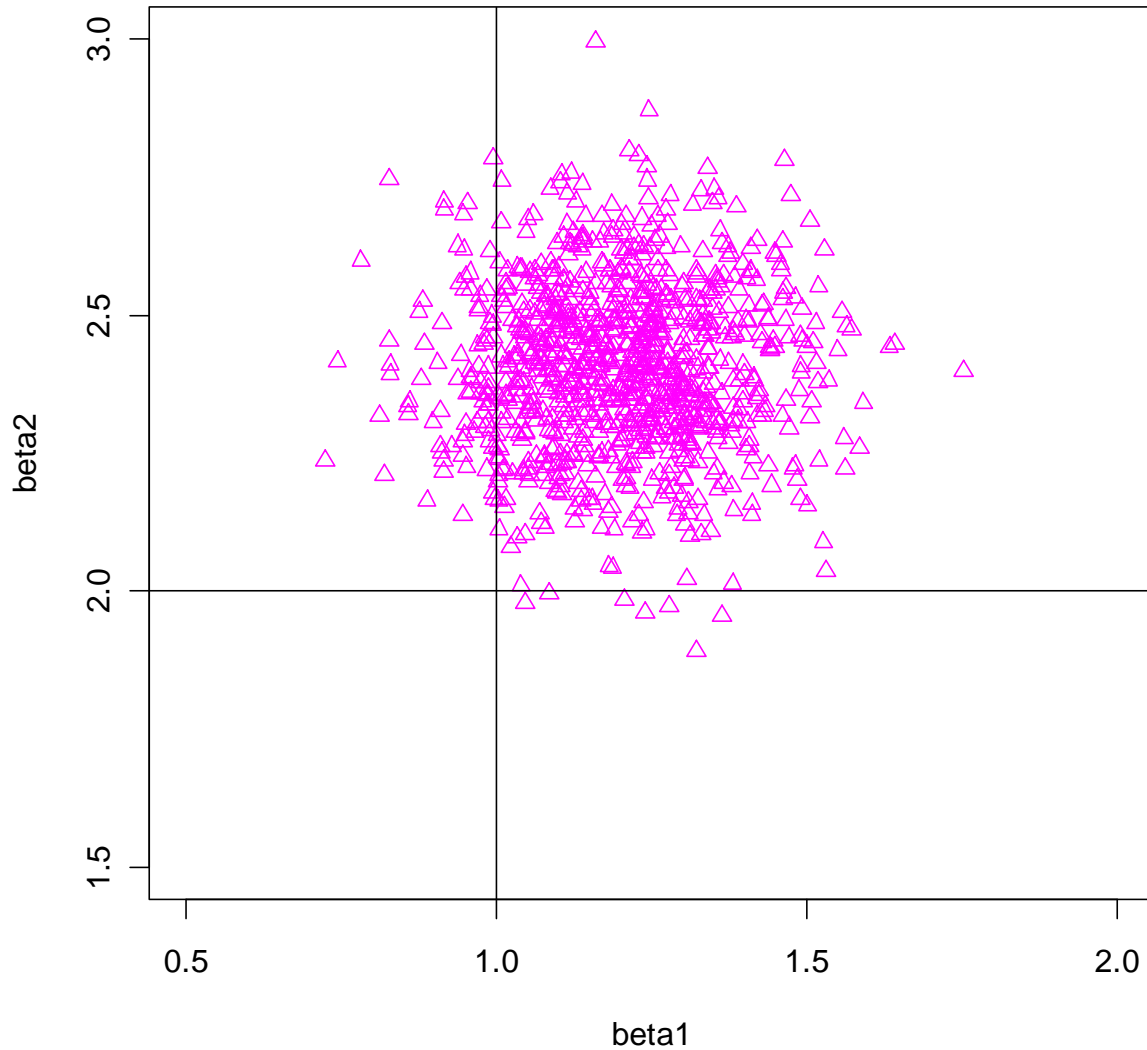


Figure 1 illustrates the presence of endogeneity bias in a sampling experiment. But, what if the data have already been collected? What if we believe that a Bayesian approach to analysis makes sense and we don't want to give much weight to arguments that investigate estimator performance across imaginary datasets? In this case, it is important to think about the effect of adaptive designs on the likelihood function in equation (4). Specifically, does the presence of endogenously determined conjoint questions change the likelihood? If not, then the endogeneity bias illustrated in figure 1 is ignorable for our data.

As it turns out, the procedure used by ACA to select informative conjoint profiles does not affect the likelihood function. The reason is due to the selection mechanism being completely determined by answers to previous questions, as in equation (7), coupled with the fact that these previous answers are also included in the likelihood. Thus, for the observed data, the likelihood of observing the next conjoint profile, *given* previous responses, is independent of the model parameters and equal to one:

$$\pi(x_t | \{y_{<t}\}, \beta, \sigma^2) = \pi(x_t | \{y_{<t}\}) = 1 \quad (8)$$

Thus, influence of endogenously determined profiles is ignorable for the given data. The likelihood is unaffected by the adaptive design, and analysis based on equation (4) is correct. Liu, Otter and Allenby (2007) demonstrate that Bayesian estimates are also unaffected by the presence of an adaptive design in a hierarchical Bayes conjoint analysis. It's worth noting that all data need to be included in the analysis. If any data that lead to the endogenously determined conjoint questions are discarded (e.g., self-explicated data), the likelihood will be altered and the endogeneity is no longer ignorable (see Liu, Otter and Allenby 2007).

5. WHAT TO DO

Bayesian analysis conditions on available data, and does not employ the concept of a sampling experiment unless the data have not yet been collected. Bayesians view sampling experiments as a useful instrument for designing experiments, but they condition on the available data once it has been collected. In addition, Bayesians as well as non-Bayesian statisticians agree that bias is one of many criteria to be considered when evaluating the performance of an estimator. Other criteria include performance in large samples (e.g., consistency of the estimator) and measures of risk such as mean squared error. Bayesian estimators can be shown to have favorable performance on these measures across all models and likelihoods (see Berger 1985 chapter 8), providing an almost universal justification for their use.

Ultimately, the answer to the question "what to do?" with data that have been collected using an adaptive design depends on whether or not you are Bayesian and adhere to the likelihood principle. Equation (8) says to Bayesians that the presence of an adaptive design in the data generating process is ignorable. Figure 1 says to non-Bayesians that an adaptive design matters in a sampling experiment and it further implies that there is no way to develop an unbiased estimator. The key is in deciding what role a sampling experiment should have in statistical analysis. This issue has been an ongoing debate in statistical theory, with researchers offering thoughtful comments in favor of, and against the likelihood principle (see Berger 1984).

More generally, dealing with endogeneity involves writing the likelihood for all dependent variables, both y and X. The likelihood for y given X is provided by Equation (4). The likelihood for X given past responses is provided by Equation (8). Thus, the likelihood for all the data for ACA's adaptive design is the product of equations (4) and (8). Likelihood-based inference for endogenously determined variables is based on a likelihood for all variables created within the system of study.

A different question is whether adaptive designs should / should not be employed in this context. This question extends well beyond our discussion here and essentially goes back to Rich Johnson's original arguments that utility balanced pairs result in more thoughtful and involved response behavior. This is, little is learned from answers to obvious questions. Utility balance aims to tradeoff bias for reduced variance, a worthy objective that has a long history in statistical science.

Many factors play into the decision to be a Bayesian. We believe the most compelling is the conceptual simplicity with which Bayes theorem treats objects of inference, whether they are conjoint part-worths, alternative models, or hypotheses. Bayes theorem applies equally to all. Moreover, the practice of marketing involves decisions and actions taken on the basis of specific data, not hypothetical data. Conditioning on the observed data is what makes Bayes theorem practically compelling.

