PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

April 2003

2003 Sawtooth Software Conference Proceedings: Sequim, WA.

FOREWORD

We are pleased to present the proceedings of the tenth Sawtooth Software Conference, held in San Antonio, TX, April 15-17, 2003. The spring weather in San Antonio was gorgeous, and we thoroughly enjoyed the ambiance of the Riverwalk and the Alamo.

The focus of the conference was quantitative methods in marketing research. The authors were charged to deliver presentations of value to both the most and least sophisticated members of the audience. We were treated to a variety of topics, including discrete choice, conjoint analysis, MaxDiff scaling, latent class methods, hierarchical Bayes, genetic algorithms, data fusion, and archetypal analysis. We also saw some useful presentations involving case studies and validation for conjoint measurement.

Authors also played the role of discussant to another paper presented at the conference. Discussants spoke for five minutes to express contrasting or complementary views. Some discussants have prepared written versions of their comments for this volume.

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

Sawtooth Software

June, 2003

CONTENTS

EXPERIMENTS WITH CBC

CBC VALIDATION

CONJOINT ANALYSIS APPLICATIONS

CONJOINT ANALYSIS EXTENSIONS

DATA FUSION WITH CONJOINT ANALYSIS

ADVANCED TECHNIQUES

SUMMARY OF FINDINGS

Nearly two-dozen presentations were delivered at the tenth Sawtooth Software Conference, held in San Antonio, TX. We've summarized some of the high points below. Since we cannot possibly convey the full worth of the papers in a few paragraphs, the authors have submitted complete written papers within this 2003 Sawtooth Software Conference Proceedings.

The Internet: Where Are We? And Where Do We Go from Here? (Donna J. Wydra, TNS Intersearch): The internet is increasingly becoming a key tool for market researchers in data collection and is enabling them to present more interesting and realistic stimuli to respondents. In 2002, 20% of market research spending was accounted for by internet-based research. Some estimates project that to increase to 40% by 2004. Although the base of US individuals with access to the internet is still biased toward higher income and employment and lower age groups, the incidence is increasingly representative of the general population. Worldwide, internet usage by adults is highest in Denmark (63%), USA (62%), The Netherlands (61%), Canada (60%), Finland (59%) and Norway (58%).

Best practices for internet research include keeping the survey to 10 minutes or less and making it simple, fun, and interesting. Online open-ends are usually more complete (longer and more honest) than when provided via phone or paper-based modes. Donna emphasized that researchers must respect respondents, which are our treasured resource. Researchers must ensure privacy, provide appropriate incentives, and say "Thank You." She encouraged the audience to pay attention to privacy laws, particularly when interviewing children. She predicted that as broad-band access spreads, more research will be able to include video, 360-degree views of product concepts, and virtual shopping simulations. Cell phones have been touted as a new promising vehicle for survey research, especially as their connectivity and functionality with respect to the internet increases. However, due to small displays, people having to pay by the minute for phone usage, and the consumers' state of mind when using phones (short attention spans), this medium, she argued, is less promising than many have projected.

Sampling and the Internet (Expert Panel Discussion): This session featured representatives from three companies heavily involved in sampling over the internet: J. Michael Dennis (Knowledge Networks), Andrea Durning (SPSS MR, in alliance with AOL's Opinion Place), and Susan Hart (Synovate, formerly Market Facts). Two main concepts were discussed for reaching respondents online: River Sampling and Panels. River Sampling (e.g. Opinion Place) continuously invites respondents using banner ads having access to potentially millions of individuals. The idea is that a large river of potential respondents continually flows past the researcher, who dips a bucket into the water to sample a new set (in theory) of respondents each time. The benefits of River Sampling include its broad reach and ability to contact difficult to find populations.

In contrast to River Sampling, panels may be thought of as dipping the researcher's bucket into a pool. Respondents belong to the pool of panelists, and generally take multiple surveys per month. The Market Facts ePanel was developed in 1999, based on its extensive mail panel. Very detailed information is already known about the panelists, so much profiling information is available without incurring the cost of asking the respondents each time. Approximately 25% of the panel is replaced annually. Challenges include shortages among minority households and lower income groups, though data can be made more projectable by weighting. Knowledge

Networks offers a different approach to the internet panel: panelists are recruited in more traditional means and then given Web TVs to access surveys. Benefits include better representation of all segments of the US (including low income and education households). Among the three sources discussed, research costs for a typical study were most expensive for Knowledge Networks, and least expensive for Opinion Place.

Online Qualitative Research from the Participants' Viewpoint (Theo Downes-Le Guin, Doxus LLC): Theo spoke of a relatively new approach for qualitative interviewing over the internet called "threaded discussions." The technique is based on bulletin board web technology. Respondents are recruited, typically by phone, to participate in an on-line discussion over a few days. The discussion is moderated by one or more moderators and can include up to 30 participants.

Some of the advantages of the method are inherent in the technology, for example, there is less bias toward people who type faster and are more spontaneously articulate, as with one-session internet focus groups. Participants also indicate that the method is convenient because they can come and go as schedule dictates. Since the discussion happens over many days, respondents can consider issues more deeply and type information at their leisure. Respondents can see the comments from moderators and other participants, and respond directly to those previous messages. Theo explained that threaded discussion groups produce much more material that can be less influenced by dominant discussion members than traditional focus groups. However, like all internet methods, the method is probably not appropriate for low-involvement topics. The challenge, as always, is to scan such large amounts of text and interpret the results.

Scaling Multiple Items: Monadic Ratings vs. Paired Comparisons (Bryan Orme, Sawtooth Software): Researchers are commonly asked to measure multiple items, such as the relative desirability of multiple brands or the importance of product features. The most commonly used method for measuring items is the monadic rating scale (e.g. rate "x" on a 1 to 10 scale). Bryan described the common problems with these simple ratings scales: respondents tend to use only a few of the scale points, and respondents exhibit different scale use biases, such as the tendency to use either the upper part of the scale ("yea-sayers") or the lower end of the scale ("nay-sayers"). Lack of discrimination is often a problem with monadic ratings, and variance is a necessary element to permit comparisons among items or across segments on the items.

Bryan reviewed an old technique called paired comparisons that has been used by market researchers, but not nearly as frequently as the ubiquitous monadic rating. The method involves asking a series of questions such as "Do you prefer IBM or Dell?" or "Which is more important to you, clean floors or good tasting food?" The different items are systematically compared to one another in a balanced experimental plan. Bryan suggested that asking 1.5x as many paired comparison questions as items measured in a cyclical plan is sufficient to obtain reasonably stable estimates at the individual level (if using HB estimation). He reported evidence from two split-sample studies that demonstrated that paired comparisons work better than monadic ratings, resulting in greater between-item and between-respondent discrimination. The paired comparison data also had higher hit rates when predicting holdout observations.

* **Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation** (Steve Cohen, Consultant): Steve's presentation picked up where Bryan Orme's presentation left off, extending the argument against monadic ratings for measuring preferences for objects or importances for attributes, but focusing on a newer and more sophisticated method called Maximum Difference (Best/Worst) Scaling. MaxDiff was first proposed by Jordan Louviere in the early 90s, as a new form of conjoint analysis. Steve focused on its use for measuring the preference among an array of multiple items (such as brands, or attribute features) rather than in a conjoint context, where items composing a whole product are viewed conjointly.

With a MaxDiff exercise, respondents are shown, for example, four items and asked which of these is the most and least important/preferable. This task repeats, for a number of sets, with a new set of items considered in each set. Steve demonstrated that if four items (A, B, C, D) are presented, and the respondent indicates that A is best and B is worst, we learn five of the six possible paired comparisons from this task (A>B, A>C, A>D, B>D, C>D). Steve showed that MaxDiff can lead to even greater between-item discrimination and better predictive performance of holdout tasks than monadic ratings or even paired comparisons. Between-group discrimination was better for MaxDiff than monadic, but about on par with paired comparisons. Finally, Steve showed how using this more powerful tool for measuring importance of items can lead to better segmentation studies, where the MaxDiff tasks are analyzed using latent class analysis.

(* Most Valuable Presentation award, based on attendee ballots.)

The Predictive Validity of Kruskal's Relative Importance Algorithm (Keith Chrzan & Joe Retzer, Maritz Research, and Jon Busbice, IMS America): The authors reviewed the problem of multicollinearity when estimating derived importance measures (drivers) for product/brand characteristics from multiple regression, where the items are used as independent variables and some measure of overall performance, preference, or loyalty is the dependent variable. Multicollinearity often leads to unstable estimates of betas, where some of these actually can reflect a negative sign (negative impact on preference, loyalty, etc.) when the researcher hypothesizes that all attributes should necessarily have a positive impact.

Kruskal's algorithm involves investigating all possible orderings of independent variables and averages across the betas under each condition of entry. For example, with three independent variables A, B, and C, there are six possible orderings for entry in the regression model: ABC, ACB, BAC, BCA, CAB, and CBA. Therefore, the coefficient for variable A is the average of the partial coefficients for A when estimated within separate regression models with the following independent variables: (A alone, occurs 2x), (BA), (BCA), (CA), and (CBA). The authors showed greater stability for coefficients measured in this manner, and also demonstrated greater predictive validity in terms of hit rates for holdout respondents for Kruskal's importance measure as opposed to that from standard regression analysis.

New Developments in Latent Class Choice Models (Jay Magidson, Statistical Innovations, Inc., Thomas C. Eagle, Eagle Analytics, Inc., and Jeroen K. Vermunt, Tilburg University): Latent class analysis has emerged as an important and valuable way to model respondent preferences in ratings-based conjoint and CBC. Latent class is also valuable in more general contexts, where a dependent variable (whether discrete or continuous) is a function of single or multiple independent variables. Latent class simultaneously finds segments representing concentrations of individuals with identical beta weights (part worth utilities) and reports the beta weights by

segment. Latent class assumes a discrete distribution of heterogeneity as opposed to a continuous assumption of heterogeneity for HB.

Using output from a commercially available latent class tool called Latent GOLD Choice, Jay demonstrated the different options and ways of interpreting/reporting results. Some recent advances incorporated into this software include: ability to deal with partial- or full-ranks within choice sets; monotonicity constraints for part worths, bootstrap p-value (for helping determine the appropriate number of segments); inclusion of segment-based covariates; rescaled parameters and graphical displays; faster and better algorithms by switching to a Newton Raphson algorithm when close to convergence; and availability of individual coefficients (by weighting group vectors by each respondent's probability of membership). The authors reported results for a real data set in which latent class and HB had very similar performance in predicting shares of choice for holdout tasks (among holdout respondents). But, latent class is much faster than HB, and directly provides insights regarding segments.

Archetypal Analysis: An Alternative Approach to Finding and Defining Segments (Andy Elder, Momentum Research Group, and Jon Pinnell, MarketVision Research): The authors presented a method for segmentation called Archetypal Analysis. It is not a new technique, but it has yet to gain much traction in the market research community. Typical segmentation analysis often involves K-means clustering. The goal of such clustering is to group cases within clusters that are maximally similar within groups and maximally different between groups (Euclidean distance). The groups are formulated and almost always characterized in terms of their within-group means. In contrast, Archetypal Analysis seeks groups that are not defined principally by a concentration of similar cases, but that are closely related to particular extreme cases that are dispersed in the furthermost corners in the complex space defined by the input variables. These extreme cases are the archetypes. Archetypes are found based on an objective function (Residual Sum of Squares) and an iterative least squares solution.

The strengths of the method are that the approach focuses on identifying more "pure" types, and those pure types reflect "aspirational" rather than average individuals. Segment means from archetypal analysis can show more discrimination on the input variables than traditional cluster segmentation. However, like K-means cluster routines, archetypal analysis is subject to local minima. It doesn't work as well in high dimension space, and it is particularly sensitive to outliers.

Trade-Off vs. Self-Explication in Choice Modeling: The Current Controversy (Lawrence D. Gibson, Eric Marder Associates, Inc.): Choice models and choice experiments are vital tools for marketing researchers and marketers, Larry argued. These methods yield the unambiguous, quantitative predictions needed to improve marketing decisions and avoid recent marketing disasters. Larry described how Eric Marder Associates has been using a controlled choice experiment called STEP for many years. STEP involves a sticker allocation among competing alternatives. Respondents are randomly divided into groups in a classic experimental design, with different price, package, or positioning statements used in the different test cells. Larry noted that this single-criterion-question, pure experiment avoids revealing the subject of the study, as opposed to conjoint analysis that asks many criterion questions of each respondent.

Larry described a self-explicated choice model, called SUMM, which incorporates a complete 'map' of attributes and levels as well as each respondent's subjective perceptions of the alternatives on the various attributes. Rather than traditional rating scales, Eric Marder

Associates has developed an "unbounded" rating scale, where respondents indicate liking by writing (or typing) "L's" , or disliking by typing "D's" (as many "L" and "D" letters as desired). Each "L" indicates +1 in "utility" and every "D" -1. Preferences are then combined with the respondents' idiosyncratic perceptions of alternatives on the various features to produce an integrated choice simulator. Larry also shared a variety of evidence showing the validity of SUMM.

Larry argued that conjoint analysis lacks the interview capacity to realistically model the decision process. Collecting each respondent's subjective perceptions of the brands and using a complete "map" of attributes and levels usually eclipses the limits of conjoint analysis. If simpler self-explication approaches such as SUMM can produce valid predictions, then why bother with trade-off data, Larry challenged the audience. He further questioned why conjoint analysis continues to attract overwhelming academic support while self-explication is ignored. Finally, Larry invited the audience to participate in a validation study to compare conjoint methods with SUMM.

Perspectives Based on 10 Years of HB in Marketing Research (Greg M. Allenby, Ohio State University, and Peter Rossi, University of Chicago): Greg began by introducing Bayes theorem, which is a method for accounting for uncertainty forwarded in 1764. Even though statisticians found it to be a useful concept, it was impractical to use Bayes theorem in market research problems due to the inability to use integrate over so many variables. But, after influential papers in the 1980s and 1990s highlighting innovations in Monte Carlo Markov Chain (MCMC) algorithms, made possible because of the availability of faster computers, the Bayes revolution was off and running. Even using the fastest computers available to academics in the early 1990s, mid-sized market research problems took sometimes days or weeks to solve. Initially, reactions were mixed within the market research community. A reviewer for a leading journal called HB "smoke and mirrors." Sawtooth Software's own Rich Johnson was skeptical regarding Greg's results for estimating conjoint part worths using MCMC.

By the late 1990s, hardware technology had advanced such that most market research problems could be done in reasonable time. Forums such as the AMA's ART, Ohio State's BAMMCONF, and the Sawtooth Software Conference further spread the HB gospel. Software programs, both commercial and freely distributed by academics, made HB more accessible to leading researchers and academics. Greg predicted that over the next 10 years, HB will enable researchers to develop more rich models of consumer behavior. We will extend the standard preference models to incorporate more complex behavioral components, including screening rules in conjoint analysis (conjunctive, disjunctive, compensatory), satiation, scale usage, and inter-dependent preferences among consumers. New models will approach preference from the multitude of basic concerns and interests that give rise to needs. Common to all these problems is a dramatic increase in the number of explanatory variables. HB's ability to estimate truly large models at the disaggregate level, while simultaneously ensuring relatively stabile parameters, is key to making all this happen over the next decade.

Partial Profile Discrete Choice: What's the Optimal Number of Attributes (Michael Patterson, Probit Research, Inc. and Keith Chrzan, Maritz Research): Partial profile choice is a relatively new design approach for CBC that is becoming more widely used in the industry. In partial profile choice questions, respondents evaluate product alternatives on just a subset of the total attributes in the study. Since the attributes are systematically rotated into the questions,

each respondent sees all attributes and attribute levels when all tasks in the questionnaire are considered. Partial profile choice, it is argued, permits researchers to study many more attributes than would be feasible using the full-profile approach (due to a reduction in respondent fatigue/confusion).

Proponents of partial profile choice have generally suggested using about 5 attributes per choice question. This paper formally tested that guideline, by alternating the number of attributes shown per task in a 5-cell split-sample experiment. Respondents received either 3, 5, 7, 9 or 15 attributes per task, where 15 total attributes were being studied. A None alternative was included in all cases. The findings indicate the highest overall efficiency (statistical efficiency + respondent efficiency) and accuracy (holdout predictions) with 3 and 5 attributes. All performance measures, including completion rates, generally declined with larger numbers of attributes shown in each profile. The None parameter differed significantly, depending on the number of attributes shown per task. The authors suggested that including a None in partial profile tasks is problematic, and probably ill advised. After accounting for the difference in the None parameter, there were only a few statistically significant differences across the design cells for the parameters.

Discrete Choice Experiments with an Online Consumer Panel (Chris Goglia, Critical Mix): Panels of respondents are often a rich source for testing specific hypotheses through methodological studies. Chris was able to tap into an online consumer panel to test some specific psychological, experimental, and usability issues for CBC. For the psychological aspect, Chris tested whether there might be differences if respondents saw brand attributes represented as text or as graphical logos. As for the experimental aspects, Chris tested whether "corner prohibitions" lead to more efficient designs and accurate results. Finally, Chris asked respondents to evaluate their experience with the different versions, to see if these manipulations altered the usability of the survey. The subject matter of the survey was choices among personal computers for home use.

Chris found no differences in the part worths or internal reliability whether using brands described by text or pictures. "Corner prohibitions" involve prohibiting combinations of the best levels and worst levels for two a priori ordered attributes, such as RAM and Processor Speed. For example, 128MB RAM is prohibited with 1 GHz speed, and 512MB RAM is prohibited with 2.2 GHz speed. Corner prohibitions reduce orthogonality, but increase utility balance within choice tasks. Chris found no differences in the part worths or internal reliability with corner prohibitions. Other interesting findings were that self-explicated importance questions using a 100-point allocation produced substantially different results for the importance of brand (relative to the other attributes) than importance for brand (relative to those same attributes) derived from the CBC experiment. However, self-explicated ratings of the various brands produced very similar results as the relative part worths for those same brands derived from the CBC experiment. These results echo earlier cautions by many researchers regarding the value of asking a blanket "how important is <insert attribute>" question.

How Few Is Too Few?: Sample Size in Discrete Choice Analysis (Robert A Hart, Jr., The Gelb Consulting Group, Inc., and Michael Patterson, Probit Research, Inc.): Researchers have argued that Choice-Based Conjoint (CBC) requires relatively larger sample sizes to stabilize the parameters relative to ratings-based conjoint methods. Given the many benefits of CBC, the sensitivity of its use to sample size seems an important issue. Mike reviewed previous work by

Johnson and Orme that had suggested that, if assuming aggregate analysis, doubling the number of tasks each respondent completes is roughly equal in value to doubling the number of respondents. However, this conclusion did not consider heterogeneity and more recent estimation methods such as Latent Class and HB.

Mike presented results for both synthetic (computer generated) and real data. He and his co-author systematically varied the number of respondents and tasks per respondent, and compared the stability of the parameters across multiple random "draws" of the data. They found that the Johnson/Orme conclusion essentially held for aggregate logit conditions. They concluded that researchers could obtain relatively stable results in even small (n=50) samples, given that respondents complete a large enough number of choice tasks. They suggested that further research should be done to investigate the effects of heterogeneity, and the effects of partial profile CBC tasks on parameter stability.

Validation and Calibration of CBC for Pricing Research (Greg Rogers, Procter & Gamble, and Tim Renken, Coulter/Renken): The authors presented results from a series of CBC studies that had been compared to actual market share and also econometric models (marketing mix modeling) of demand for packaged goods at P&G. The marketing mix models used multiple regression, modeling weekly volume as a function of SKU price, merchandising variables, advertising and other marketing activities. The models controlled for cross-store variation, seasonality and trend. The authors presented share predictions for CBC (after adjusting for distributional differences) versus actual market shares for washing powder, salted snacks, facial tissue, and potato crisps. In some cases, the results were extremely similar; in other cases the results demonstrated relatively large differences.