REFERENCES

- Berger, James O. and Robert L. Wolpert (1984). *The Likelihood Principle*, Institute of Mathematical Statistics Lecture Notes – Monograph Series.
- Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag: New York.
- Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics," *Philos. Transactions of the Royal Society, London, Series A*, 222, 309-368.
- Hauser, John R. and Olivier Toubia (2005). "The Impact of Utility Balance and Endogeneity in Conjoint Analysis," *Marketing Science*, 24, 498-507.
- Liu, Qing, Thomas Otter and Greg M. Allenby (2007). "Investigating Endogeneity Bias in Marketing," *Marketing Science*, forthcoming.
- Rossi, Peter E., Greg M. Allenby and Robert McCulloch (2005). *Bayesian Statistics and Marketing*, John Wiley & Sons.
- Sawtooth Software (2003). *ACA/Hierarchical Bayes v2.0*, Technical Paper, <http://www.sawtoothsoftware.com/technicaldownloads.shtml#acahbtech>.

CBC/HB, BAYESM AND OTHER ALTERNATIVES FOR BAYESIAN ANALYSIS OF TRADE-OFF DATA

WELL HOWELL
HARRIS INTERACTIVE

INTRODUCTION

Trade-off studies of paired comparison, ratings-based conjoint and discrete choice data offer a number of challenges during the model building phase. Gelman and Hill (2007) suggest that we “fit many models,” and that we start with very simple models. For example, a logistic regression model gives aggregate level results for paired comparison data. In the healthcare field, a “second opinion” is often recommended. Multiple analytic techniques are commonly recommended in very diverse fields (Sheth, Roscoe & Howell 1974, Walker 2007). This paper will examine some of the issues involved in using CBC/HB and other analytic alternatives to study MaxDiff or best-worst comparisons and stated-preference discrete choice data.

Rossi and McCulloch released an R (R Development Core Team 2007) package called “bayesm” for Bayesian analysis of marketing research data in 2006, associated with the Bayesian Statistics and Marketing text by Rossi, Allenby and McCulloch (2005). Other R packages, including MCMCpack (Martin & Quinn 2007), R2WinBUGS (Sturtz & Gelman 2005) and BRugs (Thomas *et al.* 2006), provide additional Bayesian modeling capabilities. R2WinBUGS and BRugs drive versions of the WinBUGS (Spiegelhalter, Thomas, Best & Lunn) or OpenBUGS (Thomas *et al.* 2007) software. Since both WinBUGS and OpenBUGS software are used for this paper, the generic term “BUGS” will be used to refer to either.

CBC/HB offers the advantages of high speed and ease of use in a market research setting, especially when used in conjunction with other Sawtooth Software packages. Other advantages of multiple packages from a single supplier can include simplified training needs for an analytic team and the option to transfer projects from one team member to another, since all members are familiar with “standard” software. However, R packages also prove useful during model development and testing. The bayesm routine “rmnlIndepMetrop” provides a model similar to that used in CBC/HB, where a single mean vector and prior variance/covariance matrix are used for the prior. A second bayesm routine, “rheirMnlRwMixture” provides a more complex model where respondent heterogeneity can be described with a mixture of normal components in the prior. All the R packages mentioned above allow for multiple MCMC chains to support multi-chain convergence diagnostics such as the Gelman and Rubin R-Hat statistic, which is available in the R “coda” (Plummer *et al.* 2007) package.

This paper describes two MaxDiff examples and a stated-preference discrete choice example. These three examples were analyzed with bayesm, BUGS, CBC/HB and Harris Interactive proprietary software (HIhbmkl). The BUGS software was not used for the additional tests reported here using various prior covariance matrix values as outlined in an ART Forum paper by Orme and Lenk (2004).

EXAMPLE DATASETS

A synthetic MaxDiff dataset on automobile tire features, similar to Howell (2004), along with disguised datasets from a physician MaxDiff task and a physician stated-preference choice task from actual Harris Interactive projects will be used to demonstrate the strengths and weaknesses of the four software packages under test (bayesm, BUGS, CBC/HB and Hihbmk1).

The synthetic tire MaxDiff data consists of the 11 features shown in table 1. Sub-groups contain 50 respondents each, and are defined by region of the US (North / South) and by age of respondent (Younger / Older). Some features, such as “Well known manufacturer” are almost constant across the covariate groups, while others, such as “Better traction in ice/snow” show group differences. Utility values for all 200 respondents were generated from a multivariate normal distribution and Gumbel noise was added to each utility drawn. With these “known” utilities available, a SAS DATA step was used to create a CBC/HB .CHO (Sawtooth Software 2005) data file using 10 “versions” of a 10 task, 3 item design created by the Sawtooth MaxDiff Designer package (Sawtooth Software 2006). Five respondents from each group “answered” each of the 10 design versions. Region and age values were included in the .CHO file as “extra” variables. Once the .CHO file was created, simple reformatting steps (AWK scripts, GREP commands and the R2WinBUGS or BRugs packages) were used to arrange the study information in the forms required by the other three software packages.

Feature	North Younger	South Younger	North Older	South Older
Better traction in ice/snow	6.62	5.02	6.64	5.55
Well known manufacturer	6.93	7.09	6.60	6.93
Less risk of blowout	6.16	6.31	6.04	6.15
Less hydroplaning	6.52	6.93	5.25	5.25
Stronger steel belts	4.18	4.73	5.80	5.65
Less need to check pressure	6.89	6.70	6.45	6.09
Better control of vehicle	4.69	4.33	4.79	4.80
Better for interstate driving	5.55	5.26	5.88	4.99
Safer at very high speeds	3.54	3.86	3.66	3.54
Easier to locate at stores	3.34	3.40	3.09	3.45
Wider range of sizes	4.42	4.48	4.42	4.61

Table 1
Mean Vectors by Sub-group

The physician MaxDiff data consisted of 18 aspects of a chronic disease shown to three specialty groups of 76, 78 and 76 physicians (230 in total). Each physician chose a most and a least important disease aspect from 13 tasks with four aspects each. The only covariate used was a pair of effects codes to represent the three specialty groups.

The physician choice data consisted of 18 tasks of 3 choice alternatives each. There were four continuous attributes, a 2-level attribute, and the two effects codes needed to represent the three choice alternatives (2 ASCs). Each of the 94 physicians answered 18 tasks. There

were 5 covariate degrees of freedom available for those software packages that used them (bayesm and HIhbmkl).

IN-SAMPLE RESULTS

Initial tests of all four software packages on all three example datasets produced timing information and in-sample estimates of model accuracy. These in-sample results were used initially to investigate differences between the packages on identical input data without regard to which result would yield the best predictions. Holdout-sample estimates will be used later in this paper to examine prior covariance matrix issues. Timing values are difficult to obtain, as different packages offer different run time measures or none at all. The computers used also showed a variability of +/- 10% in run times (total duration or “wall” time) of a single reference run. Given the large differences in time for the packages, however, the difference between computers will be ignored.

Run or wall times for the four packages are shown in Table 2. Due to different chain lengths used, the most interesting speed value is minutes per 10,000 iterations.

	bayesm	BUGS	CBC/HB	HIhbmkl
11 item MaxDiff (n=200)				
run time (minutes)	614	182	15	116
iterations (or chains x iter.)	2x100K	2x20K	100K	2x175K
min/10K iterations	30.70	45.50	1.50	3.31
18 item MaxDiff (n=230)				
run time (minutes)	412	422	43	302
iterations (or chains x iter.)	2x50K	2x20K	100K	2x100K
min/10K iterations	41.20	105.50	4.30	15.10
7 dof DCM (n=94)				
run time (minutes)	300	59	5	26
iterations (or chains x iter.)	2x50K	2x20K	100K	2x175K
min/10K iterations	30.00	14.75	0.50	0.74

Table 2
Elapsed Time Comparisons for In-Sample Runs

CBC/HB is the clear speed winner, with HIhbmkl coming in second, based on minutes of wall time needed to perform 10,000 iterations. The HIhbmkl package pays a penalty in the 18 item MaxDiff run, as it defaults to univariate draws, rather than the more usual multivariate draws in an attempt to deal with some mixing issues. The `rmnlIndepMetrop` routine in the bayesm package and BUGS vie for slowest as expected (bayesm uses some compiled C and C++ functions for speed, but not in the `rmnlIndepMetrop` function).

In-sample log-likelihood values and hit rates were calculated using the point estimates for utilities based on respondent-level averages over the second half of the chain (bayesm BUGS and HIhbmkl used averages over the second half of two chains). Both log-likelihood and hit rate calculations were performed using a modified version of the bayesm llmnl function, which calculates the multinomial log-likelihood for the bayesm package. These values are displayed in Table 3. None of the packages is a clear winner on either log-likelihood or hit rate across all three examples. This was a surprise, since two of the packages (bayesm and HIhbmkl) made use of respondent-level covariates.

	Bayesm	BUGS	CBC/HB	HIhbmkl
11 item MaxDiff (n=200)				
Log-likelihood	-149.5	-102.6	-153.9	-102.5
Hit rate	99.8	99.9	99.8	99.7
18 item MaxDiff (n=230)				
Log-likelihood	-2589.1	-3246.4	-2393.7	-3282.0
Hit rate	86.1	81.1	87.7	80.5
7 dof DCM (n=94)				
Log-likelihood	-307.1	-370.2	-297.2	-362.1
Hit rate	92.1	90.5	92.8	90.5

Table 3
In-Sample Log-likelihood and Hit Rate

HOLDOUT SAMPLE TESTS WITH DIFFERENT PRIOR COVARIANCE MATRICES

Confusion over the sample and update versions of the BUGS thin command caused very slow preliminary in-sample runs. As a result, BUGS software was used only for in-sample tests. The bayesm, CBC/HB and HIhbmkl packages will be used for the remainder of the tests reported here.

Tests were performed on the three example datasets using the three surviving software packages to examine the impact of different prior covariance matrices as suggested in the A/R/T Forum paper by Orme and Lenk (2004). Their paper found a “proper” covariance matrix was more useful than the standard identity matrix, especially when the data were sparse and for larger effects-coded variables like the two MaxDiff example datasets in this paper. While none of the three examples is extremely “sparse,” some improvements might be uncovered via a “tuning process” in which we search for a multiplier of the “proper covariance matrix” that will yield maximum holdout log-likelihood and/or hit rate. One “task” (MaxDiff task pair) was held out at random for every respondent in the three example datasets to provide a single observation per respondent on which to perform an out-of-sample estimate of log-likelihood and hit rate, and also to make each example dataset just a bit more “sparse.”

One advantage of using the R package for this paper is the simplicity of generating the sets of holdout tasks. For example, the simple R command:

```
picked = sample(10,200,replace=TRUE) -1
```

produced the vector of zero referenced tasks “picked” from the synthetic Tire MaxDiff data to be “holdout tasks.” For the two MaxDiff examples, there is a Best and a Worst holdout task, while for the DCM example, there will only be a single holdout task. All tasks not picked for holdout will be used to develop the models being evaluated on holdout log-likelihood and hit rate.

HOLDOUT SAMPLE TESTS—11 ITEM TIRE MAXDIFF

The synthetic Tire MaxDiff data now consists of 9 estimation task pairs and one holdout task pair with 11 tire attributes, which yields a 10 parameter model. All earlier runs were performed with the prior covariance matrix taken as a simple identity matrix (I). Orme and Lenk (2004) recommend a prior covariance matrix which is made up from the sum of an identity matrix and a matrix of all ones (J). This more proper prior covariance matrix can be made even “stronger” by multiplying it by a constant. This constant is often a bit larger than the number of model parameters. The tests here will use three prior covariance matrices; the identity (I); the identity plus ones (I+J); and the “stronger” identity plus ones times degrees-of-freedom plus 3 (or 13 for our 10 parameter model, giving (I+J)*13). This last matrix is shown in Table 4.

	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
U1	26	13	13	13	13	13	13	13	13	13
U2	13	26	13	13	13	13	13	13	13	13
U3	13	13	26	13	13	13	13	13	13	13
U4	13	13	13	26	13	13	13	13	13	13
U5	13	13	13	13	26	13	13	13	13	13
U6	13	13	13	13	13	26	13	13	13	13
U7	13	13	13	13	13	13	26	13	13	13
U8	13	13	13	13	13	13	13	26	13	13
U9	13	13	13	13	13	13	13	13	26	13
U10	13	13	13	13	13	13	13	13	13	26

Table 4
(Identity + Ones)*13 Prior Covariance Matrix for 11-item Tire MaxDiff

Trying these three prior covariance matrices on the synthetic Tire data with each of the three software packages gave the holdout log-likelihood and hit rate results shown in table 5. These results are for the 200 holdout task pairs (best & worst) picked with the R sample function above. The bayesm and HIhbmkl results include the region and age covariates in their upper-level model, while the CBC/HB package does not offer the capability for upper-level covariates. CBC/HB results are based on the back half of a single chain (the .CSV output), while bayesm and HIhbmkl results are averages of the back half of two chains. All chains were run for 50,000 burn-in and 50,000 estimation iterations. While the CBC/HB

package does show an improvement in log-likelihood and hit rate, the other two packages seem almost “flat.” There was no investigation of how the bayesm and Hihbmkl packages would have performed without covariates.

	bayesm	CBC/HB	Hihbmkl
Identity Matrix			
Log-likelihood	-46.00	-20.82	-24.88
Hit rate	97.5	97.0	96.5
Identity+Ones			
Log-likelihood	-53.52	-20.53	-28.93
Hit rate	97.5	97.5	96.5
(Identity + Ones) x 13			
Log-likelihood	-46.21	-13.66	-26.91
Hit rate	97.5	98.0	96.5

Table 5
Log-likelihood and Hit Rate 11-item MaxDiff

HOLDOUT SAMPLE TESTS—18 ITEM MAXDIFF

The original 18-item MaxDiff example dataset consisted of Best/Worst choices on 14 quad tasks (14x2=28 choices). Using the R sample function as above, one of the quad tasks was chosen for each of the 230 respondents as a holdout task, providing 230 x 2 = 460 holdouts and 230 x 26 = 5,980 modeling choices. Again, three prior covariance matrices were used; the identity (I), the identity plus ones (I + J) and (I + J) * 20, with this final matrix set up much like Table 4, but as a 17 x 17 matrix with values of 40 on the diagonal and 20 (17 degrees-of-freedom + 3) in off-diagonal cells. Log-likelihood and hit rate values on the 460 holdout tasks are shown in Table 6. CBC/HB continues to be the most sensitive to the strongest matrix ((I + J)*20), but now the bayesm package also shows some improvement, while Hihbmkl again remains “flat.”

	bayesm	CBC/HB	Hihbmkl
Identity Matrix			
Log-likelihood	-135.62	-104.42	-133.04
Hit rate	79.1	83.0	78.3
Identity+Ones			
Log-likelihood	-101.24	-103.56	-133.57
Hit rate	84.8	83.9	79.1
(Identity + Ones) x 20			
Log-likelihood	-101.07	-60.55	-134.45
Hit rate	83.0	90.9	79.1

Table 6
Log-likelihood and Hit Rate 18-item MaxDiff

HOLDOUT SAMPLE TESTS—7 DOF DCM

Finally, the discrete choice example needs a more complex prior covariance matrix due to the presence of both continuous and effects-coded variables. For the seven predictor variables, the strongest prior covariance matrix is shown in Table 7. The two other matrices would be the usual 7x7 identity matrix and a special identity plus ones matrix made by dividing every cell in Table 7 by 10.