The authors next compared the sensitivity of price predicted by CBC to those from the marketing mix models. After adjusting the scale factor (exponent), they found that CBC was too oversensitive to price decreases (but not price increases). Greg and Tim also calculated a scalar adjustment factor for CBC as a function of marketing mix variables (regression analysis, where the dependent variable was the difference between predicted and actual sales). While this technique didn't improve the overall fit of the CBC relative to an aggregate scalar, it shed some light on which conditions may cause CBC predictions to deviate from actual market shares. Based on the regression parameters, they concluded that CBC understates price sensitivity of big-share items, overestimates price sensitivity of items that sell a lot on deal, and overestimates price sensitivity in experiments with few items on the shelf. Despite the differences between CBC and actual market shares, the Mean Absolute Error (MAE) for CBC predictions versus actual market shares was 4.5. This indicates that CBC's predictions were on average 4.5 share points from actual market shares, and in the opinion of some members of the audience that chimed in with their assessments, reflects commendable performance for a survey-based technique.

Determinants of External Validity in CBC (Bjorn Arenoe, SKIM Analytical/Erasmus University Rotterdam): Bjorn pointed out that most validation research for conjoint analysis has used internal measures of validity, such as predictions of holdout choice tasks. Only a few presentations at previous Sawtooth Software Conferences have dealt with actual market share data. Using ten data sets covering shampoo, surface cleaner, dishwashing detergent, laundry detergent and feminine care, Bjorn systematically studied which models and techniques had the greatest systematic benefit in predicting actual sales. He covered different utility estimation

methods (logit, HB, and ICE), different simulation models (first choice, logit, RFC) and correctional measures (weighting by purchase frequency, and external effects to account for unequal distribution).

Bjorn found that the greatest impact on fit to market shares was realized for properly accounting for differences in distributional effects using external effects, followed by tuning the model for scale factor (exponent). There was just weak evidence that RFC with its attribute-error correction for similarity outperformed the logit simulation model. There was also only weak evidence for methods that account for heterogeneity (HB, ICE) over aggregate logit. There was no evidence that HB offered improvement over ICE, and no evidence that including weights for respondents based on stated purchase volumes increased predictive accuracy.

Life-Style Metrics: Time, Money, and Choice (Thomas W. Miller, Research Publishers, LLC): The vast majority of product features research focuses on the physical attributes of products, prices, and the perceived features of brands, but according to Tom, we as researchers hardly ever bother to study how the metric of time factors into the decision process. Tom reviewed the economic literature, specifically the labor-leisure model, which explains each individual's use of a 24-hour day as a trade off between leisure and work time.

In some recent conjoint analysis studies, Tom has had the opportunity to include time variables. For example, in a recent study regarding operating systems, an attribute reflecting how long it took to become proficient with the software was included. Another study of the attributes students trade off when considering an academic program included variations in time spent in classroom, time required for outside study, and time spent in a part-time job. Tom proposed that time is a component of choice that is often neglected, but should be included in many research studies.

Modeling Patient-Centered Health Services Using Discrete Choice Conjoint and Hierarchical Bayes Analyses (Charles E. Cunningham, Don Buchanan, & Ken Deal, McMaster University): CBC is most commonly associated with consumer goods research. However, Charles showed a compelling example for how CBC can be used effectively and profitably in the design of children's mental health services. Current mental health service programs face a number of problems, including low utilization of treatments, low adherence to treatment, and high drop-out rate. Designing new programs to address these issues requires a substantial investment of limited funds. Often, research is done through expensive split sample tests where individuals are assigned to either a control or experimental group, where the experimental group reflects a new health services program to be tested, but for which very little primary quantitative research has gone into designing that new alternative.

Charles presented actual data for a children's health care program that was improved by first using CBC analysis to design a more optimal treatment. By applying latent class to the CBC data, the authors identified two strategically important (and significantly different) segments with different needs. Advantaged families wanted a program offering a "quick skill tune up" whereas high risk families desired more "intensive problem-focused" programs with highly experienced moderators. The advantaged families preferred meeting on evenings and Saturdays, whereas unemployed high risk families were less sensitive to workshop times. There were other divergent needs between these groups that surfaced. The predictions of the CBC and segmentation analysis were validated using clinic field trials and the results of previously conducted studies in which families were randomly assigned to either the existing program or

programs consistent with parental preferences. As predicted, high risk families were more likely to enroll in programs consistent with preferences. In addition, participants attended more sessions, completed more homework, and reported greater reductions in child behavior problems at a significantly reduced relative cost versus the standard program ($19K versus $120K).

Complementary Capabilities for Choice, and Perceptual Mapping Web Data Collection (Joseph Curry, Sawtooth Technologies, Inc.): Joe described how advancing technology in computer interviewing over the last few decades has enabled researchers to do what previously could not be done. While many of the sophisticated research techniques and extensions that researchers would like to do have been ported for use on the Web, other capabilities are not yet widely supported. Off-the-shelf web interviewing software has limitations, so researchers must choose to avoid more complicated techniques, wait for new releases, or customize their own solutions.

Joe showed three examples involving projects that required customized designs exceeding the capabilities of most off-the-shelf software. The examples involved conditional pricing for CBC (in which a complicated tier structure of price variations was prescribed, depending on the attributes present in a product alternative), visualization of choice tasks (in which graphics were arranged to create a "store shelf" look), and randomized comparison scales (in which respondents rated relevant brands on relevant attributes) for adaptive perceptual mapping studies. In each case, the Sensus software product (produced by Joe's company, Sawtooth Technologies) was able to provide the flexibility needed to accommodate the more sophisticated design. Joe hypothesized that these more flexible design approaches may lead to more accurate predictions of real world behavior, more efficient use of respondents' time, higher completion rates, and happier clients.

Brand Positioning Conjoint: The Hard Impact of the Soft Touch (Marco Vriens & Curtis Frazier, Millward Brown IntelliQuest): Most conjoint analysis projects focus on concrete attribute features and many include brand. The brand part worths include information about preference, but not why respondents have these preferences. Separate studies are often conducted to determine how soft attributes (perhaps more associated with perceptual/imagery studies) are drivers (or not) of brand preference. Marco and Curtis demonstrated a technique that bridges both kinds of information within a single choice simulator. The concrete features are measured through conjoint analysis, and the brand part worths from the conjoint model become dependent variables in a separate regression step that finds weights for the soft brand features (and an intercept, reflecting the unexplained component) that drive brand preference. Finally, the weights from the brand-drivers are included as additional variables within the choice simulator.

The benefits of this approach, the authors explained, are that it includes less tangible brand positioning information, providing a more complete understanding of how consumers make decisions. The drawbacks of the approach, as presented, were that the preferences for concrete attributes were estimated at the individual level, but the brand drivers were estimated as aggregate parameters. Discussion ensued directly following the paper regarding how the brand drivers may be estimated at the individual-level using HB, and how the concrete conjoint attributes and the weights for the soft imagery attributes might be estimated simultaneously, rather than in two separate steps.

Combining Self-Explicated and Experimental Choice Data (Amanda Kraus & Diana Lien, Center for Naval Analyses, and Bryan Orme, Sawtooth Software): The authors described a research project to study reenlistment decisions for Navy personnel. The sponsors were interested in what kinds of non-pay related factors might increase sailors' likelihood of reenlisting. The sponsors felt that choice-based conjoint was the proper technique, but wanted to study 13 attributes, each on 4 levels. Furthermore, obtaining stable individual-level estimates was key, as the sponsors required that the choice simulator provide confidence interval estimates in addition to the aggregate likelihood shares. To deal with these complexities, the authors used a three-part hybrid CBC study.

In the first section, respondents completed a self-explicated preference section identical to that employed in the first stage of ACA (respondents rate levels within attributes, and the importance of each attribute). In the second stage, respondents were given 15 partial-profile choice questions, each described using 4 of the attributes studied (without a "would not reenlist" option). In the final section, nine near-full profile CBC questions were shown (11 of the 13 attributes were displayed, due to screen real estate constraints), with a "would not reenlist" option. The authors tried various methods of estimation (logit, latent class, and HB), and various ways of combining the self-explicated, partial-profile CBC, and near-full profile CBC questions. Performance of each of the models was gauged using holdout respondents and tasks. The best model was one in which the partial-profile and near-full profile tasks were combined within the same data set, and individual-level estimates were estimated using HB, without any use of the self-explicated data. All attempts to use the self-explicated information did not improve prediction of the near-full profile holdout CBC tasks. Combining partial-profile and full-profile CBC tasks is a novel idea, and leverages the relative strengths of the two techniques. Partial-profile permits respondents to deal with so many attributes in a CBC task, and the full-profile tasks are needed for proper calibration of the None parameter.

Creating a Dynamic Market Simulator: Bridging Conjoint Analysis across Respondents (Jon Pinnell & Lisa Fridley, MarketVision Research): The issue of missing data is common to many market research problems, though not usually present with conjoint analysis. Jon described a project in which after the conjoint study was done, the client wanted to add a few more attributes to the analysis. The options were to redo the study with the full set of attributes, or collect some more data on a smaller scale with some of the original attributes plus the new attributes, and bridge (fuse) the new data with the old.

Typical conjoint bridging is conducted among the same respondents, or relies on aggregate level estimation. However, Jon's method used individual-level models and data fusion/imputation. Imputation of missing data is often done through mean substitution, hot deck, or model-based procedures (missing value is a function of other variables in the data, such as in regression). To evaluate the performance of various methods, Jon examined four conjoint data sets with no missing information, and randomly deleted some of the part worth data in each. He found that the hot-deck method worked consistently well for imputing values nearest to the original data, resulting in market simulations approximating those of the original data. The "nearest neighbor" hot-deck method involves scanning the data set to find the respondent or respondents that on common attributes most closely match the current respondent (with the missing data), and using the mean value from that nearest neighbor(s). Jon tried imputing the mean of the nearest neighbor, two nearest neighbors, etc. He found consistently better results when imputing the mean value from the four nearest neighbors.

Using Genetic Algorithms in Marketing Research (David G. Bakken, Harris Interactive): There are many kinds of problems facing market researchers that require searching for optimal combinations of variables in a large and complex search space, David explained. Common problems include conjoint-based combinatorial/optimization problems (finding the best product(s), relative to given competition), TURF and TURF-like combinatorial problems (e.g. find the most efficient set of six ice cream flavors such that all respondents find at least one flavor appealing), Non-linear ROI problems (such as in satisfaction/loyalty research), target marketing applications, adaptive questionnaire design, and simulations of market evolution.

Genetic Algorithms (GA) involve ideas from evolutionary biology. In conjoint analysis problems, the product alternatives are the "chromosomes," the attributes are the "genes," and the levels the attributes can assume are "alleles." A random population of chromosomes is generated, and evaluated in terms of fitness (share, etc.). The most fit members "mate" (share genetic information through random crossover and mutation) and produce new "offspring." The least fit are discarded, and the sequence repeats, for a number of generations. David also mentioned simpler, more direct routines such as hill-climbing, which are much quicker, but more subject to local minima. David suggested that GAs may be particularly useful for larger problems, when the search space is "lumpy" or not well understood, when the fitness function is "noisy" and a "good enough" solution is acceptable (in lieu of a global optimum).

Adaptive Choice-Based Conjoint (Rich Johnson, Sawtooth Software, Joel Huber, Duke University, and Lynd Bacon, NFO WorldGroup): There have been a number of papers in the literature on how to design CBC tasks to increase the accuracy of the estimated parameters. Four main criteria for efficient choice designs are: level balance, orthogonality, minimal overlap, and utility balance. These cannot be simultaneously satisfied, but a measure called D-efficiency appropriately trades off these opposing aims. D-efficiency is proportional to the determinant of the information matrix for the design.

The authors described a new design approach (ACBC) that uses prior utility information about the attribute levels to design new statistically informative questions. The general idea is that the determinant of the information matrix can be expressed as the product of the characteristic roots of the matrix, and the biggest improvement comes from increasing the smallest roots. Thus, new choice tasks with design vectors that mirror the characteristic vectors corresponding to the smallest roots are quite efficient in maximizing precision. In addition to choosing new tasks in this way, additional utility balance can be introduced across the alternatives within a task by swapping levels. The authors conducted a split-sample study (n=1099, using a web-based sample from the Knowledge Networks panel) in which respondents received either traditional CBC designs, the new adaptive CBC method just described, or that new method plus 1 or 2 swaps to improve utility balance. The authors found that the ACBC (with no additional swaps for utility balance) led to improvements in share predictive accuracy of holdout choice tasks (among holdout respondents). There were little or no differences in the treatments in terms of hit rates. The authors emphasized that when priors are used to choose the design, and particularly when used for utility balancing, the information in the data is reduced, but can be re-introduced with monotonicity constraints during part worth estimation. To do so, they used HB estimation subject to customized (within respondent) monotonicity constraints.

LEVERAGING THE INTERNET

THE INTERNET: WHERE ARE WE? AND, WHERE DO WE GO FROM HERE?

DONNA J. WYDRA
TNS INTERSEARCH

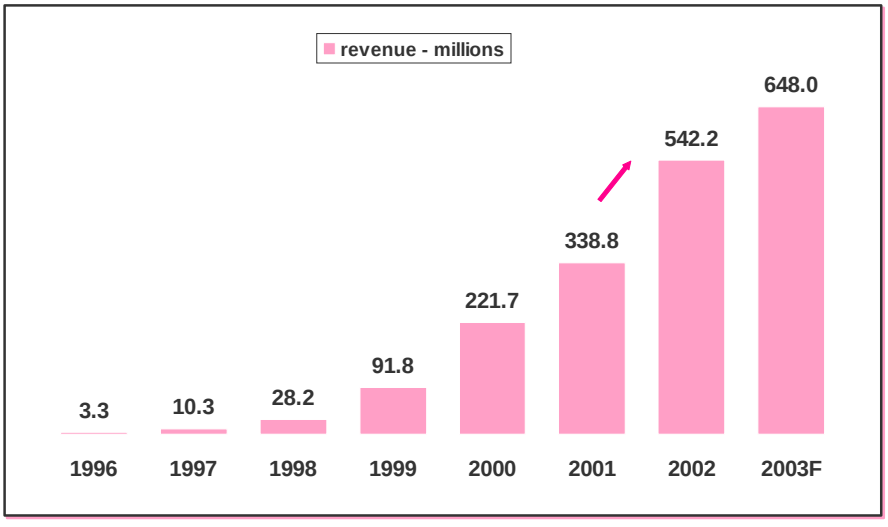
AGENDA

- Where have we been
- Where are we now
- A quick update
- Ten lessons learned
- Where are we going
- What's next

…for internet research.

WHERE HAVE WE BEEN

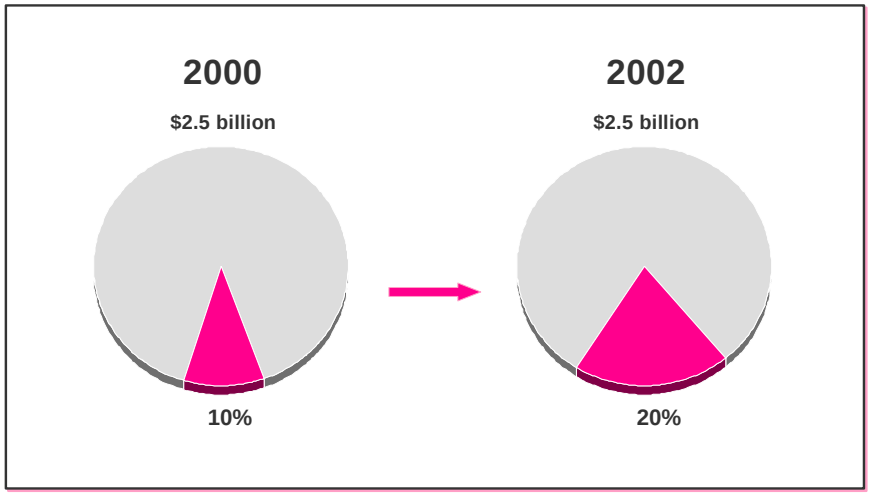
In terms of trends in online research, revenue increased 60% in 2002. Following four years of exponential growth through the mid-Nineties, the dawn of the millennium marked a decrease in exponential expenditures for online research. While not yet at saturation, forecasted 2003 online research revenue is expected to grow by 20% over last year.



Source: Inside Research, January 2003, volume 14, number 1

WHERE ARE WE NOW: AN UPDATE

The chart below shows the % of total U.S. market research spending accounted for by internet-based research.

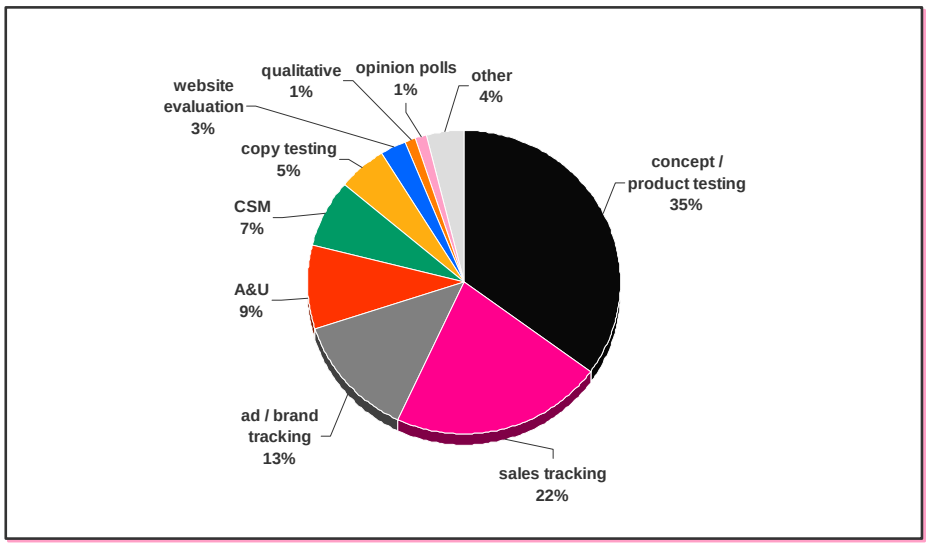


Source: Inside Research, January 2003, volume 14, number 1

Online is estimated to be approximately 40% of all MR spending in 2004.

In a flat-to-shrinking overall research expenditure environment since 2000, the growth of internet methodologies for capturing data nevertheless has been exponentially increasing as a proportion of total research dollars spent. Forecasts show it doubling again in two years.

Internet-Based Market Research Spending by Type

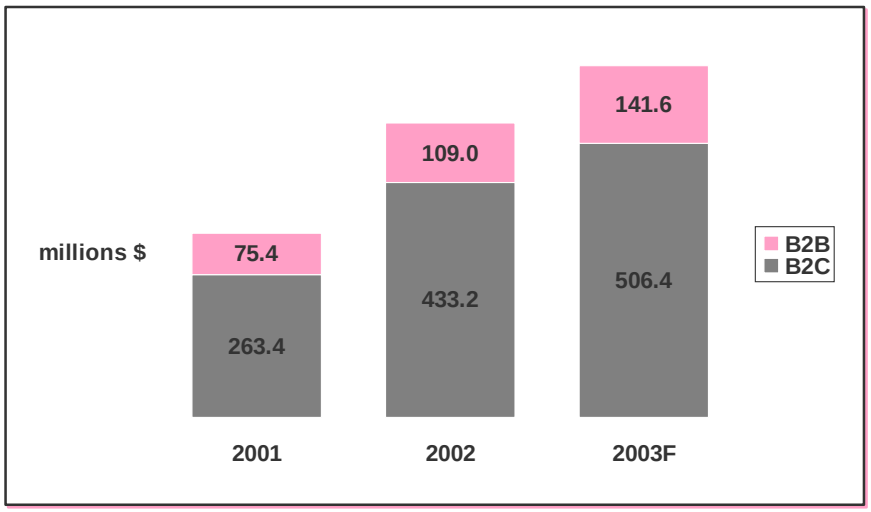


Source: Inside Research, January 2003, volume 14, number 1

As this chart shows, just over one-third of online research is allocated towards product and concept testing. The graphics-rich capabilities of internet research make it a fertile ground for evaluation of new products and concepts. Further, mass distribution of surveys and simultaneous administration give online the speed advantage over in-person or mail out administrations. Similarly, economies of scale make online clearly advantageous for ongoing tracking research, currently just over one-third of online research varieties.