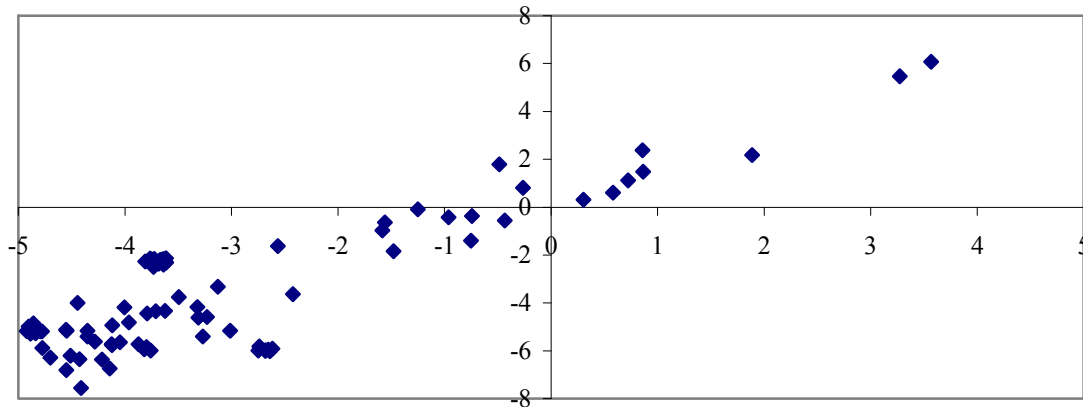
	U1	U2	U3	U4	U5	U6	U7
U1	10	0	0	0	0	0	0
U2	0	10	0	0	0	0	0
U3	0	0	10	0	0	0	0
U4	0	0	0	10	0	0	0
U5	0	0	0	0	10	0	0
U6	0	0	0	0	0	20	10
U7	0	0	0	0	0	10	20

Table 7
7 dof Special Prior Covariance Matrix
for DCM

Trying these three prior covariance matrices on the DCM data with each of the three software packages gave the holdout log-likelihood and hit rate results shown in table 8. For this example dataset, both bayesm and CBC/HB show a worse log-likelihood with the matrix in Table 7 while Hihbmkl shows an inconclusive log-likelihood. The flat hit rates are the result of identical predictions for all three matrices. Figure 1 demonstrates the major shift in respondent-level average utilities (.CSV file) for the CBC/HB analysis comparing an Identity prior (x) versus the special matrix of Table 7. The level of disagreement between these two models and the fact that a DCM holdout task offers only one holdout choice may be part of the cause for the increase in log-likelihood.

	bayesm	CBC/HB	Hihbmkl
Identity Prior Matrix			
Log-likelihood	-536.26	-440.25	-536.26
Hit rate	40.4	39.4	40.43
Table 7 / 10 Prior Matrix			
Log-likelihood	-564.58	-685.97	-564.58
Hit rate	40.4	39.4	40.43
Table 7 Prior Matrix			
Log-likelihood	-444.51	-707.65	-537.86
Hit rate	40.4	39.4	40.4

Table 8
Log-likelihood and Hit Rate 7 dof DCM



**Figure 1 -
CBC/HB Item_7 (ASC_2)
Identity (x) vs Table 7 Prior Matrix (y)**

CONCLUSIONS AND NEXT STEPS

Alternatives to CBC/HB, such as bayesm, BUGS or proprietary software seem useful for situations in which one might want control of issues like dispersed chain starting points, respondent-level covariates or a constant Metropolis step size during estimation iterations. However, the CBC/HB speed advantage over these other alternatives would suggest that initial models should be developed under CBC/HB. Most post-modeling questions, such as batched mean or time series estimates of standard errors can be handled within the R package using CBC/HB .DRA file output.

There is the possibility that one can learn to “edit” the CBC/HB restart file after just a few iterations, so that the issue of dispersed starting points might be simulated. However, given that the CBC/HB Metropolis step size can’t be held constant, a simpler approximation might be to run two chains using the existing 0-start and smart-start capabilities of CBC/HB with different “seed values.” The (possibly thinned) .ALP files from CBC/HB can then be analyzed with custom software or the R coda package to test for MCMC convergence.

The investigation of different prior covariance matrices was inconclusive, although these example datasets were not very sparse. Sawtooth Software (2006) recommends each MaxDiff item be shown at least 3 times as was approximately done in the MaxDiff examples. Next steps include a return to the synthetic Tire data where many additional holdout tasks are available and the full set of original tasks can be retained for modeling. A more closely spaced “grid” than the 3 prior matrices used here (I, I+J and (I+J)*nu) may also help understand the process. The current suggestion is to use at least the I+J prior covariance matrix on all MaxDiff projects (the Sawtooth Software MaxDiff Designer provides an .MTRX file, so the only extra step in using it is to check the appropriate box on the advanced settings tab of CBC/HB). Other packages, such as bayesm, could create the needed matrix for a MaxDiff design of size nItems+1 with the simple statement:

```
properPriorCovariance = diag(1,nItems) + matrix(1,nItems,nItems)
```

or almost as easy as checking a box.

The use of a proper covariance matrix has proved helpful when performing Bayesian analyses of Harris Interactive COMPASS® data in which tasks are pairwise comparisons. However, each item is usually shown only two times, rather than the recommended three times. Such “two-show pairs” were adequate for the aggregate-level analysis used previously, but can cause difficulty during Bayesian analysis with lists of 21 items or more. For a 21 item list, each respondent sees 21 pairs, or only 10% of the possible pairs of 21 items, causing a sparse data situation.

REFERENCES

- Gelman, Andrew, and Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, p. 547.
- Howell, Well (2004). "Reducing Respondent Burden in Ranking Tasks: Hierarchical Bayesian Analysis of Pairwise Comparisons with Covariates," JSM2004 Abstract - #301746 <http://www.whowell.biz/articlew.pdf>.
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325-337.
- Martin, Andrew D., and Kevin M. Quinn (2007). MCMCpack: Markov chain Monte Carlo (MCMC) Package. R package version 0.8-2. <http://mcmcpack.wustl.edu>.
- Orme, Bryan, and Peter Lenk (2004). "HB Estimation for 'Sparse' Data Sets: The Priors Can Matter." 2004 ART Forum, American Marketing Association, Whistler, BC.
- Plummer, Martyn, Nicky Best, Kate Cowles and Karen Vines (2007). coda: Output analysis and diagnostics for MCMC. R package version 0.12-1.
- Roscoe, A. M., J. Sheth and W. Howell (1974). "Intertechnique Cross-Validation in Cluster Analysis," in R. C. Curham (ed.) 1974 Combined Proceedings (American Marketing Association) <http://www.jagsheth.net/docs>.
- Rossi, Peter E., Greg Allenby, and Rob McCulloch (2005). Bayesian Statistics and Marketing. John Wiley and Sons, December 2005.
- Rossi, Peter, and Rob McCulloch (2007). bayesm: Bayesian Inference for Marketing/Micro-econometrics. R package version 2.1-3. <http://faculty.chicagogsb.edu/peter.rossi/research/bsm.html>.
- R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Sawtooth Software (2005). The CBC/HB System for Hierarchical Bayes Estimation Version 4.0 Technical Paper, <http://www.sawtoothsoftware.com/download/techpap/hbtech.pdf>.
- Sawtooth Software (2006). MaxDiff Designer v2, <http://www.sawtoothsoftware.com/download/techpap/bwdes.pdf>.
- Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-16.
- Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS Open. *R News* 6 (1), 12-17.
- Walker, Gabrielle (2007). "The Biggest Thing in Physics" (Discover, 8/2007).