Online business to business research holds steady at about twenty-percent of total combined B2B & B2C research.

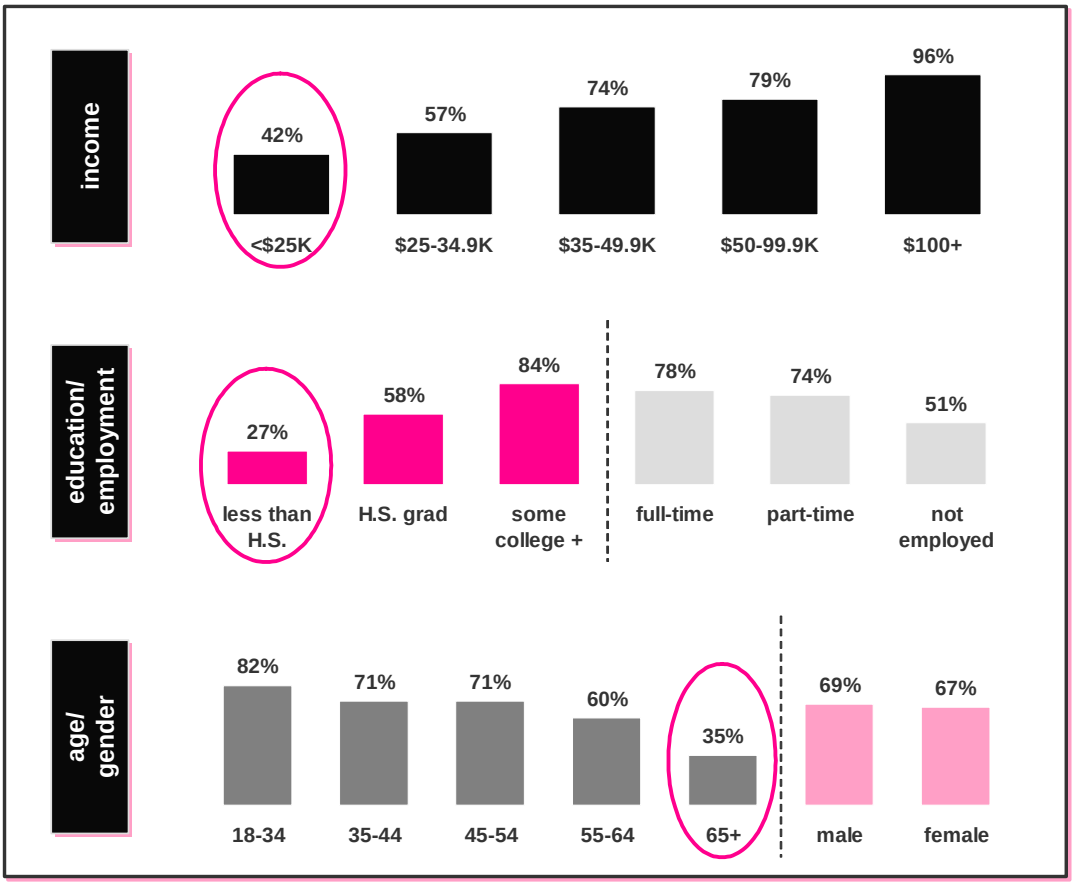
B2C versus B2B Internet Revenue



Source: Inside Research, January 2003, volume 14, number 1

While forecasted B2C online research growth is predicted to diminish markedly from 64% in 2002 to 17% in 2003, the proportion of B2B research versus total is projected to hold steady at 20% over the next year. Online business to business research is also predicted to grow more slowly in 2003 at 30% versus 45% in 2002, but not nearly as dramatically as business to consumer research spending for the same period.

U.S. Internet Users

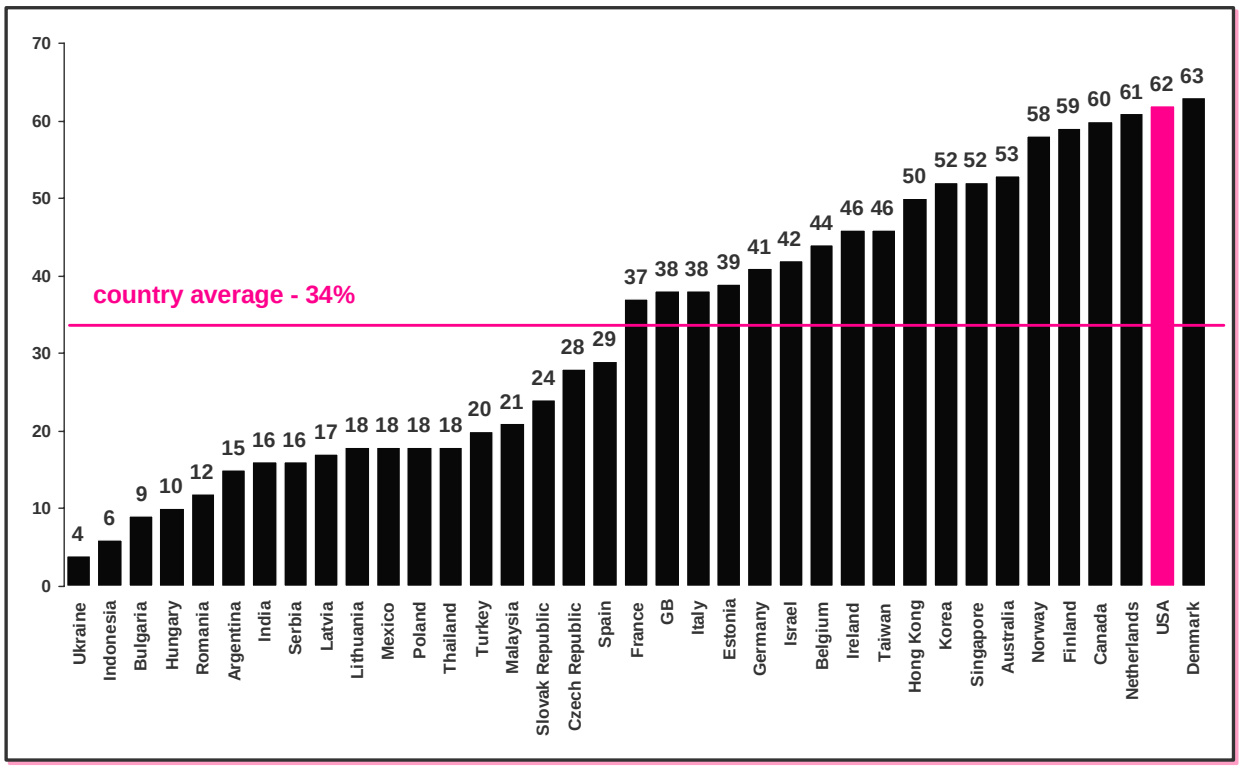


Source: TNS Intersearch Express Omnibus 2/03

Online representation for lowest income, lowest education and oldest Americans lags. Conversely, highest income, highest educated, and youngest Americans are strongly represented online.

This chart shows the percentage of the population who have personally used the internet during the past month.

Internet Users across the World



Source: TNS Interactive Global eCommerce Report, 2002

One third of the world's population has used the internet in the last month. Internet usage leads for northern European countries, US & Canada, and Asia-Pacific countries. Eastern European countries and India lag below average for internet usage.

TEN LESSONS LEARNED

#1: Sample options abound

Access to audiences you need to reach:

- large (representative) internet panels
- specialized, targeted panels
- shared
- proprietary
- many options for targeted sample frames via partnership with web communities
- "river" methodology
- online recruitment (pop-up surveys, banner ads, site links)

- client opt-in e-mail lists
- phone or mail recruitment

keep an eye on quality!

As internet penetration has increased, so have opportunities to capture these respondents. More research suppliers are building panels with multiple means for expansion. For example, specialized websites which already appeal to hard-to-read, low incidence targets are including opt-in methods for their visitors to participate in research. Further, high traffic web portals like MSN have partnered with existing suppliers like Greenfield Online to develop a "river" methodology for increasing panel sizes and individual survey respondents. This methodology leverages the high traffic of existing site visitors to capture respondents via an every *n*th-visitor pop-up windows, banner ads, or sweepstakes links on the main portal page to recruit. As with any sampling methodology, however, care must be taken to optimize the ability to project to the census under study.

#2: Keep it short, simple (& interesting)

- rule of thumb: 10 minutes or less
    - time is a precious commodity, especially online
    - quit rates rise as surveys grow
    - incent for long and/or 'grueling' surveys

Completion rates drop drastically at ten minutes and longer; in fact, online surveys longer than 20 minutes dwindle to 15% or less of qualified respondents without sufficient incentives. Recent incentive research shows that sweepstakes can be as effective as individual incentives for surveys longer than 10 minutes. Easier administration affords sweepstakes an advantage over individual incentives.

- simplicity
    - instructions succinct
    - avoid large scales
    - avoid scrolling

Like any self-administered survey, internet instructions must be short but clear to avoid confusion. Because web-enabled surveying allows for instantaneous response checking, errors can be minimized. Without sufficient explanations on complex questions, however, respondent frustration can diminish completion rates.

As a rule of thumb, seven-point likert scales should be the maximum for questions if the response scale is shown completely on the screen. Computer screen "real estate" is about 800x600 pixels for 90% of US PC users, which is diminished slightly by browser scroll bars and windows. If only part of a response or question grid is visible on screen, response bias can be introduced.

- make surveys more fun and interesting

#3: Use graphics wisely

- graphics can bring higher quality
    - clarification for respondents
    - less assumption
    - truer data

Especially for product and concept tests, the ability to bring full-color graphics to the respondent is a clear online benefit. Further, multiple views can be afforded when introducing new packaging, placements, or prototypes.

- image quality is key
    - may impact appeal
    - clarify image 'quality' from 'appeal'

Sufficient explanation must be given to respondents when introducing sketches, unfinished concepts, or animatic commercials, such that the appeal of the graphic is not confounded with the solicited metric. The cleanest read for likeability, purchase intent, etc., is given by screening out respondents who describe problems viewing the graphic in separate, additional questions.

- distracts from 'work' of survey

#4: Leverage online open ends

- more mentions online
    - one more mention online vs offline
- rich verbiage
    - gets to consumer language
    - assistance with copy
- more honest
    - gather negative feedback
- program probes into questionnaire
    - ROR proves additional mentions

Online administration also proves a rich ground for open-ended responses, in that the anonymity of administration gives the respondent a feeling of freedom to say more than may generally be given via a traditional in-person or CATI method. The caveat associated with this finding is that the response bias associated with providing "pleasing" answers to the interviewer is also gone: online respondents are more likely to give negative feedback due expressly to the anonymous nature of internet surveys. In the context of an ongoing tracker, this can temporarily mean that a drop in certain metrics might be expected online. The savvy researcher moving to the internet arena for tracking research will always conduct simultaneous, parallel tests to understand and calibrate for this effect.

#5: Leverage economies of scale

Leveraging large sample sizes:

- finer statistical differences
- analysis of subgroups
- analytic needs
    - conjoint/choice
    - segmentation
    - hybrid methodologies
- spectra/prizm coding of results
- recruitment for future research needs

Since online administration has, in general, reduced the overall cost per complete, larger sample sizes can be afforded for the same or fewer research dollars. This means that more attention can be paid to the sampling plan and quota schemes to dive deeper than before into sub-segments of the sample. Another benefit of this economy is the opportunity for more sophisticated statistical techniques. Once confined only to costly in-person administration with multiple cards, conjoint and discrete choice methods can be administered online more efficiently with sufficient pre-testing. However, just as larger sample sizes allow finer statistical precision to reveal significant differences at a given alpha error level, care must be given to balance against the power of a test and its trade offs. Evaluate the risks associated with insufficient statistical power: can the researcher afford to throw out a good idea overlooked by too much precision and not enough power?

#6: Respect respondents

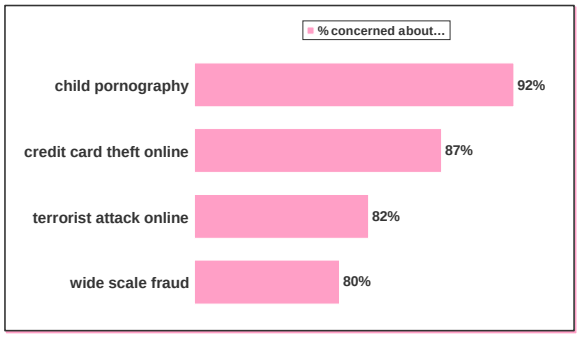
- a treasured resource
    - still motivated to 'make a difference'
    - short attention span
    - burned by offline research; prefer online
- guarantee their rights
    - confidentiality
    - privacy
- preserve respondent experience
    - ensure willingness to participate
    - don't wear out your welcome
    - say 'thank you'
    - provide support
    - appropriate incentive programs

The online respondent is not necessarily the same type of respondent who might volunteer for phone or other research methods. The rise of internet usage has witnessed the simultaneous abbreviation of participants' attention spans. Respondents expect fewer questions per page of online survey. However, since internet surveys are self-administered, researchers can expect a 15%-30% decrease in the time expected to run an identical phone survey.

Panel management systems have become more sophisticated with proliferation of online sample streams. Respondents expect full identity privacy associated with their emails and individual answers. Panel management should not be taken for granted: just as participation must be fully "opt-in," the right to "opt-out" is crucial. Further, survey invitation rates should be equalized to minimize wear-outs from too-frequent invitations versus defection rates from too-rare invitations.

#7: Keep it secure

Security concerns are real for internet users:



Source: PEW Internet Tracking Report

Consumers:

- across the world, consumers seek privacy
- assure their trust
    - strict security guidelines
    - place privacy statement on website

Surveys:

- must secure
    - admittance to survey
    - content and logic
    - physical environment

Concepts:

- marketers need security too

Mass, simultaneous survey administration for online new product and concept testing means that marketers must be more careful to protect their ideas from reaching competitors. More

sophisticated programming methods are being used to accomplish these goals: competitive domain name filtering for respondents, zip code filtering, and disabling graphical concept copying / printing from on screen can all work in tandem for marketer security.

#8: Know the laws

- vary dramatically by geography
    - US: more laws coming soon
    - Europe: stricter than US
- COPPA
    - Children's Online Privacy Protection Act
- Spam police
    - not the law, but….

Respondent "opt-in" & "opt-out" policies are paramount to adhere to existing online research standards. Further, online snowball sampling methods are frowned on, if not illegal, in some geographies. Laws for the online landscape are developing and changing rapidly. Again, savvy suppliers are proactive. Today's guidelines could be tomorrow's laws.

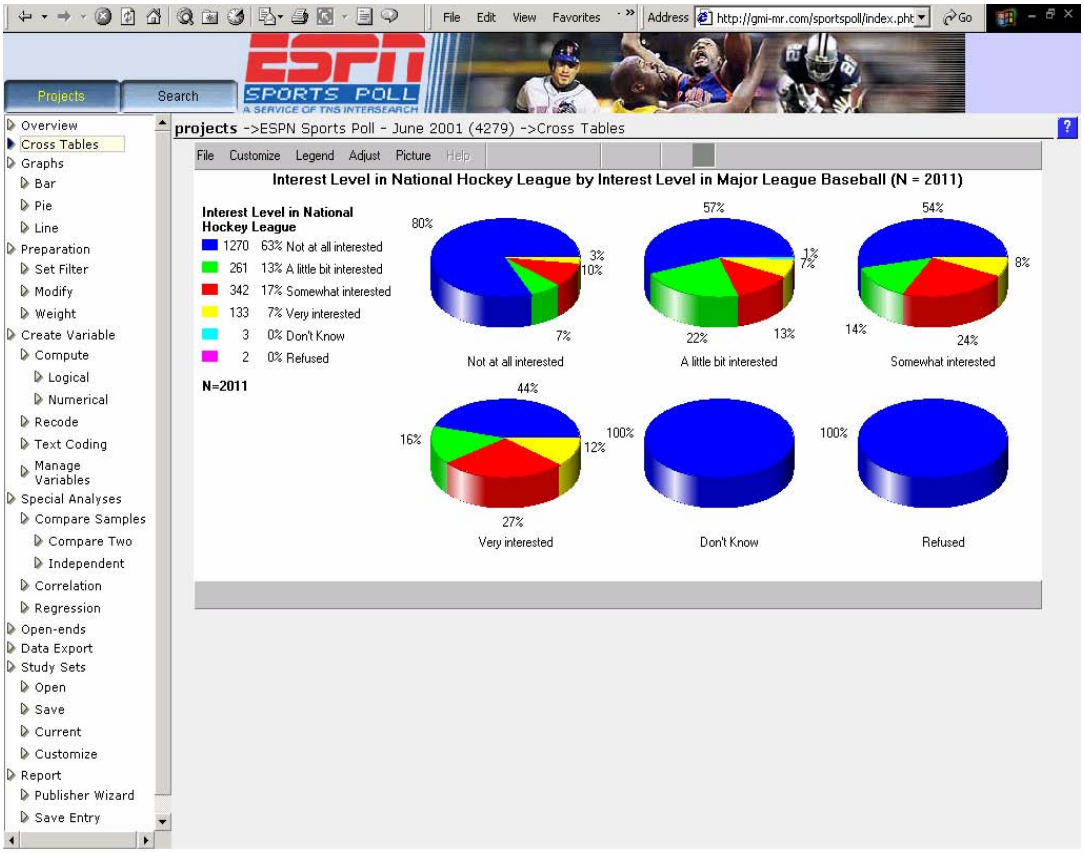
#9: Online data delivery & management

- Makes data
    - available
    - manageable
    - flexible
    - consistent
- Common platform across
    - business units; companies
    - countries
    - multiple data sources
- Reign in multiple data sources
    - surround traditional data delivery with contextual, relevant information

Just as online survey administration mushrooms, online data, results, and analytic capabilities are growing. More research suppliers are making secure internet client portals for final data delivery available. These allow for cross-platform compatibility and even decrease email traffic by housing results data at a centralized server.