RESPONDENT WEIGHTING IN HB

JOHN HOWELL
SAWTOOTH SOFTWARE

INTRODUCTION

CBC/HB is often used to analyze CBC studies. A recent Sawtooth Software survey indicates that among those who used CBC in the previous year, 69% of users employed CBC/HB for their final models (Sawtooth Solutions, Fall 2007). Often these CBC studies cover different market segments that may have heterogeneous preferences. These different preferences could be related to differing demographics such as country of residence, gender, or whether the respondent has bought specific products in the past. One of the properties of CBC/HB is that individual respondent's utilities are shrunk toward the average utility of the population for each attribute level. This has been a concern for samples that are not representative. An over-sampled group (customers for instance) could have a disproportionately large impact on the utilities of the under-sampled group. The effect of this is that products appealing to the larger group in a market simulator will have their shares overstated while products appealing to the smaller group would have their shares understated. Since the bias is inherent in the utilities, weighting the market simulator will not completely correct for the misrepresentation of the shares.

The purposes of this paper are to examine the severity of the bias that comes from non-representative samples and to propose a possible solution. Because the problem manifests itself as bias in the utilities, the paper will use a simulation study to examine the results.

BACKGROUND

At Sawtooth Software, we occasionally have customers approach us with questions about CBC/HB's ability to handle studies with different groups with different preferences. These customers will often have a sample that comes from two lists, a list of customers and a list of non-customers, for example. Because of the nature of the problem it is often easier to elicit interviews from customers than non-customers, so customers make up a larger portion of the responses than their actual market share. Since HB is commonly said to "borrow" information from the other respondents, the concern is that the large proportion of customers will bias the utilities of the non-customers. The natural desire is to weight the sample so that the non-customer responses weigh more heavily in the borrowing procedure and the customers contribute less weight.

One thing to note is that since CBC/HB is an individual-level model, the output from the estimation is one set of utilities for each respondent. This means that the data still need to be weighted in any analysis of the utilities (assuming a disproportionate sampling plan as above), including market simulations and average utility computations.

Information-borrowing in CBC/HB occurs through the use of a prior distribution. (For more details on the way CBC/HB works, see "CBC/HB Technical Paper" and "Understanding HB: An Intuitive Approach" from the Sawtooth Software Technical Papers Library at <http://www.sawtoothsoftware.com/education/techpap.shtml>).

In CBC/HB, the prior distribution assumes that *a priori* the betas can be drawn from a multivariate normal distribution. (This does not mean that the final estimate of the betas will actually be distributed multivariate normal.) CBC/HB also uses an uninformative prior distribution for the mean of this prior distribution. This means that the *a priori* mean of this distribution is 0 and the variance is infinite. This uninformative distribution implies that all information about the location of the parameters comes from the data. It also means that the prior distribution for each respondent's betas is the mean of all the respondents' betas. In practical terms this means that all the respondents shrink toward the overall sample mean. This works well for situations where the respondent utilities are fairly normally distributed. When there is more than one group with different utilities, shrinking to the overall mean could lead to a worse fit since the mean is no longer the most likely value for any single respondent. Figure 1 is a hypothetical illustration of what this would look like. Because there is a complex interplay in the model, a real situation would not necessarily behave exactly like this.

CBC/HB Shrinkage Example

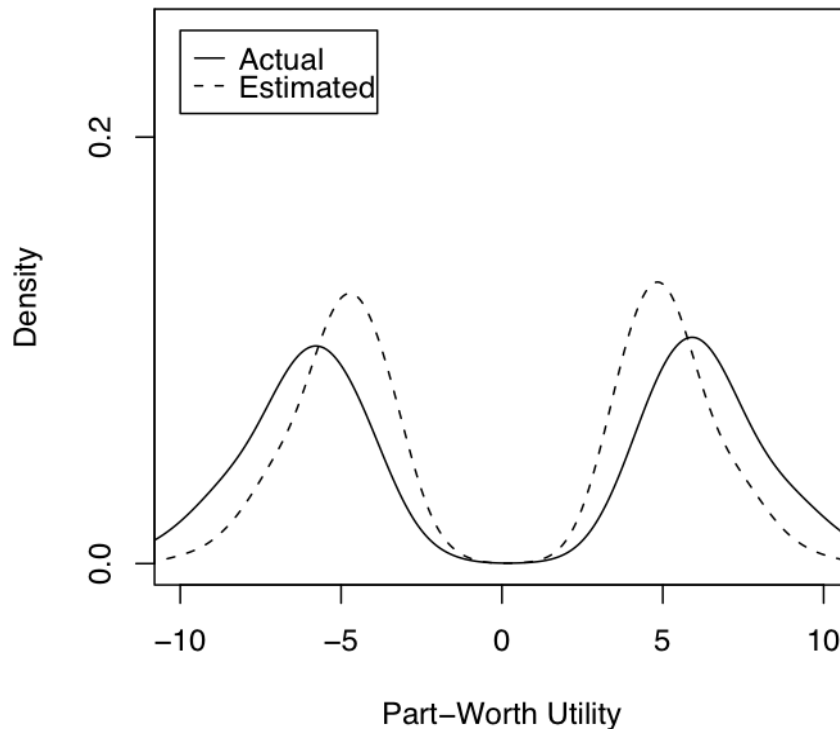


Figure 1
Hypothetical Shrinkage Example

With two groups of equal size, the shrinkage causes a similar effect on both groups and is often counteracted by the increase in the variance of the model over a single group. However, the concern we have heard voiced by our customers is that this effect can be a concern when the two groups are not the same size. The smaller group must shrink much more than the larger group if the mean is to stay in the same location as shown in Figure 2. Additionally, the variance

is smaller when you have two unequal groups than when the groups are equal. This smaller variance leads to greater shrinkage.

CBC/HB Shrinkage Example

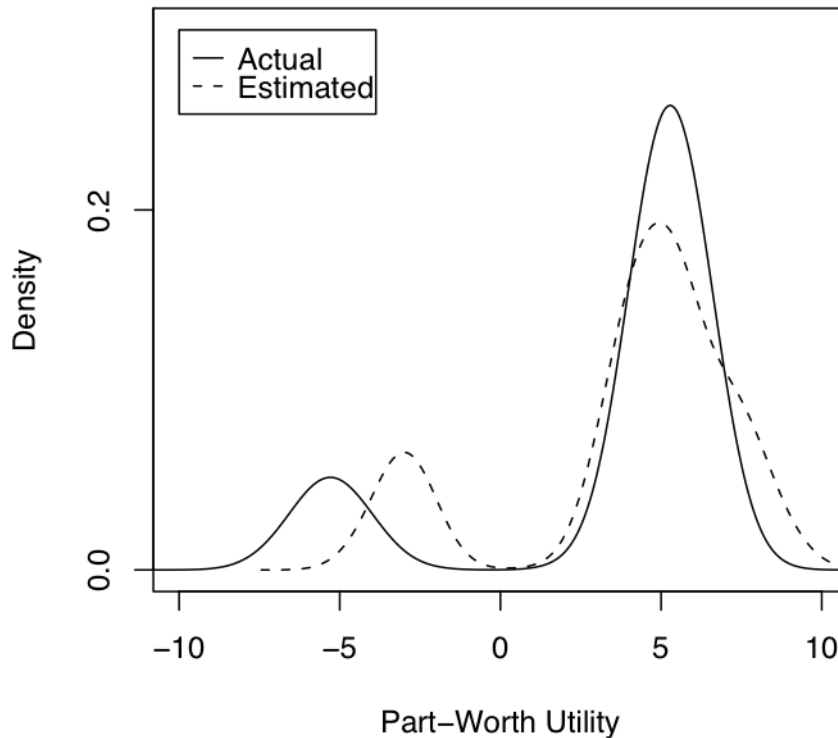


Figure 2
Hypothetical Shrinkage for Two Groups of Unequal Size

The result in the market simulator is that if there is a product preferred principally by the smaller group and a second product that mainly appeals to the larger group, the product for the larger group will receive a greater portion of the share than it normally would and the share for the product appealing to the smaller group would be understated.

This paper attempts to quantify the severity of the problem and then attempts a simple weighting scheme to correct for the issue. The solution is incomplete as it only seeks to adjust the mean of the prior distribution by weighting the groups so that the shrinkage does not influence the smaller group so dramatically. Because the paper examines bias in the individual-level estimates of the utilities, it is necessary to know the true utilities. Therefore, we use simulated respondent data where the true utilities are known. A couple of data sets are used, but both reflect a fairly standard design generated by CBC/Web. The design has 4 attributes with 4 levels, 5 levels, 2 levels, and 7 levels. Each design contains 12 tasks and 300 versions (blocks). The respondents' utilities were generated based on known utilities. First, utilities were determined for each group. Then each respondent's utilities were generated by making a random draw from a multivariate normal distribution with mean vector equal to their group's utilities and variance equal to 1. The respondents were then assigned to a random version of the questionnaire and simulated as if they were taking the questionnaire. Each respondent chose the concept that had the highest utility determined by summing the respondent's utility for each level

plus Gumbel error. The analysis was then done using either standard CBC/HB or a customized version of CBC/HB that employs the weighting scheme described later in the paper under the heading *Weighting in CBC/HB*.

ONE GROUP

As a base case, we ran a series of models with a single group. The purpose of this exercise was to establish a base line prior to testing multiple groups. A series of simulations was done at various samples sizes. All respondents in all runs had the same base utilities. The sample sizes varied with a run at 50 respondents, 100 respondents, 200 respondents, 300 respondents, 600 respondents, 900 respondents, 1800 respondents, and 3600 respondents. The purpose of these runs was to establish a base line against which to compare the other groups' shrinkage. Three measures are used as an evaluation of the fit. The first is the correlation between the average estimated utilities and the average actual utilities. Previous experience has shown that this should be .99 or higher for CBC/HB runs. The second measure is the root mean squared error between the true utilities and the estimated utilities. It is calculated by taking the difference between the estimated value and the expected value, squaring this difference, taking the mean of the squares, and then taking the square root of the mean. Lower numbers are better and the larger the sample, the smaller this number should be. The final measure is the average bias. It is calculated by taking the average of the absolute value of the bias for each parameter. The bias for each parameter is just the mean of the difference between the estimated utility and the actual utility for each respondent. The averaged differences will be both positive and negative. In order to prevent a positive value from canceling a negative value the absolute value is taken. This number should capture the amount of shrinkage that occurs when calculated at the group level.

Table 1
Base Case Results

	50	100	200	300	600	900	1800	3600
Correlation	.996	.997	.999	.999	1.00	1.00	1.00	1.00
RMSE	5.12	3.75	2.73	2.28	1.68	1.46	1.09	0.60
Avg. Bias	4.19	2.97	2.22	1.77	1.30	1.09	0.76	0.30

In Table 1 are the results from the base case run. The data show that the Average Bias is fairly high for all group sizes, but the correlations are also very high. This generally means that the relationship between the estimated and the actual utilities is maintained, but the scale factor applied in Multinomial Logit (MNL) analysis is different. In MNL analysis the amount of error that a respondent answers with is directly related to the range of the utility estimates. The range of the utility estimates is called scale. When respondents answer the survey consistently, the scale factor is larger, since the error is reduced. Since CBC/HB infers a large amount of information from relatively few responses and respondents are often fairly consistent within an interview, it is common for CBC/HB scale factors to be larger than they would otherwise be. This is often exacerbated in simulation studies since the respondents' answers are more structured and usually a little more consistent than they are in real life. By examining the bias on the individual level, it appears that scale factor is indeed the issue, since the negative utilities are generally more negative and positive utilities are more positive.

It is possible to account for the scale factor by multiplying the utilities by a positive constant. Sawtooth Software’s SMRT market simulator does this by multiplying all the individual utilities by a global multiplier. It does not allow for individual respondents to have unique scale factors. The recommendation is that you tune the scale factor against fixed hold out tasks by choosing a scale factor that minimizes either MSE or MAE of the hold out tasks versus matching simulations. With simulated data it is also possible to tune the scale factor at the individual level by using a regression model to determine what scale factor multiplier minimizes the difference between the estimated utilities and the actual utilities. Table 2 duplicates the output from Table 1 using both an aggregate and an individual level scale factor adjustment.

Table 2
Tuned Results for Base Case

		50	100	200	300	600	900	1800	3600
Aggregate Tuning	RMSE	0.62	0.62	0.56	0.59	0.54	0.52	0.49	0.45
	Avg. Bias	0.22	0.17	0.10	0.13	0.06	0.05	0.04	0.03
Individual Tuning	RMSE	0.67	0.65	0.61	0.63	0.58	0.57	0.54	0.50
	Avg. bias	0.22	0.17	0.09	0.13	0.06	0.05	0.04	0.03

Tuning the estimates has a dramatic decrease in the bias of the estimates as well as the RMSE. It appears that with this model the aggregate tuning method is equivalent to the individual tuning method. The RMSE is very similar between the two methods and the Average Bias is nearly identical. This implies that all the respondents in the simulation have a similar scale factor. Examination of the individual scale factors reveals that this is the case. The individual scale factors have a mean equal to the aggregate scale factor and a very narrow variance.

TWO EQUAL SIZED GROUPS

The next base case to consider is two groups of equal size. The simulation is very similar to the previous one except that the respondents are split into two groups. The utilities for the groups are nearly opposite. The base utilities for the two groups are shown in Table 3.

Table 3
Utilities for the Two Groups

	Group 1	Group 2
Att. 1 Level 1	-5.25	5.25
Att. 1 Level 2	-1.25	1.25
Att. 1 Level 3	2.75	-2.75
Att. 1 Level 4	3.75	-3.75
Att. 2 Level 1	-2.2	0.8
Att. 2 Level 2	-5.2	4.8
Att. 2 Level 3	1.8	1.8
Att. 2 Level 4	4.8	-5.2
Att. 2 Level 5	0.8	-2.2
Att. 3 Level 1	-1	-1
Att. 3 Level 2	1	1
Att. 4 Level 1	-6	-5
Att. 4 Level 2	-4	-3
Att. 4 Level 3	-2	-1
Att. 4 Level 4	0	0
Att. 4 Level 5	2	1
Att. 4 Level 6	4	3
Att. 4 Level 7	6	5

The results in Table 4 are much the same as the one-group results. The correlation is very high for all solutions. This suggests that the results are a good match, but there is a scale issue. The significantly lower RMSE and Average bias for the tuned results suggest that adjusting for scale factor accounts for most of the bias in the results. A couple of other things stand out. The bias for the tuned results in this simulation is higher than for the single-group case. It appears that there is some residual bias that is not accounted for by tuning the results. This is consistent with the theory of Bayesian shrinkage since the utilities for both groups should move slightly toward the mean. Another interesting finding is that the individually tuned results have a slightly lower bias and RMSE than the aggregate tuned results. While the difference is slight it appears at all sample sizes. This is expected since the raw utilities for the two groups have a slightly different range and would thus be expected to have a slightly different scale factor.