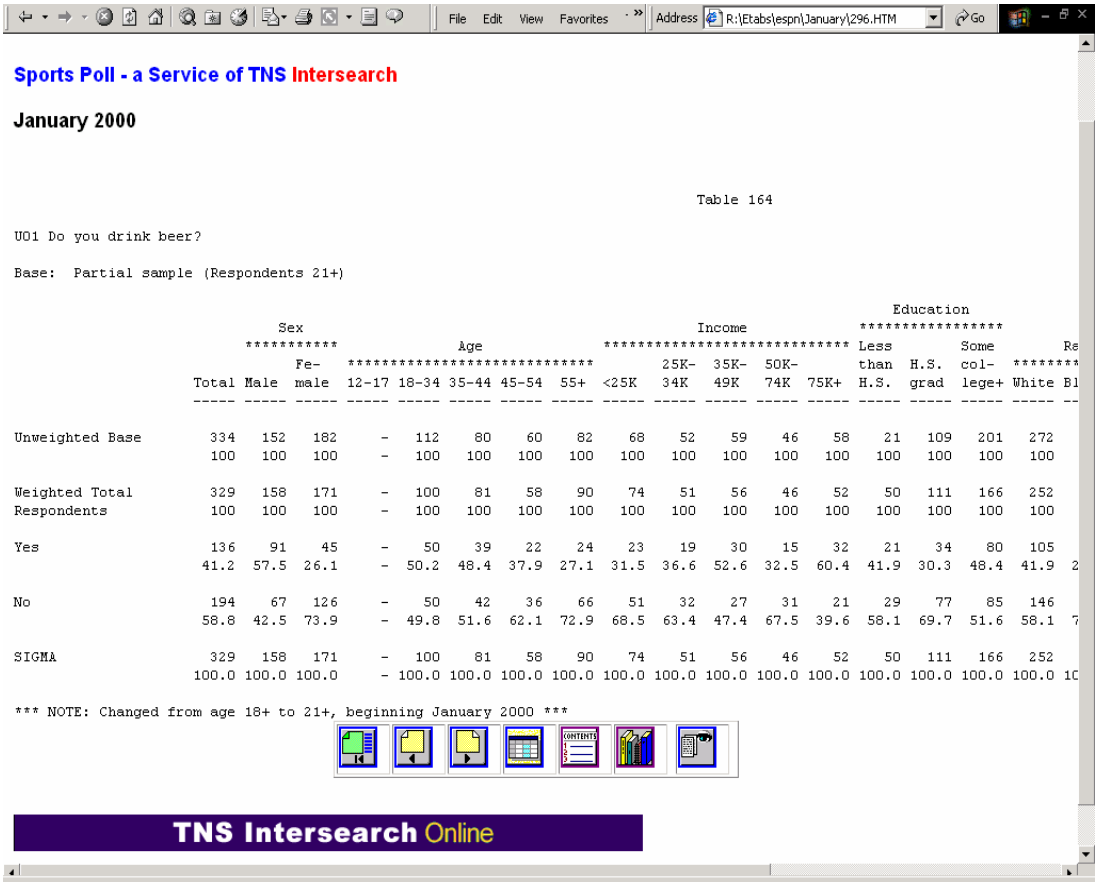
Online analytic applications allow real-time reporting and graphs of respondent data:

Analysis & graphics



Four-inch cross tab binders will become historical artifacts with e-tabs:

Online tabulations



Sports Poll - a Service of TNS Intersearch

January 2000

Table 164

UO1 Do you drink beer?

Base: Partial sample (Respondents 21+)

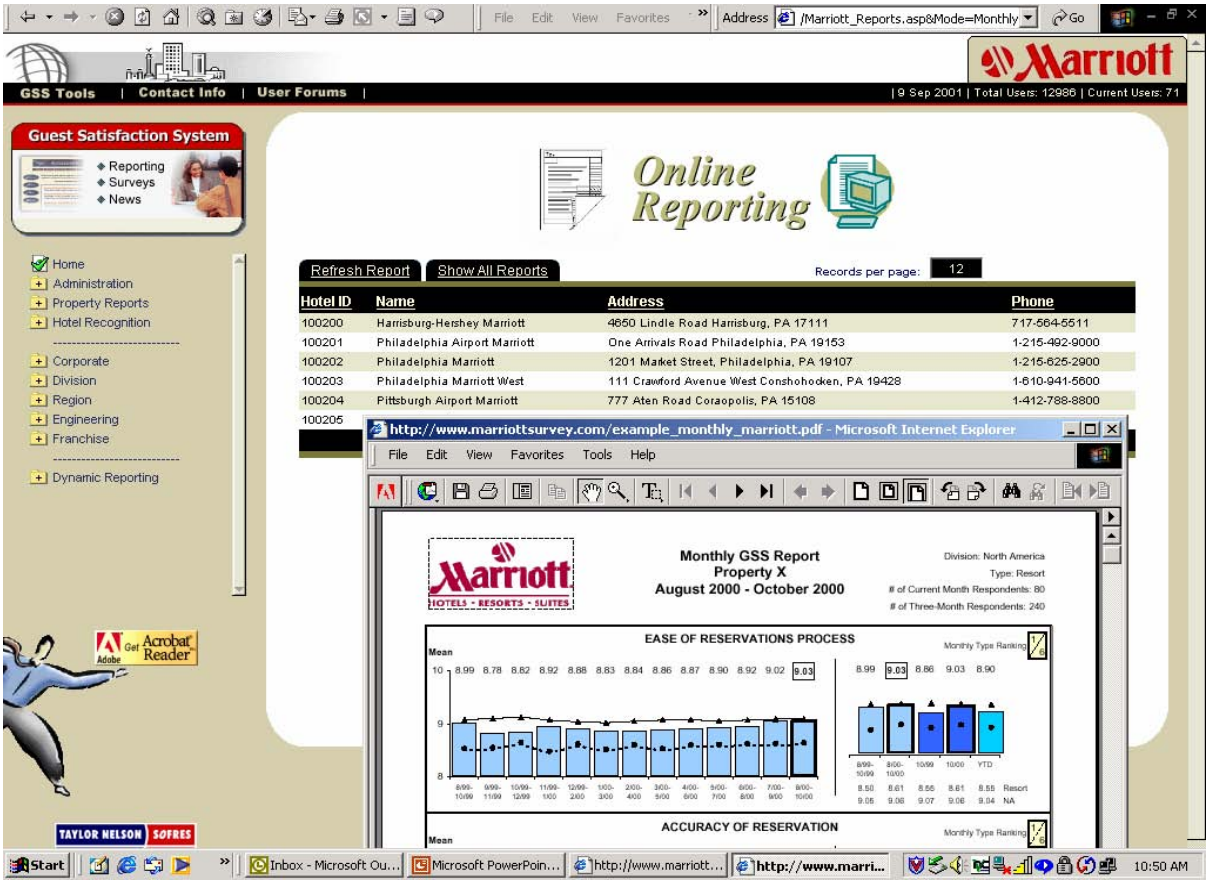
		Sex		Age					Income						Education			
	Total	Male	Female	12-17	18-34	35-44	45-54	55+	<25K	25K-34K	35K-49K	50K-74K	75K+	Less than H.S.	H.S. grad	Some college+	White	Bl
Unweighted Base	334	152	182	-	112	80	60	82	68	52	59	46	58	21	109	201	272	
	100	100	100	-	100	100	100	100	100	100	100	100	100	100	100	100	100	
Weighted Total Respondents	329	158	171	-	100	81	58	90	74	51	56	46	52	50	111	166	252	
	100	100	100	-	100	100	100	100	100	100	100	100	100	100	100	100	100	
Yes	136	91	45	-	50	39	22	24	23	19	30	15	32	21	34	80	105	
	41.2	57.5	26.1	-	50.2	48.4	37.9	27.1	31.5	36.6	52.6	32.5	60.4	41.9	30.3	48.4	41.9	2
No	194	67	126	-	50	42	36	66	51	32	27	31	21	29	77	85	146	
	58.8	42.5	73.9	-	49.8	51.6	62.1	72.9	68.5	63.4	47.4	67.5	39.6	58.1	69.7	51.6	58.1	7
SIGMA	329	158	171	-	100	81	58	90	74	51	56	46	52	50	111	166	252	
	100.0	100.0	100.0	-	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	10

*** NOTE: Changed from age 18+ to 21+, beginning January 2000 ***

TNS Intersearch Online

Customized client portals allow graphical results to be cut-and-pasted into reporting applications:

Information portal



#10: International one step at a time

- Wide variance remains across countries
  - penetration
  - usage habits
  - demographics
  - customs
- Cannot paint global online research with broad brush
- Internet growth has slowed
  - expansion worldwide may not be quick

While international internet penetration has grown faster than the ability to collect data from traditional in-person or telephone methods, the risks associated with violating geographic laws, local customs, and multi-language translation problems have also multiplied.

WHERE ARE WE GOING

Extended analytic applications:

- Multi-phase research designs
    - same or different respondents
    - quantitative or qualitative
    - compress timeline by combining phases
    - larger sample allows for in-depth analysis

Online qualitative research has offered new advantages for gaining rich open-ended responses. Virtual focus groups remove geographical boundaries, anonymity affords discussion of more sensitive topics, and a greater number of clients can simultaneously attend groups. Further, password-protected discussion board groups can be moderated via topical questions and threaded responses posted over a period of time for very low overhead costs.

Mixed-mode methodologies:

- Fit methodology to research need
    - offline recruit to online
    - give respondents a choice
    - vary by study or within study
- leverage benefits of environment in new ways
    - Affinova example

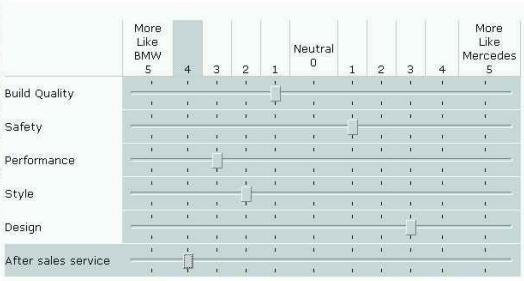
Extended graphics capability, today…

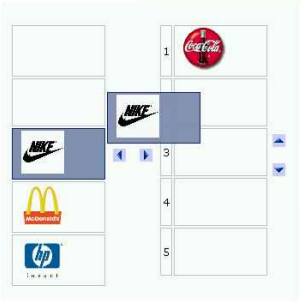
- beyond flat concepts
    - flip, rotate images
    - 360-degree views
    - thumbnails
    - package graphics & functionality

Three dimensional rendering of products or packaging is becoming more commonplace for online research. Technologies such as QuickTime VR allow zooming and an interactive interface for respondents to "experience" a prototype.

constant sum checking

interactive slider grids





graphical check boxes

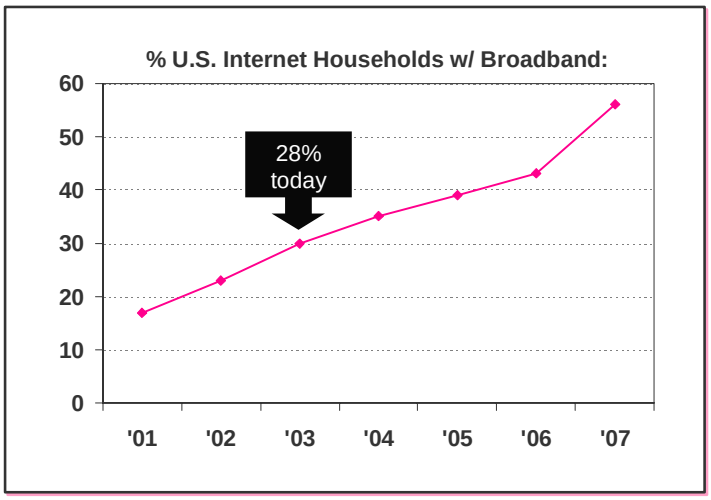
drag-and-drop ranking

Graphics-rich surveys, tomorrow…

- broadband opens new doors

Broadband taking off

- as penetration grows
  - sample bias shrinks
  - stimulus options multiply



Source: Jupiter Research; Gartner Dataquest Survey

Growing steadily at a rate of about 10% per year, U.S. broadband penetration is projected to significantly increase with a 30% yearly in 2007.

Broadband will hasten new techniques:

- online ad testing
    - single ads being tested today
    - broadband will increase viability of clutter reel

With current state-of-the-art video compression protocols, only single commercial administrations are rolling out for online advertising testing. When broadband reaches sufficient penetration, long-format programming quality will reach satisfactory levels for internet clutter reel testing.

- virtual shopping
    - view entire stores or shelf sets
    - navigate the aisles
    - examine packages / products
    - make product choices

Since high-bandwidth is necessary to render stores or shelf sets, only in-person administration for virtual shopping testing is currently available. Online virtual shopping tests, however, are currently in development.

- Virtual dressing room
    - retailers using this today
    - even faster with broadband

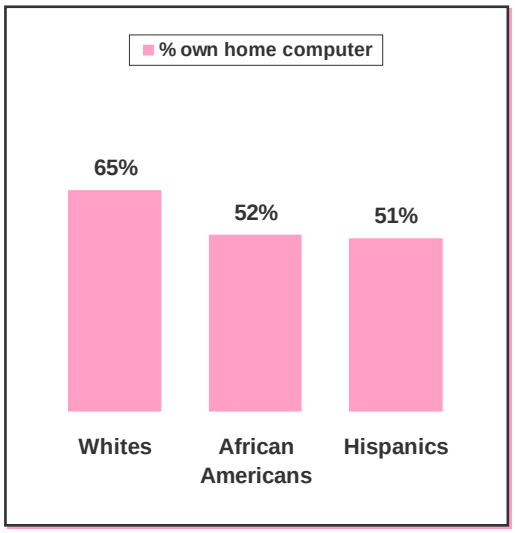
Promise in mobile?

- "A cell phone in every hand" soon?
    - two-thirds of American households have cell phone and growing
- Research applications?
    - high penetration
    - low functionality
    - per minute fees
    - low cooperation

Push-technologies and text / multimedia messaging on cell-phone or RIM / Blackberry platforms hold promise for the future of cell-enabled wireless research.

Reaching ethnic groups

- Becoming a more viable option
  - online usage growing
  - remains a challenge
    - penetration still lags
    - non-acculturated Hispanics
    - lower income segments
- Urban
  - online may not be your best option
  - consider design carefully



Some low-penetration targets are underrepresented online. Internet research cannot immediately replace all conventional methods, but is improving rapidly.

KEY TAKE AWAYS

Researchers

- imagined the possibilities ten years ago
- tested the possibilities six years ago
- realized the possibilities four years ago
- are maximizing the possibilities today
- will continue to evolve the possibilities into tomorrow

PANEL DISCUSSION: SAMPLING AND THE INTERNET

SUMMARIZED BY KARLAN WITT
CUSTOMER METRICS GROUP

For the conference this year, we invited representatives from three leading firms who offer very different approaches to online sampling to participate in a panel discussion. They included:

- Andrea Durning, SPSS MR Online
- Susan Hart, Synovate (formerly Market Facts)
- Michael Dennis, Knowledge Networks

This paper summarizes the information about each approach as provided by the presenters, as well as the moderator's remarks made during the conference.

The three approaches presented can be thought of very simplistically as:

- "River" sampling
- "Pool" sampling
- "Representative" online sampling (via a "pool" approach)

Presenters provided a brief background defining their approach, the resulting representativeness, and recommendations on uses for their panels. Below is a brief summary of each of the three approaches.

"RIVER" SAMPLING

SPSS MR Online/Opinion Place Overview

Opinion Place is one of the oldest and most heavily trafficked online research areas. Each week tens of thousands of respondents are drawn to Opinion Place via tens of millions of promotion impressions across AOL Time Warner's Internet properties. Respondents are screened in real-time and randomly assigned to surveys based on their responses to the screening questions. Over the past seven years, more than 10 million interviews have been conducted in Opinion Place.

The Opinion Place method has been compared to a river, where the constant flow of impressions yields a consistently fresh supply of respondents available for research. Clients find this method particularly compelling for sophisticated research (and especially for tracking studies), because AOL's perpetual promotion plan is always tapping into new, previously unengaged respondents. The promotion plan supporting Opinion Place is much larger and wider in scope than that of any other online research method.

Opinion Place Representativeness

Drawing from the numerous and diverse Internet properties available to Opinion Place, researchers are able to find sufficient sample sizes among the various demographic groups to include in many sample designs. Below are tables showing the ability to attract respondents across the most common demographic variables requested.

Internet & Opinion Place Demographics

	Internet %	AOL %	Opinion Place %	U.S. %
Gender				
♦ Male	47	43	34	48
♦ Female	53	57	66	52
Age				
♦ 18-24	12	14	10	13
♦ 25-34	20	18	23	19
♦ 35-44	25	23	26	22
♦ 45-54	24	25	21	18
♦ 55+	19	20	18	28
Married				
♦ Yes	66	62	56	53

Source: Internet & AOL Data from @plan Winter 2003 which represents 7/02-9/02; Opinion Place Data from Opinion Place Screener Oct-Dec 2002; US Data from US Census Bureau Statistics

Internet & Opinion Place Demographics

	Internet %	AOL %	Opinion Place %	U.S. %
Income				
♦ $25K or Less	9	10	19	29
♦ $25K - $49K	28	29	37	28
♦ $50K - $99K	41	40	34	29
♦ $100K +	22	22	10	13
Children In HH				
♦ Yes	43	44	46	33
Region				
♦ Northeast	20	23	23	19
♦ Midwest	23	20	20	23
♦ South	32	34	37	36
♦ West	25	23	19	22

Source: Internet & AOL Data from @plan Winter 2003 which represents 7/02-9/02; Opinion Place Data from Opinion Place Screener Oct-Dec 2002; US Data from US Census Bureau Statistics

Opinion Place Recommended Uses

SPSS MR Online recommends using Opinion Place for a broad range of research applications in both the B2B and B2C spaces. Among those discussed at the conference were:

- Concept / brand testing
- Conjoint studies
- Profiling & segmentation
- Website evaluations
- Customer satisfaction
- Brand image, awareness & usage
- Commercial / TV program testing
- Print ad testing
- Multimedia evaluations
- Tracking studies

"POOL" SAMPLING

Synovate Overview

Synovate (formerly Market Facts) presented their panel methodology for their ePanel. The ePanel is recruited in part from their larger Consumer Opinion Panel (COP), and hosts approximately 375,000 U.S. households.

ePanel members are recruited in two ways: by screening the COP households for e-mail addresses and willingness to participate in online surveys, and by an affiliated marketing program. Importantly, there is only a 20% overlap in COP and ePanel membership, which provides a large base of exclusive respondents in both panels. The affiliated marketing program recruits partner web sites and portals that promote their offer to join the ePanel to give opinions. The partner sites promote the offer by use of banner ads and by sending their subscribers e-mails with the explanation of the offer.

With a current base of approximately 375,000 respondents, the ePanel projects adding 20,000 new ePanel households each month in 2003.

Synovate ePanel Representativeness

Synovate has the longest history of the three companies in operating "pool" type panels for market research. They have extensive experience in recruiting, maintaining, and balancing panel samples. Although in their panels they may over-recruit target respondents in high demand to ensure that respondents are not over-used, their panel can be balanced to national representation. Below is a chart showing that balancing on key demographic variables.

ePanel Demo Household vs US Census				
EPANEL		**Census**		**DIFFER**
HH Region	**%**	**HH Region**	**%**	**%**
New England	4.72%	New England	5.30%	-0.58%
Middle Atlantic	13.45%	Middle Atlantic	14.10%	-0.65%
East North Central	17.85%	East North Central	16.70%	1.15%
West North Central	8.14%	West North Central	7.20%	0.94%
South Atlantic	19.43%	South Atlantic	19.10%	0.33%
East South Central	5.95%	East South Central	6.20%	-0.25%
West South Central	11.23%	West South Central	10.60%	0.63%
Mountain	6.86%	Mountain	6.40%	0.46%
Pacific	12.38%	Pacific	14.40%	-2.02%
HH Income		**HH Income**		
Under $17,500	11.77%	Under $17,500	17.90%	-6.13%
$17,500 - $32,499	22.96%	$17,500 - $32,499	19.90%	3.06%
$32,500 - $49,999	24.75%	$32,500 - $49,999	18.20%	6.55%
$50,000 - $74,999	21.79%	$50,000 - $74,999	18.50%	3.29%
$75,000 +	18.72%	$75,000 +	25.50%	-6.78%
HH Age Type		**HH Age Type**		
Family		Family		
Under 30 Years	11.87%	Under 30 Years	9.30%	2.57%
30-39 Years	21.23%	30-39 Years	16.30%	4.93%
40-49 Years	22.95%	40-49 Years	17.50%	5.45%
50-59 Years	14.57%	50-59 Years	12.20%	2.37%
60 Years +	6.28%	60 Years +	13.10%	-6.82%
Non Family		Non Family		
Male Under 40	3.50%	Male Under 40	5.90%	-2.40%
Male 40 Years & Older	4.23%	Male 40 Years & Older	8.20%	-3.97%
Female Under 40	7.18%	Female Under 40	4.00%	3.18%
Female 40 Years & Older	8.19%	Female 40 Years & Older	13.50%	-5.31%

Synovate ePanel Recommended Uses

With their long history in the off-line world, Synovate brings a deep understanding of appropriate panel uses. Their recommendations included:

- Attitudes & Usage
- Brand Equity
- Brand Tracking
- Concept Screening and Testing
- Copy / Logo Testing
- Market Segmentation
- Message Testing

- Pricing Research
- Product Placement
- Purchase Process
- Technology Adoption
- Low incidence screening for recontact studies

"REPRESENTATIVE" ONLINE SAMPLING

Knowledge Networks Overview

Knowledge Networks combines the best of traditional methodology — probability sampling — with the power, reach and multimedia capabilities of the Internet, so that you can ask the "right people" the "right questions" using the speed of Web interviewing. Unlike most of today's Web surveys, which are limited to users who already have Internet access, Knowledge Networks offers full population coverage by providing free hardware and Internet access to all selected households.

They begin with Random Digit Dialing (RDD) selection of households. This provides a scientifically valid list of potential participants. It is important to note that they're not just selecting computer users, or people who are already online. Knowledge Networks uses a random sample of all U.S. households with telephones, regardless of whether they have Web access, or even a computer.

This allows customers to choose from large, nationally representative random samples, or to select highly targeted customer groups that are impossible to reach using traditional survey methods.

Knowledge Networks Representativeness

With their representative recruiting approach, Knowledge Networks is able to build a panel that closely reflects the demographic profile of the U.S. population:

	Knowledge Networks Panel*	**US Adult Population****
Women	51%	51%
Age 18-34	34%	33%
Married	63%	61%
Working	74%	67%
Homeowner	74%	71%
Less than college	62%	69%
African-American	11%	12%
Hispanic	11%	11%
Over $75,000	25%	22%

* Weighted ** Source: U.S. Census

Knowledge Networks Recommended Uses

As the only approach that is described as truly representative, not only of the online user universe but also of the overall US population, Knowledge Networks recommends a broader set of uses for their panel. Rather than list specific types of recommended applications, they recommend using this solution whenever you would normally use an RDD sample, especially if there is a need to show the respondent the stimulus (as in a conjoint study, ad test, etc.).

COMPARISON OF APPROACHES

Each of the firms has a different approach to attracting and retaining relevant potential respondents. Below is a table that briefly summarizes the various approaches:

Design Element	Opinion Place	Synovate ePanel	Knowledge Networks
Population Base	Access to over 80% of the online population — 97.9 million unique monthly visitors	375,000 and growing by approx. 20,000 per month in 2003	40,000
Respondents	River	Pool	Pool
Recruitment Process	Ongoing promotions allow participation when convenient for respondent.	Multiple methods to leverage existing resources and partner affiliate marketing programs	RDD telephone listings sent pre-notification letter via special delivery mail
Frequency of Participation	Not more than 1 survey every 14 days; 1 per quarter in a given category	Not more than 2 to 3 per month per household; 1 per 6 weeks in a given category	Not more than once per week per household
Incentive	Multiple programs provide universally relevant incentives to each respondent	Multiple programs provide universally relevant incentives to each respondent	Multiple programs provide universally relevant incentives to each respondent
Typical Response Rates	85% +	35%	70%

As with any panel, there are certain sub-groups that tend to be under-represented, as shown in the tables above. From a sampling perspective, you can still request that a certain number of these be included for your study, but the concerns regarding representativeness increase as participation within a sub-group decreases.

SUMMARY

If you think about objectives, each of these firms is offering options to help you balance several aspects to meet your objectives. In any study, to some extent, you have to balance the

quality of the research, the speed with which results are delivered, and the cost. Historically, we believed (and told clients) you had to pick two of these three. With online research, the hope has been to achieve all three. The three online methods presented here all offer options that yield similar turnaround times. In the interest of time, we'll focus on issues related to quality and cost choices.

Quality

While as researchers we think about quantifying "error" in a statistical sense, error represents a quality failure that can impact the conclusions from the data. But maximal quality is not the primary objective of every study. In rapidly evolving markets, a CEO may simply want to confirm that the market is headed "this" way instead of "that" way and do so quickly while conserving precious capital. So let's focus on what impacts quality so that we can make conscious choices about what types of compromises may be acceptable given the study objectives.

We'll categorize the types of error into four areas:

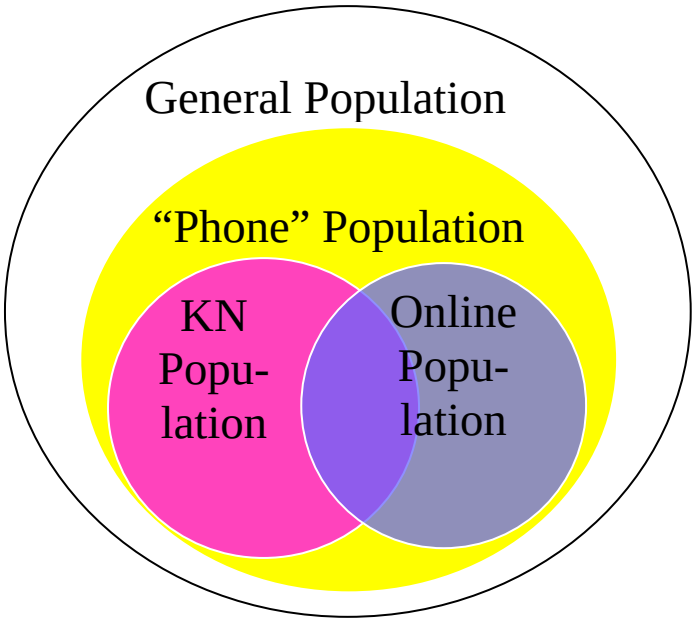
Coverage error is an interesting one for internet panels. It refers to issues arising when everyone you define in your target population is not represented in your sample frame. In this case, it may be trying to conduct a general population survey online, when there are still portions of the population who are not online.

Sampling error simply refers to issues arising from drawing a sample of the population rather than a census. Depending on how it is drawn, it may not be representative of the overall population. While sampling error can occur in any type of sample, it can be greatly pronounced in online samples.

Non-response error is an issue with every modality. However, with telephone and other types of studies, survey research was an established and accepted practice before telemarketing really took hold. Unfortunately, with the web, if anything, folks are getting vastly more email solicitations than invitations to genuine studies, and response rates have declined precipitously for many types of online recruitment.

Finally, **measurement error** is something we look at in all types of research, and we've heard more about it at this conference. Are we measuring what we think we are measuring? This is primarily a design issue and the key issue to note when comparing the options shown here is that panel members (in a traditional "pool" approach) are known to exhibit some level of "panel conditioning" such that their answers are not necessarily the same as non-panel members. This is not an issue unique to internet panels, however, so we won't spend time on this today.

Let's look briefly at a coverage issue, which some believe to be the biggest issue with online research, and others dismiss entirely.

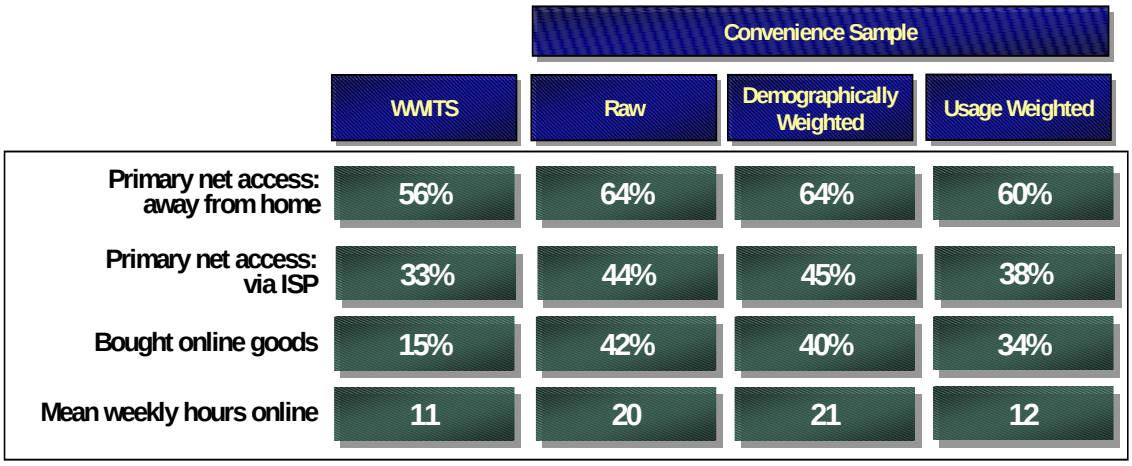


For some surveys, clients may decide that they only want the more affluent and technology savvy folks who are online, and like the other benefits of using web-based research enough to make that compromise. Depending on the target audience and study objectives, they may select an online pool approach such as the Synovate solution, or a river sampling approach such as Opinion Place.

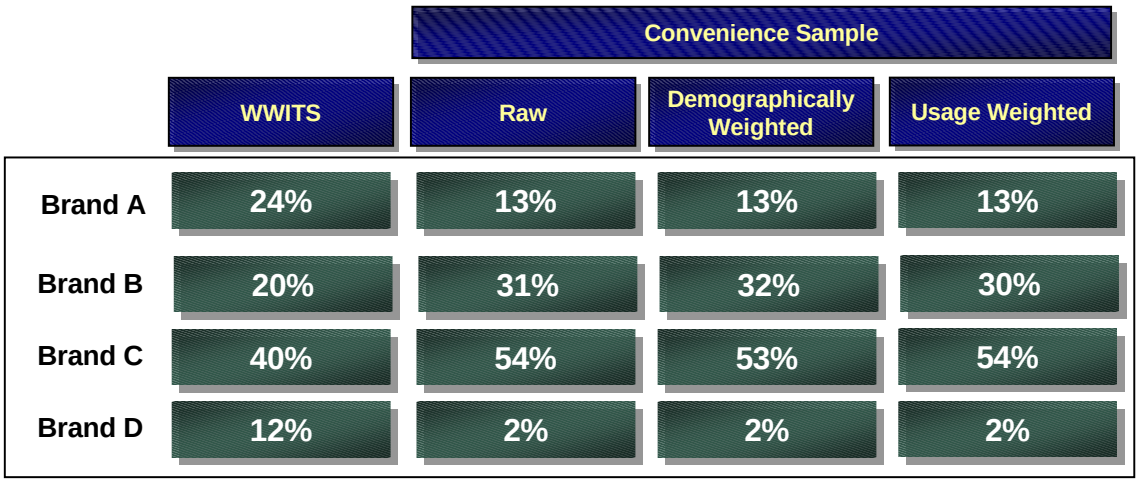
For others, that compromise would undermine the objectives, but they may require the visual display capabilities of an online study, or the speed with which it can be completed, so they opt for a Knowledge Networks (KN) solution.

Moving on to Sampling Error, I'd like to revisit data from a paper shown at this conference a few years ago, so that the details are available to everyone[1]. Briefly, in this study, a representative sample was compared to one recruited from invitations on the top 6 portal sites. The raw results had some imbalances. Keith Chrzan, at IntelliQuest at the time, was recruited to attempt to weight the convenience sample data to reflect the data captured from the representative study. The results of this were an entire paper unto themselves, but basically, numerous weighting schemes were employed with mixed results. While some variables were brought into line, others were not.

[1] Source: IntelliQuest World Wide Internet Tracking Study and Online Snapshot Study Q1 1998.



The criteria we often hear from clients is "would I make a different business decision using this data?" So, the next variable examined was brand share data. In both data sets, we see that Brand C is a strong leader in the market. However, after that, the rank orders vary, suggesting that when the composition of our sample shifted, so did the resulting brand shares which were obtained.



When I hear vendors (especially online sample vendors) promote how large their samples are, or the high number of completes you can get for your budget dollar, I am reminded of the 1936 presidential election, in which the *Literary Digest* conducted a poll and achieved a sample of over 2 million to predict of Alf Landon winning the presidential election. Roosevelt won in all but two states. In this case, quantity definitely did not make up for quality. Our challenge in this electronic age is to find useful applications for online surveys and the various sampling options.

Cost

Each of the three vendors offers methodologies that for certain situations offer a compelling way to meet your objectives. To complete the ROI calculation, let's consider the cost side of the equation.

Each of the vendors was asked to bid on two studies. The specifications for those bids are included in Appendix A.

Here are the costs each provided to conduct those studies. These costs include questionnaire programming, sample costs, incentives, web hosting, and cross tabs.

Vendor	**Investor Study**	**Status of Women N=500**	**in America Study N=1000**
Opinion Place	$14,800	$8,900	$12,250
Synovate	$30,400	$20,300	$26,800
Knowledge Networks	$53,940	$36,300	$57,800

In each case, Opinion Place was the least expensive option, Synovate the next, and Knowledge Networks the most expensive with prices about on par with traditional telephone research.

In cases where a representative audience of the type that an RDD telephone survey would produce is a must, such as market sizing and forecasting, but with stimuli that is best served through a web interface, Knowledge Networks offers a stronger solution than other on-line or off-line agencies. However, in cases where the objectives allow more flexibility and budgets are constrained, Opinion Place and Synovate each offer unique benefits at differing costs to help researchers meet their goals.

There are a variety of other firms that have related approaches for sampling. We've reviewed three main approaches here. We encourage you to carefully research the variety of quality options out there. We hope this review gives you some points to think about and some background for making a decision.

APPENDIX A

REQUEST FOR PROPOSAL FOR PUBLIC CONFIDENCE SURVEY

Sample Source

You will provide a random sample list for use as the research sample. Below are criteria for the sample source:

- Qualified respondents will all be 18 years or older.
- Respondents should represent requisite demographic groups so that they can be weighted back to the 2000 U.S. Census.
- Using Prizm clustering or other similar methodology or panel data, we need to oversample "active investors" who currently have a minimum of $100,000 invested in the securities market. We need at least 250 completed interviews within this target audience.

Sample Size

The total sample size is 1,250, including the 250 "active investors" described above.

Questionnaire

The questionnaire will be provided to you and should be ready for programming. The survey will be approximately 15 minutes in length.

The surveys should all be conducted via the internet.

Due to the nature of this project, we would like the option of surveys being completed in Spanish as well as English to minimize terminates due to language barriers. Please provide pricing for English-only, as well as English and Spanish.

It is assumed that there will be two open-ended questions that will need to be coded and tabulated. You should propose a coding scheme which we approve before coding is done. The coding scheme may be done off of partial data to facilitate timing.

Data Analysis

For the data analysis, we believe we are looking at the following activities:

- Clean the data to check that all skip patterns are followed, numeric questions are within range, etc.
- Assume two sets of banner tables.
- By way of advanced analyses, we anticipate one regression analysis to identify predictors of public confidence.
- Data should be weighted back to reflect the 2000 U.S. Census.

REQUEST FOR PROPOSAL FOR STATUS OF WOMEN IN AMERICA STUDY

Sample Source

You will provide a random sample list for use as the research sample. Below are criteria for the sample source:

- Qualified respondents will all be 18 years or older and must be women.
- Half of the women should be <43 years old and half 43 or above with the overall sample having a median age of 43. Within the above and below 43 groups, ages should be spread so that all age groups can be represented.

Sample Size

Final sample size will depend on cost. Please provide cost estimates for 500 and 1,000 completed interviews.

Questionnaire

The questionnaire will be provided to you and should be ready for programming. The questionnaire will include topics covering the status of women on topics such as:

Health	Career	Education	Safety
Finance	Family	Travel	Media
Technology	Housing	Sports	

Due to the broad nature of the topics to be covered, the survey will be approximately 25 minutes in length. To increase response rate, we will disclose the name of the non-profit sponsor of the research to respondents.

It is assumed that there will be two open-ended questions that will need to be coded and tabulated. You should propose a coding scheme which we approve before coding is done. The coding scheme may be done off of partial data to facilitate timing.

Data Analysis

For the data analysis, we believe we are looking at the following activities:

- Clean the data to check that all skip patterns are followed, numeric questions are within range, etc.
- Assume two sets of banner tables
- Since we are potentially using quotas, we will need to weight the data back to reflect the 2000 U.S. Census.

RECOMMENDED READING

Batagelj, Zenel, Bojan Korenini, and Vasja Vehovar. 2002. "The integration power of the intelligent banner advertising network and survey research." Paper presented at the Net Effects 5 Conference, Berlin, Germany, February.

Comley, Peter and Ray Poynter. 2003. "Beyond Online Panels." Paper presented at Technovate: CRM, Internet Research, and New Media, Cannes, France, January.

Couper, M. P., M. Traugott, and M. Lamias. 1999. "Effective Survey Administration on the Web." Paper presented at the Midwest Association for Public Opinion Research, November.

Dommeyer, C.J., and E. Moriarty. 2000. "Increasing the Response Rate to E-Mail Surveys." Paper presented at the annual meeting of the American Association for Public Opinion Research, Portland, OR, May.

Groves, R. M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Krotki, K. P. 2000 "Sampling and Weighting for Web Surveys." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

Rivers, D. 2000. "Probability-Based Web Surveying: An Overview." Paper presented at the annual conference of the American Association for Public Opinion Research, Portland, OR, May.

ONLINE QUALITATIVE RESEARCH FROM THE PARTICIPANTS' VIEWPOINT

THEO DOWNES-LE GUIN
DOXUS LLC

Despite an increasing number of misgivings about abuse of focus groups for different research objectives, all evidence points to the fact that the traditional group is alive and well. As Kevin Roberts of Saatchi & Saatchi put it so eloquently at the 2002 ESOMAR conference in Barcelona,

> "We all know that focus groups are a miserable way to understand anything. And we all know why. And yet we can't seem to leave them behind. Why not? For a deeply emotional reason. For a brief moment we can touch consumers. Flawed, skewed and inaccurate the results may be, but for that moment real people talking feels true."

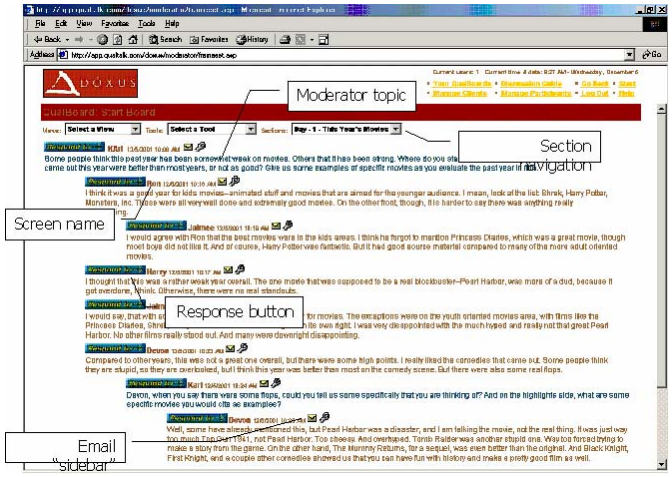
The dominant research videoconference vendor estimates that a total of 224,000 US focus groups took place in 2001, generating an estimated expenditure of more than US$1 billion, up from 213,000 in 2000 and nearly double the norm for most of the 1990s (source: FocusVision Worldwide).[1] Worldwide, the traditional focus group has experienced a similar rate of growth. Even as the bloom is off the methodological rose and corporate travel budgets are under scrutiny, focus groups continue to consolidate their standing as a dominant research modality.

But what happened to the online focus group? Adoption estimates for online qualitative methods are harder to come by, but by most anecdotal accounts, online focus groups have not seen the "hockey stick" adoption curve that many vendors hoped for a few years ago. Many researchers have found the process and results limiting or disappointing. Concerns with online focus groups range from user-friendliness to security to inadequate cost savings or show rates, but the complaint we hear most frequently is about the quality of the interaction. With many topics and populations, online focus groups create an environment that encourages a casual tone (inspired by chat room norms) at the expense of more informed discourse.