Table 4
Results for Two Equal Groups

		300	600	900	1800	3600
Untuned Results	Correlation	.999	.999	.999	.999	.999
	RMSE	2.39	1.87	1.90	1.56	1.15
	Avg. Bias	1.64	1.18	1.17	0.87	0.48
Aggregate Tuned Results	RMSE	0.85	0.85	0.88	0.86	0.83
	Avg. Bias	0.17	0.17	0.17	0.17	0.14
Individual Tuned Results	RMSE	0.75	0.74	0.73	0.71	0.69
	Avg. Bias	0.13	0.12	0.11	0.11	0.08

TWO GROUPS WITH DIFFERENT GROUP SIZES

The third example we considered is with two groups that have different group sizes. Various samples were generated with a group size ratio of 1 to 5. So for 300 respondents, 50 would be in the first group and 250 would be in the second group. The group utilities were the same as those in Table 3. The sample sizes were also consistent with the previous experiment. Table 5 contains the results from this exercise.

Table 5
Results for 1:5 Ratio of Groups with Various Sample Sizes

		300	600	900	1800	3600
Untuned Results	Correlation	.995	.998	.999	.998	.999
	RMSE	2.86	1.92	1.93	1.83	1.41
	Avg. Bias	1.89	1.19	1.23	1.08	0.80
Aggregate Tuned Results	RMSE	1.03	0.98	1.01	1.01	0.98
	Avg. Bias	0.32	0.29	0.32	0.32	0.28
Individual Tuned Results	RMSE	0.75	0.70	0.68	0.70	0.65
	Avg. Bias	0.14	0.10	0.09	0.10	0.07

These results are similar to the previous results. Tuning for scale seems to remove most of the bias from the estimates. There is however a larger difference between the aggregate tuning and the individual tuning. This is because there is a large difference between the scale factor for the larger group and the smaller group. For the 3600 respondent simulation, the tuning multiplier for the large group has a mean of 0.77 and a variance of 0.02. The tuning scale multiplier for the smaller group has a mean of 1.27 with a variance of 0.02. If the sample were simply tuned at the aggregate level, the tuning coefficient would be 0.77. This would place the large group on the proper scale, but the smaller group would actually be tuned in the wrong direction. Instead of reducing the bias in the smaller group, the bias would be increased.

The bias appears to be fairly small and is smaller than the bias for the equal group sizes case. When examining the bias at the group level however, the bias for the larger group is almost non-existent and the bias for the smaller group accounts for almost all of the reported bias. This suggests that the smaller group is indeed shrunk toward the larger group while the larger group experiences very little shrinkage.

WEIGHTING IN CBC/HB

One commonly proposed solution to the shrinkage issue is to weight the respondents from each group so that the mean is an accurate representation of the proper sample sizes. This is especially appealing when the different group sizes are due to sampling constraints. It is a common practice in traditional analysis to weight the respondents to account for the disproportionate sampling. One possible solution is a simple weighting scheme for the prior distribution of the betas. This should allow the betas to shrink to a weighted mean instead of the raw mean. By default the estimates for alpha (the prior mean for the betas) are draws from a multivariate normal with mean $\bar{\beta}$ and variance D/n where $\bar{\beta}$ is the mean of the individual betas and D is the prior variance-covariance matrix for the sample. Instead draw alpha from a distribution:

$$\alpha \sim MVN\left(\sum_{i=1}^n w_i \beta_i, D/n\right)$$

where w_i is a weighting applied to the i th individual. A weight of $1/n$ would give the same solution as the standard CBC/HB algorithm.

Using a customized version of CBC/HB, we ran the previous two unequal group solution adjusting the weights to compare against the unweighted run. The results are presented in Table 6 below.

Table 6
Weighted Results

		300	600	900	1800	3600
Untuned Results	Correlation	.997	.997	.999	.998	.999
	RMSE	3.49	2.74	2.53	3.21	3.24
	Avg. Bias	2.50	1.87	1.71	2.22	2.30
Aggregate Tuned Results	RMSE	0.90	0.88	0.88	0.94	0.91
	Avg. Bias	0.19	0.17	0.20	0.22	0.18
Individual Tuned Results	RMSE	0.78	0.75	0.71	0.76	0.66
	Avg. Bias	0.15	0.12	0.09	0.12	0.09

There are a few surprising results in this test. The first is that the average bias for the untuned utilities actually increases as the sample size increases. This may be due to the increased shrinkage of the larger group as the smaller group exerts more weight. It also becomes even more important to tune the model. The variance of the estimates is inflated since the mean is adjusted. This leads to slightly greater overfitting and thus requires greater tuning.

Figure 3 is a comparison of the average bias of the unweighted and weighted results. The longer bars represent the untuned results. The weighted model appears to have a higher bias and that bias does not improve as the sample size increases. For the individually tuned models, the bias seems to be equivalent between the two methods. A closer look at the individual respondent's bias shows that the smaller group is slightly less biased while the larger group is slightly more biased.

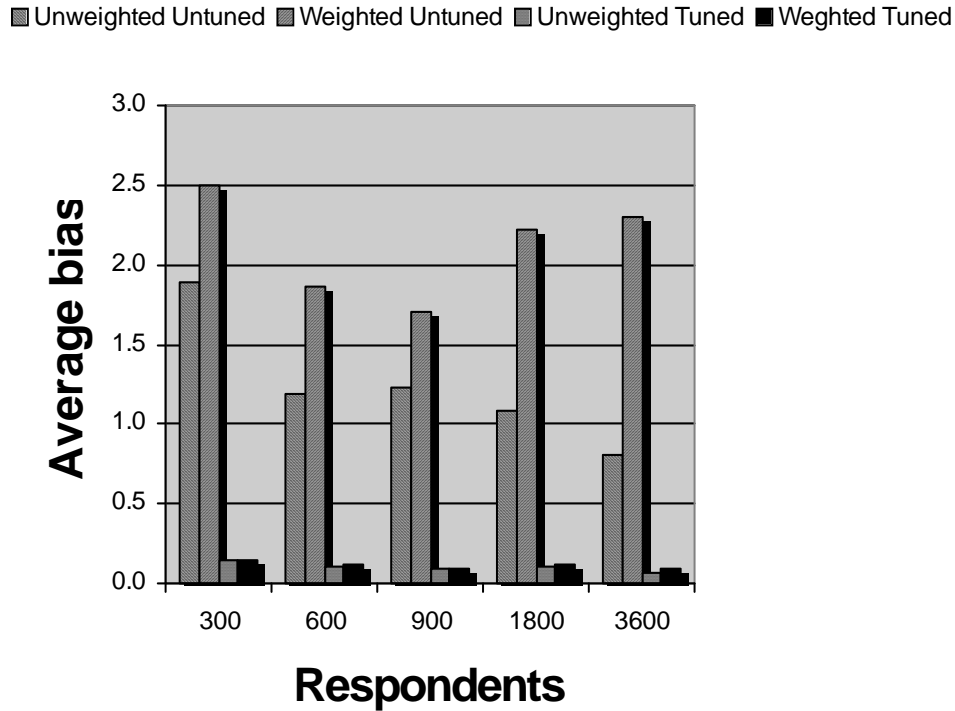


Figure 3
Weighted vs. Unweighted Average Bias

Since the results are generally used in a market simulator, the primary question is how this will affect the results of any simulations. To test this, a simulation run was constructed with three products. The first product appealed primarily to the smaller group, the second product appealed to the larger group, and the final product was a low share product that was equally appealing. The results are presented in Table 7 below.