Like so many ideas that have arisen from the internet, however, online focus groups have nevertheless borne interesting fruit. The threaded discussion (or bulletin-board focus group) seems to be filling the gaps that online focus groups have left.[2] The basic technology and process is much like the online focus group: participants are recruited by email (if a source is available) or telephone or traditional means; they visit a secure website and contribute ideas to a moderated session (see below for an example of the user interface).

[1] Estimates comprise qualitative research conducted by individual researchers and smaller companies as well as multinational vendors and CASRO member organizations.

[2] QualTalk, one of several vendors for the web technology for threaded discussion, has seen demand for bulletin boards quadruple from 2001-2002.



Unlike the online focus group, however, the threaded discussion takes place over multiple sessions and days (typically three sessions a day for three or four days). Participants log on at their convenience throughout a one- or multi-day period and respond to questions. Each participant can see the answers that others have given and is encouraged to interact with other participants, not just with the moderator. The ideal result is a rich, developed dialogue about the given topic.

Threaded discussions overcome the shortcomings of the synchronous online groups in a number of ways:

- Threaded discussions offer the opportunity for more, and more leisurely, participant/moderator interaction because the content is not being posted in time-constrained environment. In addition, posts in threaded discussions are typically longer and more articulate.
- Just as participants have more time to ponder their input and the moderator's questions, clients have more time to scrutinize what they are "hearing" and relay ideas for new questions — even new stimulus — to the moderator. Especially when evaluating new product ideas or marketing deliverables, the format offers a lot of flexibility in altering the path of inquiry or showing iterated stimuli that are driven by previous responses.
- Technical and keyboard skills of the participant are less likely to affect participation rates. Because the ability to get online and type fast no longer privileges those with fast connections and faster fingers, the threaded discussion allows all participants to contribute at an equal level.

Needless to say, the method is not appropriate for every research need any more than focus groups are universally appropriate (though they are often treated as such). Over multiple projects we have found that threads are most successful from participants', researchers' and clients' perspectives when focused on:

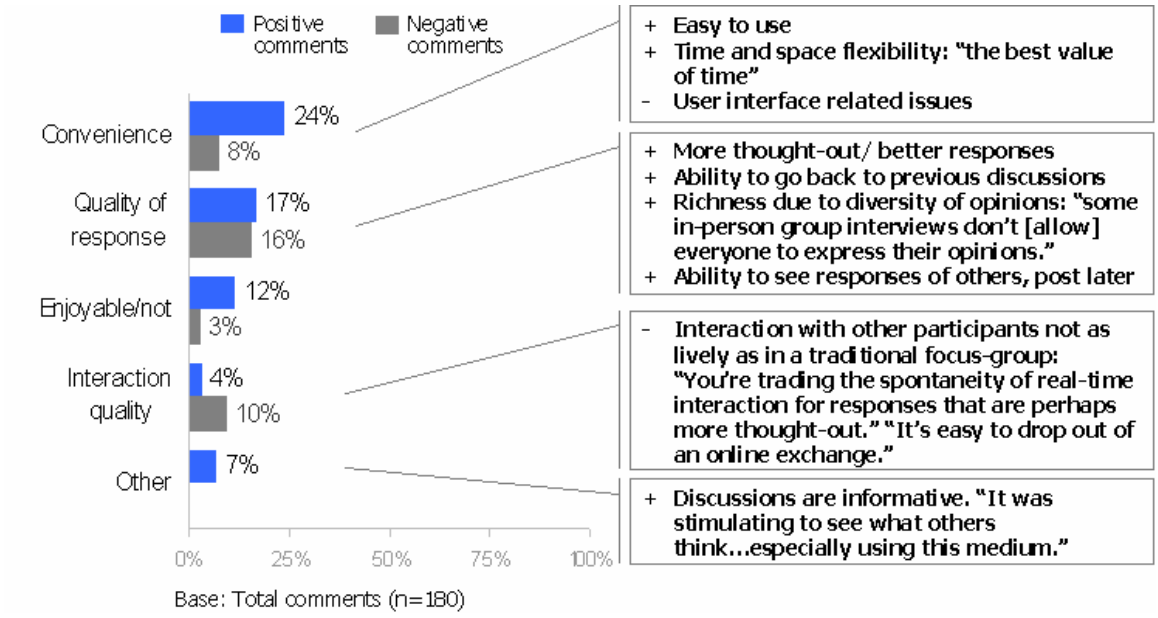
- High involvement topics, especially (but not only) those related to technology. Low involvement topics will not sustain participants' interest over a two-or-more day period, resulting in a steady drop-off in participation.

- Geographically-dispersed populations. Because one is not constrained to a few cities and because of the asynchronous nature of the discussion, discussions can be comprised of national or international audiences without regard to time zone while still fostering a true discussion
- Populations in which cooperation rates are declining. The novelty and, as we discuss below, convenience of the method alone makes it quite appealing to highly-researched populations like IT managers.
- Stimuli-intensive research needs (e.g., multiple concepts or websites) that take time to review and absorb.

Concomitantly, technology-averse or low web penetration populations are obviously a poor target for any web-based mode. In addition, interpretation of many qualitative research processes rest heavily on the researchers' ability to read beyond words and tone to body language, facial expression and physical interaction between participants — which is simply not an option in threaded discussions.

To date, most literature on threaded discussions has described the method from the researcher's perspective without reference to the respondent experience. To address this concern, Doxus has inserted questions at the end of several bulletin board discussions, representing 110 participants, to learn about their experience of participating in this format as compared to other forms of research in which they may have taken part, including surveys and focus groups.[3]

Our questions about the bulletin board experience were open-ended, but generally, we talked to participants about their experience in terms of convenience and interaction quality. The table below shows the results of a reduction and coding of verbatim responses into several macro categories.



[3] The research populations and topics for these projects vary widely, but for the most part we talked to professionals and a few consumers regarding technology products and services. We chose the bulletin board method for these projects because of its appropriateness to the population and topic, so our findings are in no way a proxy for a robust, experiment-based investigation of participants' methods preferences.

Many of the participants we interviewed had participated in traditional, in-person focus groups or standardized surveys in the past, with a smaller subset having participated in other on-line qualitative methods such as online focus groups. On the whole, participants express a preference for threaded discussions in terms of convenience. Many comments regarding convenience underscore a general appeal to the methodology:

- "I liked it quite a bit actually. Flexible schedule, lots of opinions, and good questions. On line provides [the] best value of time"
- "I liked this format. I have done online focus groups and I have done in person research groups. I hope more research groups will take this approach in the future."
- "I really like the idea of being able to do this without leaving the office and to be able to log in whenever time permits."
- "It is a luxury to be able to access the discussion at my convenience instead of rearranging my schedule to participate."
- "[This method] lets me find time in much smaller chunks as opposed to taking a full afternoon off somewhere."

In sum, participants appear to recognize and appreciate that threaded discussions are very different from the traditional approach to research, which involves a contained interaction that may or may not happen at the respondent's convenience.

Comments also suggest that threaded discussions provide a higher quality of discussion, at least by the criterion of participant thoughtfulness and articulation. Most participants we interviewed say that the quality of the individual responses is better than in other methods (with the most typical comparison being in-person focus groups) because the threaded discussion format allows time to give thought to the moderator questions and other participant postings before responding. As one participant put it, "It allows me to think a question through and answer thoroughly without being rushed or interrupted by other participants!"

Time for thought results in a relaxed, less competitive experience for the participant — and thus an ability to contribute more mature thought. Rather than "merely repeating some already articulated ideas" or posting comments not really applicable to the topic of discussion, several say that threaded discussions allow them to post more substantive comments. From the researcher perspective, however, this benefit may be fairly subjective; across many studies we have noticed that some participants take the time to consider and edit their responses, while others treat the method with the immediacy and spontaneity of a chat session.

Finally, a few of our research subjects mention the lack of dominance of strong personalities as a merit of threaded discussions compared to an in-person focus group. Because of dominant respondents, "some in-person group interviews don't give enough opportunity for everyone to express their opinions," while in a threaded discussion all participants have (theoretically) equal opportunity to post their comments without being "dominated by persons that love to hear themselves talk."

Despite these advantages, however, many participants recognize that the threaded discussion is not always able to achieve the interpersonal richness and immediate gratification (for networking and socialization purposes) of an in-person focus group. The concern is not so much

the absence of spontaneity of an online discussion as it is missing the liveliness of an in-person focus group. Some “missed the immediate interaction and the ability to read facial expressions of an in-person focus group” and the “opportunity for detailed interactions with other participants.” As one neatly summarized the trade-offs, “You’re trading off the spontaneity of real-time interaction for responses that are perhaps more thought-out.”

A related concern is the difficulty of maintaining participants’ attention over a longer period of time, and through multiple interactions rather than just one. While show-up rates for participants are a consistent source of anxiety for most qualitative researchers, once participants are in the room they are extremely unlikely to leave (though they may try to tune out of the discussion). Some online participants comment that the online experience simply didn’t grab and hold them as an in-person group would. In terms of sheer physical stimulation, most participants with a basis for comparison agree that in-person groups are optimal for keeping participants engaged: “In-person…you're in a controlled environment without the distractions of the office.” As with other modes, however, the researcher is ultimately responsible for maintaining participation rates by making sure that the appropriate respondents are recruited, the discussion is relevant and interesting, and any reasonable means are employed to keep participants engaged throughout the process.

Participants’ comments regarding quality of discussion and interaction highlight the fact that the method is not appropriate for every topic or every participant. But in some settings, threaded discussions may lead to greater cognitive involvement for some topics, while erasing barriers to equal participation such as typing skills or verbal articulateness.

ITEM SCALING AND MEASURING IMPORTANCES

SCALING MULTIPLE ITEMS: MONADIC RATINGS VS. PAIRED COMPARISONS

BRYAN ORME
SAWTOOTH SOFTWARE, INC.

INTRODUCTION

One of the most common tasks for market researchers is to measure people's preferences for things such as brands, flavors, or colors, or to measure the importance of product features. For these kinds of problems, the monadic rating is a very popular approach. For example:

> How important is each of the following in your decision to purchase an automobile? (Use a scale where a "10" means it is Extremely Important and a "1" means it is Not Important At All.)
>
> It is luxurious
>
> It is useful for variety of tasks
>
> It offers a smooth ride
>
> It offers a quiet ride
>
> I like the way I feel when I ride in it
>
> I like the way I look when I ride in it
>
> I get a good value for the money
>
> etc…

There are a variety of formats for the monadic rating. A "grid" style with rows of checkboxes for each item is quite common for both paper and Web-based surveys.

Usually, the respondent is asked to rate many items. Respondents are notorious for rating items very rapidly, using simplification heuristics. How can we blame them when we present so many items for their consideration? Despite our best efforts to offer many scale points, respondents tend to use only a subset of them, resulting in many ties. Moreover, there are many response styles. Some respondents use just the top few boxes of the scale, some respondents religiously refuse to register a top score for any item, while others conscientiously spread their ratings across the entire range.

Response style bias poses problems for statistical tests of significance and multivariate modeling. To deal with this, the data can be standardized such that the mean rating within each respondent is zero, and standard deviation is 1, but these transformations often make it difficult to present the data to managers. This, and the added work of standardizing data, leads to rare use of standardization in practice.

By using rating scales in a blunt way, respondents provide less statistical information by way of discrimination among the items than the researcher would desire. An extension of the monadic ratings that forces discrimination is to require that each rating only be used once. This is cognitively more difficult, and is a middling approach between the monadic rating and the also common ranking procedure.

With ranking questions, respondents order the items from best to worst (with no ties). To answer the question, respondents *must* discriminate among items. However, this procedure has two main drawbacks:

1. Respondents often find it difficult to rank items — particularly more than seven.
2. The resulting scores are on an ordinal (rather than metric) scale. We cannot know whether the difference (e.g. in preference, importance) between ranks 1 and 2 is the same as the difference between ranks 2 and 3. This limits the researcher to non-parametric procedures that are not always as statistically powerful or commonly used as the parametric tests and procedures.

Researchers have experimented with many techniques to achieve the benefits of metric scaling while also encouraging respondents to discriminate among the items. Another common approach is the "constant sum" (allocation, or chip allocation) scale. With constant sum scales, respondents allocate a certain number of points or chips across an array of items. Some researchers prefer 100 points, while others prefer prime numbers like 11. With prime numbers, it is impossible to tie all the items (unless the number of items equals the number of points to be allocated).

Some researchers prefer the constant sum scale, but it also has its limitations:

- Some respondents find it difficult to make the allocations sum to the required number (though the researcher can relax this requirement). This hurdle may get in the way of respondents accurately reflecting their preferences.
- As with rankings, it is difficult to do — especially with more than about seven items.
- Whereas the researcher desires that the points (chips) be allocated in an independent fashion, it is usually not the case.

There are many other techniques that researchers have developed over the years to deal with the problems we have been discussing. The Method of Paired Comparisons (MPC) is one of the oldest and is the subject of this paper.

THE METHOD OF PAIRED COMPARISONS

In its simplest form, respondents are shown just two items (objects) at a time and asked in each case to choose one of them. Consider the case of three items: a, b, and c. With just three items, there are three possible comparisons: a,b; a,c; and b,c (assuming a,b and b,a, etc. are equivalent.) Generally, with t items, the number of possible comparisons is $\frac{1}{2}\, t(t\text{-}1)$. For example, with 10 items, there are $\frac{1}{2}(10)(9) = 45$ possible comparisons.

The beauty of MPC is that by using a series of simple comparisons, we can place many items on a common interval scale. Interval scales are valuable in marketing research, as they let the researcher use a variety of parametric statistical techniques. Even a child unable to understand a rating scale could perform a series of paired comparison choices reliably, yielding a statistically informative assessment of all the items. Being able to translate simple comparisons to an interval scale also has valuable implications for cross-cultural research, where controlling for response style bias is desirable.

MPC questions indeed require more thought and time to complete than simple monadic ratings. But these more challenging questions are also less transparent to respondents. With monadic ratings, respondents more easily settle into a less motivated, patterned response strategy. MPC discourages inattention and seems to elicit more insightful responses.

With many items, the number of possible comparisons can become very large. However, it is not necessary to ask respondents to make all comparisons to obtain reasonably stable interval-scale estimates for all items. Incomplete (fractional) designs showing just a carefully chosen subset of the universe of comparisons are adequate in practice. This is particularly the case if Latent Class or hierarchical Bayes (HB) is used to estimate the parameters. Indeed, the availability of Latent Class and HB are breathing new life into this old method.

In my experience (and also based on some personal correspondence with researcher Keith Chrzan), it seems that useful individual-level results can be achieved by asking as few as 1.5*t* questions, where *t* is the number of items in the MPC experiment. If the researcher is willing to forego the requirement of individual-level estimation in favor of aggregation, each respondent may be asked only a few comparisons. With randomized designs, in the limit (given a very large sample size), just one MPC question could be asked of each respondent.

Fundamental to MPC is the law of *transitivity*. Transitivity holds that if a>b (where ">" means is preferred to) and b>c, then a>c. In practice, it has been demonstrated that errors in human judgment can lead to seemingly contradictory relationships among items (intransitivities, or *circular triads*), but this almost invariably is attributable to error rather than systematic violations of transitivity.

The Method of Paired Comparisons has a long and rich history, beginning with psychophysicists and mathematical psychologists and only eventually becoming used by economists and market researchers. The history extends at least to Fechner (1860). Fechner studied how well humans can judge the relative mass of a number of objects by making multiple comparisons: lifting one object with one hand and a second object with the other. Among others, Thurstone (1927) furthered the research, followed by Bradley and Terry (1952). History shows that Paired Comparisons is a much older technique than the related Conjoint Analysis (CA) though CA has achieved more widespread use, at least in marketing research. Later, I'll discuss the similarities and differences between these related techniques.

There are two recent events that increase the usefulness of Paired Comparisons. The first is the availability of computers and their ability to randomly assign each respondent different fractions of the full balanced design and randomize the presentation order of the tasks. When using techniques that pool (or "borrow") data across respondents, such designs usually have greater design efficiency than fixed designs. The second and probably more important event is the introduction of new techniques such as Latent Class and HB estimation. Latent Class and

HB have proven superior to previously used methods for conjoint analysis and discrete choice (CBC), achieving better predictions with fewer questions, and are equally likely to improve the value and accuracy of Paired Comparisons.

PROS AND CONS OF PAIRED COMPARISON

Pros of MPC:

- MPC is considered a good method when you want respondents to draw fine distinctions between many close items.
- MPC is a theoretically appealing approach. The results tend to have greater validity in predicting choice behavior than questioning techniques that don't force tradeoffs.
- Relative to monadic rating procedures, MPC leads to greater discrimination among items and improved ability to detect significant differences *between items* and differences *between respondents* on the items.
- The resulting scores in MPC contain more information than the often blunt monadic ratings data. This quality makes MPC scores more appropriate for subsequent use within a variety of other multivariate methods for metric data, including correlation, regression, cluster analysis, latent class, logit and discriminant analysis, to name a few.

These points are all empirically tested later.

Cons of MPC:

- MPC questions are not overly challenging, but the number of comparisons required is often demanding for respondents. Holding the number of items constant, an MPC approach can take about triple the time or more to complete over simple monadic ratings, depending on how many MPC questions are employed.
- MPC is analytically more demanding than simpler methods (monadic ratings, rankings, or constant sum). The design of the experiment is more complex, and estimation techniques such as hierarchical Bayes analysis can require an hour or more to solve for very large problems.
- The resulting scores (betas) from an MPC exercise are set on a relative interval scale. With monadic ratings, a "7" might be associated with "somewhat important." With MPC, the scores are based on relative comparisons and don't map to a scale with easy to understand anchor points. This makes it more challenging to present the results to less technical individuals.
- In MPC, every respondent gets the same average score. If a respondent truly either hates or loves all the items, the researcher cannot determine this from the tradeoffs. This is sometimes handled by including an item for which most every respondent should be indifferent.

COMPARING PAIRED COMPARISONS AND CONJOINT ANALYSIS (CA)

Because the audience likely to read this paper has extensive experience with conjoint analysis, it may be useful to draw some comparisons between these techniques.

How is MPC similar to CA?

1. Both MPC and CA force tradeoffs, which yield greater discrimination and should in most applications lead to better predictive validity.
2. Rather than ask respondents to directly score the items, with either method we observe actual choices and derive the preferences/importances.
3. For both MPC and CA, responses using either choices or a rating scale yield interval scaled betas (scores) for items (levels).
4. Scores can be estimated at either the group or the individual level for MPC and CA.
5. By specifying prohibitions in MPC, some of the items can be made mutually exclusive. But, there is no compelling reason to require items to be mutually exclusive as with CA.

How is MPC different from CA?

1. With MPC, concepts are described by a single item. In CA, product concepts are described by multiple (conjoined) items. Since MPC does not present conjoined items for consideration it is not a conjoint method.
2. Because each item is evaluated independently in MPC (not joined with other items), it is not possible to measure interaction effects as can be done with conjoint analysis.
3. MPC has the valuable quality that all objects are measured on a common scale. In CA, the need to dummy-code within attributes (to avoid linear dependency) results in an arbitrary origin for each attribute. With CA, one cannot directly compare one level of one attribute to another from another attribute (except in the case of binary attributes, all with a common "null" state).
4. MPC is suitable for a wider variety of circumstances than CA: preference measurement, importance measurement, taste tests, personnel ratings, sporting (Round Robin) tournaments, customer satisfaction, willingness to pay (forego), brand positioning (perceptual data), and psychographic profiling, to name a few. Almost any time respondents are asked to rank or rate multiple items, paired comparisons might be considered.