From the results it appears that there is a little bit of bias due to the disproportionate shrinkage. As expected, product 1 does indeed have a smaller share than expected and product 2 has a larger share than expected—but the difference is small. The simulation results are from the 300 respondent utilities that showed the highest bias and thus should exhibit the largest amount of shrinkage. There does not seem to be much difference between the results, however, as none of the simulations show significant differences. The standard errors for the products are approximately 1.3, 1.5, and 0.5 respectively, so there is not a statistically significant difference between the results for the different estimation and tuning methods.

Table 7
Simulation Results 300 Respondents

	Actual	Raw Estimate	Aggregate Tuning w/o Weights	Individual Tuning w/o Weights	Aggregate Tuning with Weights	Individual Tuning with Weights
Product 1 (appeals to smaller group)	15.48	10.61	12.61	12.93	12.59	12.45
Product 2 (appeals to larger group)	83.06	87.73	85.33	85.46	85.58	85.89
Product 3 (neutral)	1.46	1.65	2.07	1.61	1.84	1.66
RMSE		3.90	2.14	2.02	2.22	2.40

A SECOND TEST OF WEIGHTING

In a second test of the weighting algorithm we tested a set of utilities that had more realistic utilities. One of the groups had a strong preference for a certain level such as brand and the other group was fairly ambivalent toward that level and attribute. This is the type of data that could potentially arise from a situation where the sample was a mix of customers and non-customers. The customers would likely have a stronger preference for their purchased brand where the non-customers would have only mild brand preferences. There were again just two groups and the total sample size was 1000 respondents for each cell. The proportion of the two groups was varied between an extreme of 100:900 to equal group sizes of 500:500 in 100-respondent increments. The report on the bias and the RMSE is found in Table 8.

Table 8
Second Test of Weighting

		100:900	200:800	300:700	400:600	500:500
Unweighted Individual Tuned Results	RMSE	0.65	0.66	0.68	0.69	0.67
	Avg. Bias	0.04	0.06	0.07	0.08	0.08
Weighted Aggregate Tuned Results	RMSE	0.65	0.65	0.66	0.66	0.65
	Avg. Bias	0.07	0.08	0.07	0.08	0.08
Weighted Individual Tuned Results	RMSE	0.68	0.67	0.69	0.69	0.68
	Avg. Bias	0.07	0.07	0.07	0.08	0.09

It appears that neither the weighting nor the tuning method affect the results in a significant way. The utilities were not nearly as widely separated, leading to less overfitting and therefore a smaller need to tune the exponent. From Table 8 it appears that the disproportionality of the samples has only a limited influence on the amount of bias in the estimates.

A market simulation run of the 100:900 study is presented in Table 9. This run should exhibit the most shrinkage since the 900 respondents should provide a lot more signal than the 100 respondents. Again there are three products and the simulation method is Randomized First Choice, but this time the products do not have a specific meaning.

Table 9
100:900 Respondents Simulation

	Actual	Raw Estimate	Aggregate Tuning w/o Weights	Individual Tuning w/o Weights	Aggregate Tuning with Weights	Individual Tuning with Weights
Product 1	37.99	34.98	35.49	36.02	38.37	38.55
Product 2	6.63	4.44	5.11	5.08	6.12	6.16
Product 3	55.37	60.58	59.39	58.90	55.52	55.29
RMSE		3.70	2.88	2.50	0.37	0.42

In this study the weighting does appear to provide a benefit to the simulations. It decreases the Root Mean Squared Error by about 2 points. As in the last study it does not appear that individually tuning the model provides much benefit, especially in the weighted case. The results suggest that there may be a benefit in weighting the model while the previous simulation found almost no effect. Additional research is needed to determine when weighting will make a difference.

CONCLUSIONS

The results from this exercise are inconclusive. Looking just at the measures of bias, it appears that the largest source is the inconsistent scale factor between the generated utilities and the estimated utilities. If it is suspected that the different groups have large differences in the utilities, then the scale factor should be adjusted at the group level at a minimum and at the individual level if possible. Tuning had the greatest single effect on the bias of the utility estimates. Weighting had minimal effect on the bias of the utilities, but had mixed results on the simulation tests. Because of the inconsistent results, additional investigation into weighted CBC/HB is needed. Currently if there are multiple groups in a sample and the sample size is large enough it is best to split the sample into the respective groups and estimate separate models

for each group. This will completely avoid any possible shrinkage issues. Even with weighting, the CBC/HB model will still shrink the results and lead to potentially biased outcomes.

Sentis and Li addressed a similar issue in the 2001 Sawtooth Software Conference (2001). Using data from actual studies they compared the hit rates between a combined estimations and a segmented estimation. They concluded that segmenting the sample did not provide any improvement in the hit rates. The total sample sizes were not especially large, ranging from 280 to 800 respondents. The results may have been different if the sample sizes for each segment were larger. One of the other conclusions that they reached was that it was most likely that there were no clearly differentiated groups. Real datasets often exhibit the “Watermelon Theory” of market segmentation. Different groups may be detected, but those differences are really based more on the segmentation strategy than actual clearly defined market segments. If the watermelon theory is indeed correct, then splitting the sample will not provide any real benefit. On the other hand it did not seem to hurt the results either, so it is largely a matter of choice.

One of the issues with the weighted implementation of CBC/HB is that it ignores some important aspects of the model that may prove important. It only adjusts the prior mean for the betas and ignores the effect that these adjustments will have on the prior variance. It also requires prior knowledge of the appropriate groups, which may not always be possible. It also does not incorporate the group information directly into the Bayesian model.

There are a couple of other possible solutions to the problem that are more promising, although a little more complicated to implement. The most common would be to incorporate the different group information into the upper level of the model as covariates. This is quite commonly done in other software packages like bayesm and WinBUGs, but is not currently possible in CBC/HB. This would in effect estimate a separate upper level model for each group and should most closely resemble the current approach of splitting the sample into groups prior to the run. The downside of this is that the run time for the estimation could be greatly increased as well as the number of parameters estimated.

Other approaches would involve changing the specification of the prior distribution. One possible distribution is the student’s-t distribution. Since it has wider tails than a normal distribution, it will cause the data to shrink less toward the mean. Another possible prior distribution is a mixture of normals model. This has also been used in practice and would not necessarily require the analyst to specify the groups in advance.

Sawtooth Software has so far avoided implementing more complex models. The CBC/HB software has tried to use the simplest model that can provide accurate results for our users. This has made CBC/HB more accessible to users as it has simplified the input requirements and allows it to work with a large variety of datasets. We are not aware of any studies that show strong gains in accuracy from more complicated modeling procedures. Because bias due to Bayesian shrinkage has been a common concern, however, we are planning to continue to look into this issue and are always working to improve our algorithms while at the same time keeping the software easy to use and broadly applicable.

REFERENCES

Sawtooth Solutions (Fall 2007), "2007 Customer Survey Results." Available at www.sawtoothsoftware.com/ssolutions.shtml.

Sentis, Keith and Lihua Li (2001), "One Size Fits All or Custom Tailored: Which HB Fits Better?" *Sawtooth Software Conference Proceedings*, Sawtooth Software, Sequim, WA.