DESIGN OF EXPERIMENT

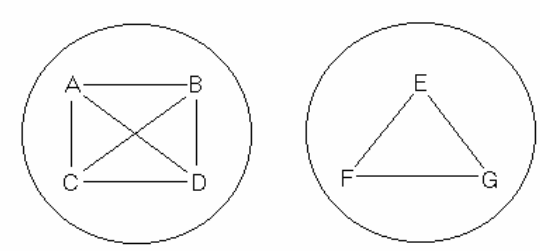
With t items, the number of possible paired comparisons is $\frac{1}{2}\, t(t\text{-}1)$. A design containing all possible combinations of the elements is a *complete* design. With many items, the number of possible comparisons in the complete design can become very large. However, it is not necessary for respondents to make all comparisons to obtain unbiased, stable estimates for all items. Fractional designs including just a carefully chosen subset of all the comparisons can be more than adequate in practice.

As recorded by David (1969), Kendall (1955) described two minimal requirements for efficient fractional designs in MPC:

1. *Balance*: every item should appear equally often.
2. *Connectivity*: the items cannot be divided into two sets wherein no comparison is made between any item in one set and another item from the other.

Imagine an MPC design with seven elements, labeled A through G. An example of a *deficient* design that *violates* these two principles is shown below:

Figure 1



In Figure 1, each line between two items represents a paired comparison. Note that items A through D appear each in three comparisons. Items E through G appear each in only two comparisons. In this case, we would expect greater precision of estimates for A through D at the expense of lower precision for E through G. This deficiency is secondary to the most critical problem. Note that there is no comparison linking any of items A through D to any of items E through G. The requirement of connectivity is not satisfied, and there is no way to place all items on a common scale.

DESIGNS FOR INDIVIDUAL-LEVEL ESTIMATION

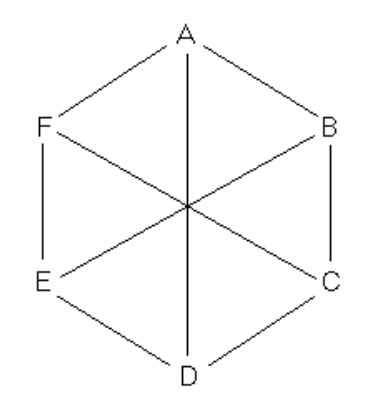
Cyclic designs have been proposed to ensure both *balance* and *connectivity*. Let t equal the number of items, and k equal the number of paired comparisons. Consider a design where $t = 6$, with items A through F. There are $k = 1/2(t)(t\text{-}1) = 15$ possible combinations. However, an incomplete design with just nine comparisons (Figure 2) can satisfy both *balance* and *connectivity*.

Figure 2

	A	B	C	D	E	F
A						
B	X					
C		X				
D	X		X			
E		X		X		
F	X		X		X	

We can represent this design geometrically in Figure 3:

Figure 3



Each item is involved in three comparisons, and the design has symmetric connectivity. In practice, cyclic designs where $k \geq 1.5t$ comparisons often provide enough information to estimate useful scores at the individual level. The fractional design above satisfies that requirement.

With an odd number of elements such as 7, asking 1.5*t* comparisons leads to an infeasible number of comparisons: 10.5. One can round up to the nearest integer. A cyclical design with 7 elements and 11 comparisons leads to one of the items occurring an extra time relative to the others. This slight imbalance is of little consequence and does not cause bias in estimation when using linear methods such as HB, logit, or regression. When we consider a customized survey in which each respondent can receive a randomized order of presentation, the overall pooled design is still in nearly perfect balance. There is no pattern to the imbalance across respondents.

For the purposes of simplicity, the examples presented thus far have illustrated designs with 1.5*t* comparisons. If respondents can reliably complete more than 1.5*t* questions, you should increase the number to achieve even more precise results.

DESIGNS FOR AGGREGATE ANALYSIS

If the number of comparisons *k* is less than the number of items *t*, there isn't enough information to estimate reliable individual-level parameters (though with HB estimation, it would be able to provide an estimate for items not shown to a respondent based on a draw from the population distribution). In that case, respondents generally are pooled for group-based analysis. At the respondent level, one must abandon the principle of connectivity[1] in favor of the goal that each respondent should be exposed to as many items, as many times as possible. Across respondents, connectivity still holds (suitable for aggregate analysis). Given enough respondents, randomized designs approximate the complete design.

If one assumes sample homogeneity, with a large enough sample size each respondent would only need to answer a single question to lead to stable aggregate estimates. Given the advantages of capturing heterogeneity, it makes sense whenever possible to ask enough comparisons of each respondent to utilize latent class or HB.

[1] When $k<(t-1)$, connectivity is no longer attainable at the individual-level.

RESPONSE SCALE: CHOICE VERSUS RATINGS?

Choice is more natural and easy for respondents to complete than ratings. However, ratings seem to provide more statistical information for each comparison. Not only do we learn which item is preferred, but we learn by how much it is preferred.

The following conditions would favor the use of choice rather than ratings:

Relatively large sample size

More emphasis on aggregate rather than individual-level estimation

Concern that respondents may have difficulty using a rating scale

Desire to rescale the data to probability scaling (described below)

The following conditions favor the use of ratings rather than choice:

Relatively small sample size

Desire for stable individual-level estimation

Respondents are not confused by the rating scales

DEALING WITH THE RELATIVE SCALE

Depending on your viewpoint, one weakness of MPC is that the scores reflect a relative (rather than absolute) interval scale. Furthermore, they include negative and positive values. With monadic ratings, a "7" might be associated with "somewhat important," or a 9 with "extremely desirable." With MPC, the scores are based on relative comparisons among the items included in the study. The respondent might very much like or dislike *all* items, but MPC cannot determine this.

There are clever ways to include items in an MPC experiment tied to something meaningful, such as a monetary equivalent. Including the item "Receive $20 off" in a preference experiment scales the items with respect to something more concrete, such as the value of $20. With mutually exclusive prohibitions, you could include multiple monetary equivalents (it wouldn't make sense to compare these monetary-based items directly, as the response is obvious).

If the MPC experiment uses choices rather than ratings, one can easily transform the derived scores (assuming the parameters were fit using a logit model specification) to a probability scale. Assuming zero-centered raw scores, one can use the transform:

$$P^x = 100\ [e^x/(1+e^x)]$$

This transform places the scores on a 0 to 100 scale. The interpretation of a "70" is that this item would be chosen 70% of the time on average when compared to the other items we measured.

CASE STUDY #1: PREFERENCES FOR FAST FOOD RESTAURANT ATTRIBUTES

I conducted a pilot study using a paper-based questionnaire and a convenience sample of 54 respondents. Respondents were instructed to rate 10 items related to fast food restaurants:

1. Has clean eating area (floors, tables and chairs)
2. Has clean bathrooms
3. Has some health food items on the menu
4. Typical wait is less than 5 minutes in line
5. Prices are very reasonable
6. Typical wait is about 10 minutes in line
7. Your order is always filled right
8. Has a play area for children
9. Food tastes wonderful
10. Restaurant gives generously to charities

Items 4 and 6 were prohibited from being paired. Respondents rated the items using both monadic ratings (9-point desirability scale) and Paired Comparisons (8-point graded comparison scale, rather than choice). I used a cyclical design for the MPC questions and employed two versions of the questionnaire (27 respondents per version) to control for order effects:

Version 1	***Version 2***
I. Monadic ratings (10 items) II. MPC (15 pairs) III. Monadic ratings (10 items, repeat of section I, but with rotated order) IV. MPC (15 pairs, 3 pairs were repeats of pairs from section II)	I. MPC (15 pairs, same as version 1 section IV) II. Monadic ratings (10 items, same order as version 1 section III) III. MPC (15 pairs, same as version 1 section II) IV. Monadic ratings (10 items, same order as version 1 section I)

With a 10-item design, there are 45 possible paired comparisons. Respondents completed 27 of these possible pairings, and three of the pairings were repeated to assess test-retest reliability (for a total of 30 paired comparisons). Most respondents completed the entire questionnaire in about 8 to 10 minutes.

SCALE UTILIZATION

Do respondents use the monadic rating and MPC scales differently? I examined how many unique scale points respondents used within the first 10 questions of each question type. For monadic ratings, respondents used on average 4.6 of the 9 available scale points (4.6/9 = 51% of the scale points). For MPC, respondents used on average 5.4 of the 8 scale points (5.4/8 = 68% of the scale points).

It is worth noting that most of the 10 items studied were considered quite desirable in a fast food restaurant. It shouldn't surprise us that respondents tended to rate many items near the top of the scale in the monadic ratings. This resulted in fewer overall scale points used and many ties for the most desirable items. Over the first monadic ratings section (10 questions), respondents tied 4.1 items on average at the most desirable rating used.

In summary, for this study, respondents used the monadic rating scale in a more limited manner than MPC, using fewer scale points than MPC and tying many of the items.

SCALE USE RELIABILITY

Which type of scale do respondents use more reliably? MPC questions are more challenging to answer than monadic ratings, so one might question how reliably respondents use the scale.

Respondents repeated all 10 monadic ratings, and 3 of the MPC questions. The test-retest root mean squared error (RMSE) for monadic ratings and MPC was 1.07 and 0.99, respectively. A RMSE of 1.07 means that when respondents rated the same item a second time, their responses were on average 1.07 scale units apart. From this, one can only conclude that respondents completed both types of questions with about equal reliability. (Note that the monadic ratings used a 9-point scale, and MPC 8. Therefore, monadic ratings are at a slight disadvantage in this analysis. Generally, it is harder to achieve a high test-retest reliability when there are a greater number of scale points.)

In summary, this study suggests that respondents use the MPC scale at about the same degree of reliability as monadic ratings.

SIMILARITY OF PREFERENCE SCORES

Are the preference scores different depending on the method used (monadic ratings vs. MPC)? Table 1 shows the estimated scores from the two methods, sorted by preference. I've scaled the data so that the scores are zero-centered, and the range from best to worst item is 100 points.

Table 1
Desirability Scores by Method

	Monadic Ratings	Paired Comparisons
Clean eating area	30.1	50.0
Tastes wonderful	27.5	32.2
Clean bathrooms	27.4	16.0
Get order right	26.5	15.8
Reasonable prices	23.6	26.1
Wait 5 minutes	16.6	9.6
Health food items	-14.8	-23.4
Play area	-24.2	-30.1
Gives to charities	-42.7	-46.1
Wait 10 minutes	-69.9	-50.0

With the exception of one item (Reasonable prices), the rank order of the items is the same between the two methods[2]. For this data set, MPC seems to show greater discrimination among the top few items than monadic ratings.

In summary, for this study, monadic ratings and MPC result in nearly the same rank ordering of the items. Whether the relative scores themselves are equivalent is questionable. For example, MPC seems to find greater differences among the top few items for this data set.

PREDICTIVE VALIDITY

The last five paired comparison questions were held out for validity testing. This form of validity testing is more a test of internal reliability than true validity (given that data are used from the same respondents and the same questionnaire instrument). Nonetheless, I'll still use the term *validity*, since it is commonly referred to as such in methodological research studies.

I calculated a hit rate using scores from the monadic ratings and MPC to predict whether a respondent would be expected to choose the left or right item in each of the five holdout pairs. If the prediction matches the respondent's actual choice, a hit is scored; otherwise, a miss is scored. In the case of a predicted tie (there were many ties when using monadic ratings) I scored the hit rate for that comparison as 50%.

The scores from monadic ratings resulted in a hit rate of 72.2%. Scores from the paired comparison produced a much higher hit rate of 85.0%. Of course, this comparison naturally favors MPC, since the holdout criterion is also paired comparisons.

DISCRIMINATION AMONG ITEMS

As mentioned earlier in this paper, achieving discrimination among items is very important. Clients commonly ask whether there is a statistically significant difference between items. I hypothesized that MPC should result in greater discrimination relative to monadic ratings. To test this assertion, I compared the most preferred (on average) item (#1 — Clean eating area)

[2] My experience is that due to social desirability bias, price importance is sometimes lower for self-explicated than derived methods of preference measurement. While this is a possible explanation for this difference in the rank order of scores between the methods, I'd like to see this finding replicated across other studies before drawing that conclusion here.

with all other items using matched sample t-tests. The results for monadic ratings and MPC under individual-level OLS estimation is shown in Table 2.

Table 2
Discrimination among Items
T-Tests, Most Preferred (Item 1) versus All Others

	Monadic Ratings	Paired Comparisons
Item 1 vs. Item 2	0.9	6.1
Item 1 vs. Item 3	7.1	10.7
Item 1 vs. Item 4	2.0	5.6
Item 1 vs. Item 5	0.9	4.0
Item 1 vs. Item 6	11.9	15.1
Item 1 vs. Item 7	0.6	5.5
Item 1 vs. Item 8	6.2	10.8
Item 1 vs. Item 9	0.3	3.4
Item 1 vs. Item 10	9.3	15.1

For MPC, all comparisons are significant beyond the 95% confidence level. For monadic ratings, four of the comparisons are not statistically significant. The t-values are generally higher on all items for MPC vs. monadic ratings. The average t-value for monadic ratings is 4.4 as opposed to 8.5 for paired comparisons. (Though my choice of contrasting all levels with respect to the most preferred item may have influenced the outcome, in the second case study to follow I used a central item for contrasts and confirmed these general findings.) This suggests that MPC can achieve greater discrimination among the items given the same sample size. This of course also means that by using MPC, one can achieve equivalent discrimination as monadic ratings using a smaller sample size.

Discrimination between Groups

Another common question clients ask is whether the preference scores differ among respondent groups. Can the use of paired comparisons improve our ability to find differences among segments? I collected three demographic variables: Gender, Age, and whether respondents had children between the ages of 2 to 9. Respondents were divided into two groups for each of these variables, and F-tests were conducted on all of the items for the two measurement methods.

With 3 F-tests for each of 10 items, there are a total of 30 comparisons. Using critical values associated with the 90% confidence level, one would expect 3 comparisons to be significant even from random data.

With monadic ratings, I found just one significant difference between groups. Preference for whether the restaurant has a play area for children was significantly different (p=.07) for respondents who had small children versus those that did not. Under MPC analysis, I also found this relationship, but with a p-value of 0.003. With MPC, I found 5 significant differences between segments.

This suggests that paired comparisons improve our ability to find significant differences among segments relative to monadic ratings.

CASE STUDY #2: IMPORTANCE OF SERVER ATTRIBUTES

Together with Michael Patterson, at HP, we conducted another methodological experiment, focusing on measuring the importance of a list of 20 attributes related to servers. We interviewed 374 respondents (from a list provided by HP) using a web-based survey.

Respondents were randomly given either monadic ratings (n=137), paired comparisons (n=121), or another newer technique called Maximum Difference Scaling (n=116), also known as Best/Worst Scaling. Additional holdout tasks were given, in which respondents ranked four sets of three items. In Steve Cohen's paper in this same volume, he reports the results of the Best/Worst scaling. We'll focus on the monadic vs. paired comparisons here.

Because we were interviewing using computers, we could time how long it took respondents to complete each scaling exercise. For the paired comparison experiment, we showed each respondent 30 pairs (1.5 times the number of items). The average time to complete was 1.5 minutes (about 5 seconds per "click") for monadic ratings and 5 minutes for paired comparisons (about 11 seconds per "click").

In addition, we asked respondents a number of questions related to their experience with the monadic ratings and paired comparison questions. (We used similar items as reported by Huber *et al*., 1991.)

Table 3
Qualitative Evaluation

Using a scale where a 1 means "strongly disagree" and a 7 means "strongly agree", how much do you agree or disagree that the <previous section>…

	Monadic (n = 137)	Paired Comparison (n = 121)
…was enjoyable*	4.3	4.0
…was confusing*	2.4	2.9
…was easy*	5.6	5.2
…made me feel like clicking answers just to get done	3.2	3.1
…allowed me to express my opinions	4.9	4.6

(*significant difference between groups, $p<0.05$)

Even though there was a statistically significant difference between the monadic and paired comparison groups for the first three items, I don't believe these are very consequential. For example, even though respondents found paired comparison to be a bit more confusing than monadic, it still rates a relatively low 2.9 on a 7-point scale, meaning that respondents didn't find either task very confusing. In general, these evaluations suggest that paired comparisons are quite doable over the web, in a computer interviewing format.

PREDICTIVE VALIDITY

With the server study, we also included some ranking questions, for three items at a time. We asked four sets of these ranking questions, both before the scaling task and after (test-retest). Each ranking of three items leads to three implied pairs. Using the monadic and paired comparison importance scores, we predicted the preference for these implied pairs for the holdout ranking tasks.

Relative to test-retest reliability, the predictive hit rates were 85% and 88% for monadic ratings and paired comparisons, respectively. As with the restaurant study, paired comparisons are more accurate than monadic ratings in predicting the holdouts.

Between-Item Discrimination

We again repeated the exercise of conducting paired t-tests between items, though we chose to contrast the item with the average (rather than the highest) importance versus all others. The average t values were 3.3 and 6.3 for monadic and paired comparison groups, respectively. For the restaurant case study, recall that paired comparison data also resulted in a t value about double that of monadic ratings.

Discrimination between Groups

We included a number of usage, attitudinal and profiling questions in the server study in addition to the item measurement questions. Using these questions, we formed 19 segmentation variables and divided respondents into two categories for each. We then performed 19 x 20 = 380 separate F-tests (one for each of the 19 segmentation variables for each of the 20 importance scores.) With critical values associated with the 95% confidence level, we'd expect to find 19 significant differences by chance. We found 30 significant differences for monadic (22 if standardizing the data within respondent), and 40 for paired comparisons. As with the restaurant case study, we again found that paired comparison ratings show greater discrimination between groups.

Conclusion

The Method of Paired Comparisons (MPC) has a long history for measuring, among other things, the importance or preference for items. MPC can be useful in a variety of marketing research, psychometric, social research, and econometric settings. MPC questions offer a good alternative to the commonly-used monadic ratings scales.

For a modest methodological study of 54 respondents and a real-world study with a larger sample size, we compared monadic ratings to MPC with respect to three measures of performance: validity (hit rates), discrimination among items, and discrimination among respondent groups. On all three measures and for both studies, MPC was superior to monadic ratings.

Before deciding to use the method of paired comparisons in future studies, the reader should review Steve Cohen's paper on Maximum Difference (Best/Worst) Scaling in this same volume. Maximum Difference Scaling can be considered a more sophisticated extension of paired comparisons. Steve's results suggest that Maximum Difference Scaling may work even better than paired comparisons.

APPENDIX
ESTIMATION AND CODING PROCEDURES FOR PAIRED COMPARISON EXPERIMENTS

RATINGS-BASED RESPONSE

Paired Comparison Experiments can use either a choice (left or right) or a graded ratings response (for example, a 1 to 8 scale, from strongly prefer item on left to strongly prefer item on right). With a ratings-based response scale, among other techniques, individual-level OLS or hierarchical Bayes regression are possible estimation methods.

Construct a vector of independent variables of length of the number of items minus one. In dummy coding, to avoid linear dependency, one of the items is arbitrarily chosen as the reference ("omitted" variable), and is fixed at a beta of zero. The other parameters are measured as contrasts with respect to that zero reference point. Each paired comparison question is coded as a single row in the independent variable matrix. If an item is shown on the left, it receives a -1, if on the right, a +1, otherwise the independent variable is 0. We usually center the respondent's rating, by subtracting off the midpoint of the scale (for example, with a 1 to 8-point scale, a 1 becomes -3.5, a 2 becomes -2.5, etc.). (Though, if one estimates the constant, centering or not has no effect.)

For example, assume an experiment with 8 items, where the 8th parameter is the omitted (reference) item. Further assume an 8-point response scale. The following represents coding for the first three rows (questions):

	Independent Variables	Dependent Variable
Pair #1: item 1 on left, item 4 on right Response = 4 (slightly prefer left)	-1 0 0 1 0 0 0	-0.5
Pair #2: item 8 on left, item 3 on right Response = 8 (strongly prefer right)	0 0 1 0 0 0 0	+3.5
Pair #3: item 7 on left, item 5 on right Response = 1 (strongly prefer left)	0 0 0 0 1 0 -1	-3.5

Even though we have centered the dependent variable, it is probably useful to estimate a constant in addition to the coded independent variables shown above (which constant is ignored during subsequent analysis). The constant reflects any remaining left- vs. right-hand response tendency, after accounting for all the information explained by the items shown in the pairs.

In the example above, after estimation, the betas are the scores for the 7 explicitly coded variables, and the beta for the 8th (reference) item is 0.

CHOICE-BASED RESPONSE

With a choice-based response, among other possibilities, logit, latent class or hierarchical Bayes estimation (logit-based "lower" model) are recommended approaches. Dummy coding is again suggested. If coding for use with Sawtooth Software's Latent Class or CBC/HB v2 systems (using the new "#" notation in the .EFF file), each paired comparison is coded as two separate rows in the design matrix. Dummy coding is again used, with 1s coded if the level is present, and 0 if absent. An additional dummy code reflecting whether the current item is the left (1) or right concept (0) can also be included in the vector of independent variables, and has a similar function and interpretation as the constant as described in the previous ratings-based example. It is similarly ignored during subsequent analysis.

A NOTE ON DUMMY VERSUS EFFECTS CODING

Effects-coding is a type of dummy-coding procedure in which all elements of the independent variable vector are equal to -1 rather than 0 when the omitted item is present. With effects-coding, the estimated betas are zero-centered (the reference beta is negative the sum of the other betas). With OLS or logit estimation, the model fit is identical whether using dummy-coding or effects-coding, and the parameters are also identical (within a constant).

But, with HB, the results, while extremely similar, are not equivalent (Johnson, 1999). Given enough information within the unit of analysis, the point estimate of the betas should be identical whether using either method (again, within a constant). However, if using effects-coding, the variance of the reference parameter is inflated (relative to the variance of the other parameters). With dummy-coding, there is evidence (that can be seen after zero-centering all parameters) that the variance of the reference parameter (relative to the other parameters) is deflated. Based on preliminary evidence, the understatement of variance in the dummy-coding case seems less extreme than the overstatement of variance in the effects-coding case. This is particularly the case as the number of items in the experiment increases. Dummy-coding for paired comparison experiments seems more widely used and accepted in the industry than effects coding, though theoretically either should be equally useful.

In the weeks following the Sawtooth Software conference, I have been in touch with Peter Lenk (University of Michigan) and our chairman Rich Johnson regarding the issues of dummy coding versus effects coding for HB. We are investigating the possibility that the assumptions of the prior distribution within the HB routines used here may be causing the effects noted directly above. Perhaps we'll be able to share more regarding that later.

REFERENCES

Bradley, R. A. and Terry, M. E. (1952), "The Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons," *Biometrika*, 39, 324-45.

David, H. A. (1969), *The Method of Paired Comparisons*, Charles Griffin & Company Limited, London.

Fechner, G. T. (1860), *Elemente der Psychophysik*. Leipzig: Breitkopf and Hartel.

Huber, Joel C, Dick R. Wittink, John A. Fiedler, and Richard L. Miller (1991), "An Empirical Comparison of ACA and Full Profile Judgments," *Sawtooth Software Conference Proceedings*, 189-202.

Johnson, Richard M. (1999), "The Joys and Sorrows of Implementing HB Methods for Conjoint Analysis," technical paper available at www.sawtoothsoftware.com.

Kendall, M. G. (1955), "Further Contributions to the Theory of Paired Comparisons," *Biometrics*, 11, 43-62.

Thurstone, L. L. (1927), "A Law of Comparative Judgment," *Psychology Review* 34: 273-86.

MAXIMUM DIFFERENCE SCALING: IMPROVED MEASURES OF IMPORTANCE AND PREFERENCE FOR SEGMENTATION

STEVEN H. COHEN
SHC & ASSOCIATES

INTRODUCTION

The measurement of consumer preferences has long been an area of interest to both academic and practicing researchers. Accurate measurement of preferences allows the marketer to gain a deeper understanding of consumers' wishes, desires, likes, and dislikes, and thus permits a better implementation of the tools of the marketer. After measuring preferences, a common activity is market segmentation, which permits an even more focused execution of the marketing mix.

Since the mid-1950s, marketing researchers have responded to the needs of management by conducting market segmentation studies. These studies are characterized by the collection of descriptive information about benefits sought, attitudes and beliefs about the category, purchase volume, buying styles, channels used, self, family, or company demographics, and so on. Upon analysis, the researcher typically chooses to look at the data through the lens of a segmentation basis. This basis is either defined by preexisting groups – like heavy, medium, and light buyers or older versus younger consumers – or defined by hidden groups uncovered during an in-depth statistical analysis of the data – benefits segments, attitude segments, or psychographic segments. Finally, the segments are then cross-tabulated against the remaining questions in the study to profile each group and to discover what characteristics besides the segmentation base distinguish them from one another.

Quite often, researchers find that preexisting groups, when different, are well distinguished in obvious ways and not much else. Wealthier consumers buy more goods and services, women buy and use products in particular categories more than men, smaller companies purchase less *and* less often than larger companies, and so on. However, when looking at buying motivations, benefits sought, and their sensitivity to the tools of marketers (e.g. price, promotions, and channel strategies), members of preexisting groups are often found to be indistinguishable from one another.

This realization has forced researchers to look to *post hoc* segments formed by a multivariate analysis of benefits, attitudes, or the like. This focus on benefits, psychographics, needs and wants, and marketing elasticities as means of segmentation has gained favor since the early work of Haley (1985) and is the mainstay of many market segmentation studies currently conducted. Product benefits are measured and then people with similar sets of benefits are termed "benefit segments." The utility of a focus on *post hoc* methods has been widely endorsed by marketing strategists (Aaker, 2001):

> *"If there is a 'most useful' segmentation variable, it would be benefits sought from a product, because the selection of benefits can determine a total business strategy."*

Using Benefits Segmentation as our example, we compare three methods of measuring preferences for benefits using a split-sample design. Twenty benefits were presented to IT managers in an online survey. The first method uses a traditional ratings task. Each person performed 20 "mouse clicks" to rate the items on a 1-9 scale to fulfill the task. The second method uses 30 paired comparisons (cyclical design, chosen from the 20*19 = 380 possible pairs), yielding 30 mouse clicks. The third method uses Maximum Difference Scaling (described below), showing 20 sets of four benefits (quads) and asking the respondent to choose the Most Important and Least Important from each quad, resulting in 30 mouse clicks.

This paper is organized as follows. We first briefly review the standard practices of benefit measurement and benefit segmentation and, along the way, point out their deficiencies. We then introduce the reader to Maximum Difference Scaling, a method that we believe is a much more powerful method for measuring benefit importance – a method that is *scale-free.* We then present the results of the split-sample study described above. After that we describe how Maximum Difference Scaling can be combined with Latent Class Analysis to obtain benefit segments. We then describe an example of how both the traditional and the newer methods were used in a cross-national segmentation study of buyers of an industrial product conducted several years ago.

TRADITIONAL SEGMENTATION TOOLS

The two-stage or "tandem" segmentation method has been used for over twenty years (Haley, 1985), and has been described by Myers (1996) as follows:

1. Administer a battery of rating-scale items to a group of consumers, buyers, customers, etc. These rating scales typically take the form of agree/disagree, describes/does not describe, important/not important ratings. Scales of five, seven, ten, or even 100 points are used.

2. The analyst then seeks to reduce the data to a smaller number of underlying dimensions or themes. Factor Analysis of the rating scale data, using either the raw ratings or some transformation of the ratings (like standardization) to obtain better statistical properties, is most often performed. The analyst then outputs the factor scores, one set of scores for each respondent.

3. The factor scores are passed to a Cluster Analysis, with k-means Cluster Analysis being the most preferred and the most often recommended by academic researchers (Punj and Stewart (1983). K-means is implemented in SAS as Proc Fastclus and in SPSS as Quickcluster.

4. The clusters are profiled. A cross-tabulation of group, cluster, or segment membership is created against all the other significant items in the survey.

Many of us have used rating scale data in factoring and in segmentation studies. The major problem tends to be response scale usage. Quite often we choose positively-worded important items to include in a survey. The result is that the range of mean item scores is small and tends to be (at least in the USA) towards the top-end of the scale.

The best-known response styles are acquiescence bias, extreme responding, and social desirability (Paulhus, 1991). There is ample evidence (Chen, Lee, and Stevenson, 1995; Steenkamp and Baumgartner, 1998; ter Hofstede, Steenkamp, and Wedel, 1999; Baumgartner

and Steenkamp, 2001) that countries differ in their response styles. We note that scalar inequivalence is *less likely* to occur when collecting constant sum or ranking data. Constant sum data forces trade-offs and avoids yea-saying. However, constant sum data may be difficult to collect if there are many items. Another alternative may be ranking the benefits. However, the major advantage of ranking – each scale point is used once and only once – may be outweighed by the fact that ranking suffers from order effects, does not allow ties, and is not appropriate when absolute scores are needed (e.g. purchase intent ratings).

Hence, we conclude that we would like a rating method that does not experience scale use bias, forces trade-offs, and allows each scale point to be used once and only once.

For grouping people, the tandem method of segmenting respondents using factor scores followed by Cluster Analysis is a very common practice. Cluster Analysis may be characterized as a *heuristic* method since there is no underlying *model* being fit. We contend that, while using Factor Analysis may get rid of the problems associated with correlated items, it introduces the problems of which factoring method to use, what type of rotation to use, factor score indeterminacy, and the selection of the final number of factors.

Deriving patterns from Factor Analysis becomes problematic when ratings have systematic scale use biases and large item inter-correlations owing to scale use. For example, when using a rating scale in a segmentation analysis, the first dimension uncovered in a Factor Analysis often tends to be a general factor. Using this factor in a Cluster Analysis will often uncover a "high rater" segment or a "general" segment. Additional partitions of the data may uncover meaningful groups who have different needs, but only after separating out a group or two defined by their response patterns. This approach is especially dangerous in multi-country studies, where segments often break out on national lines, more often due to cultural differences in scale use, than to true differences in needs. Indeed, as noted by Steenkamp and ter Hofstede:

> *"Notwithstanding the evidence on the biasing effects of cross-national differences in response tendencies, and of the potential lack of scalar equivalence in general, on the segmentation basis, it is worrisome to note that this issue has not received much attention in international segmentation research. We believe that cross-national differences in stylistic responding is one of the reasons why international segmentation studies often report a heavy country influence."*

Using Cluster Analysis alone has a number of limitations. These include forcing a deterministic classification (each person belongs absolutely to one and only one segment) and poor performance when the input data are correlated. In such situations, highly correlated items are "double counted" when perhaps they should be counted only once.

Even more egregious is the sensitivity of Cluster Analysis to the *order of the data*. Simply put, sort the data in one direction and obtain a solution. Then sort the data in the opposite way, specify the same number of clusters as in the first analysis. Now compare them. Our experience shows that using the clustering routines found in SAS and SPSS often yield an overlap of the two solutions in the 60% - 80% range. Not a very satisfying result, we contend.

Academic research has rightly pointed out other deficiencies of the two-stage or tandem approach of Factor Analysis followed by Cluster Analysis [see DeSarbo et al (1990); Dillon,

Mulani, & Frederick (1989); Green & Krieger (1995); Wedel & Kamakura (1999); and, Arabie & Hubert (1994)]. While the frequent use of the tandem method is unmistakable because of its ease of implementation with off-the-shelf software, most practicing researchers have simply failed to hear or heed these warnings. The bluntest assessment of the weakness of the tandem method may be attributed to Arabie & Hubert (1994):

> "*Tandem clustering is an out-moded and statistically insupportable practice.*" (italics in original).

While Chrzan and Elder (1999) discuss possible solutions to the tandem problem and attempt to dismiss Arabia & Hubert's concerns, the fact remains that their solution requires a heavy dose of analysis even before attempting to factor or cluster. The final segmentation analysis may use all or a selection of the raw variables, or may use the tandem method, depending upon the items, their intercorrelations, and other characteristics of the data.

The next section describes the use of Maximum Difference Scaling instead of rating scales to measure the *relative importance* of benefits and then we discuss the results of the split-sample study, comparing the IT managers' responses across ratings, simple paired comparisons, and the MaxDiff method.

We follow that section with a brief discussion of the advantages of Latent Class Analysis (LCA) over Cluster Analysis, as a method for uncovering market segments with similar benefit importances. We conclude with an illustration of using benefit segmentation with LCA in an international segmentation study of IT managers (different sample than the one used in the earlier analysis).

MAXIMUM DIFFERENCE SCALING

Maximum Difference Scaling (*MaxDiff*) is a measurement and scaling technique originally developed by Jordan Louviere and his colleagues (Louviere, 1991, 1992; Louviere, Finn, and Timmermans, 1994; Finn & Louviere, 1995; Louviere, Swait, and Anderson, 1995; McIntosh and Louviere, 2002). Most of the prior applications of MaxDiff have been for use in Best-Worst Conjoint Analysis. In applying MaxDiff to B-W Conjoint, the respondent is presented with a full product or service profile as in traditional Conjoint. Then, rather than giving an overall evaluation of the profile, the respondent is asked to choose the attribute/level combination shown that is most appealing (best) and least appealing (worst).

We apply this scaling technique instead to the measurement of the importance of product benefits and uncovering segments. This discussion follows the one made by Cohen & Markowitz (2002).

MaxDiff finds its genesis in a little-investigated deficiency of Conjoint Analysis. As discussed by Lynch (1985), additive conjoint models do not permit the separation of importance or weight and the scale value. Put another way, Conjoint Analysis permits *intra-attribute* comparisons of levels, but does not permit *across attribute* comparisons. This is because the scaling of the attributes is unique to each attribute, rather than being a method of global scaling.

Maximum Difference Scaling permits intra- and inter-item comparison of levels by measuring attribute level utilities on a common, interval scale. Louviere, Swait, and Anderson

(1995) and McIntosh and Louviere (2002) present the basics of MaxDiff, or Best-Worst scaling. To implement maximum difference scaling for benefits requires these steps.

- Select a set of benefits to be investigated.
- Place the benefits into several smaller subsets using an experimental design (e.g. 2^k, BIB, or PBIB are most common). Typically over a dozen such sets of three to six benefits each are needed, but each application is different.
- Present the sets one at a time to respondents. In each set, the respondent chooses the most salient or important attribute (the best) and the least important (the worst). This best-worst pair is ***the pair*** in that set that has the Maximum Difference.
- Using four items in the task (for example) and collecting the most and least in each task will result in recovering 5 of the 6 paired comparisons. For example, with items A, B, C, and D in a quad there are 4*3/2 = 6 pairs. If A were chosen as most and D as least, the only pair that we do not obtain a comparison of is the B-C pair.
- Since the data are simple choices, analyze the data with a multinomial logit (MNL) or probit model. An aggregate level model will produce a total sample benefit ordering. Using HB methods will result in similar results as in an aggregate MNL model.
- Analyze pre-existing subgroups with the same statistical technique.
- To find benefit segments, use a Latent Class multinomial logit model.

The MaxDiff model assumes that respondents behave *as if* they are examining every possible pair in each subset, and then they choose the most distinct pair as the best-worst, most-least, *maximum difference* pair. Thus, one may think of MaxDiff as a more efficient way of collecting paired comparison data.

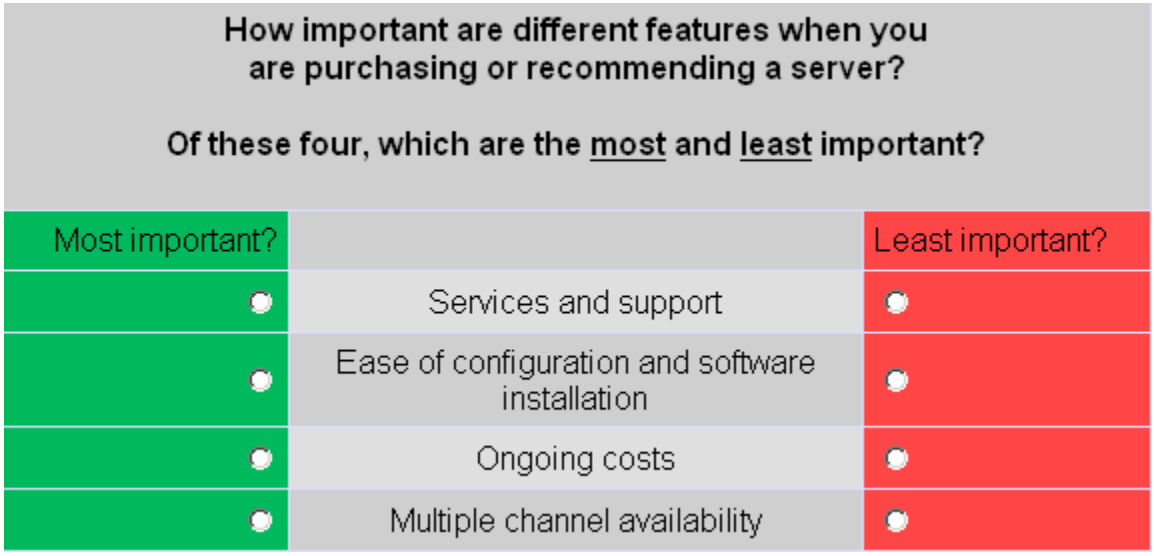
Properly designed, MaxDiff will require respondents to make trade-offs among benefits. By doing so, we do not permit anyone to like or dislike all benefits. By definition, we force the relative importances out of the respondent. A well-designed task will control for order effects. Each respondent will see each item in the first, second, third, etc. position across benefit subsets. The design will also control for context effects: each item will be seen with every other item an equal number of times.

The MaxDiff procedure will produce a unidimensional interval-level scale of benefit importance based on nominal level choice data. Because there is only one way to choose something as "most important," there is no opportunity whatsoever to encounter bias in the use of a rating scale. Hence, there is no opportunity to be a constant high/low rater or a middle-of-the-roader. The method forces respondents to make a discriminating choice among the benefits. Looking back to the observations by Steenkamp and ter Hofstede, we believe that this method overcomes very well the problems encountered in cross-national attribute comparisons that are due to differences in the use of rating scales across countries. The MaxDiff method is easy to complete (respondents make two choices per set), may also control for potential order or context biases, and is rating scale-free.

COMPARISONS OF RESULTS FROM THE THREE METHODS USED IN THIS STUDY

IT managers from an online panel were recruited and assigned to do one of the benefits evaluation tasks: 137 did ratings, 121 did paired comparisons, and 116 did the MaxDiff method.

Below is an example of a MaxDiff task for this study:



Immediately after the benefits evaluation, we asked the respondents to tell us their perceptions of the task they performed. As can be seen in Table 1, on a seven point scale of agree-disagree, all tasks were evaluated at about the midpoint of each scale, with ratings being slightly higher rated (e.g. easier) than the paired comparison or MaxDiff tasks. On average, the paired comparisons and MaxDiff task took about three times as long to complete than the ratings, but on the basis of "seconds per click," the ratings task took about ½ as long as the other two tasks, indicating the greater involvement and thought that is required.

Table 1
Qualitative Evaluation

Using a scale where a 1 means "strongly disagree" and a 7 means "strongly agree", how much do you agree or disagree that the <previous section>…

	Monadic (n = 137)	**Paired Comparison (n = 121)**	**Best/Worst (n = 116)**
…was enjoyable	4.3 (b, c)	4.0 (a)	3.8 (a)
…was confusing	2.4 (b, c)	2.9 (a)	3.2 (a)
…was easy	5.6 (b, c)	5.2 (a)	5.1 (a)
…made me feel like clicking answers just to get done	3.2	3.1 (c)	3.6 (b)
…allowed me to express my opinions	4.9 (c)	4.6	4.3 (a)

("a" means, significantly different from column a, $p<0.05$, etc.)

	Monadic (n = 137)	Paired Comparison (n = 121)	Best/Worst (n = 116)
Mean time to complete exercise	97 seconds	320 seconds	298 seconds
Seconds per mouse click	4.9 sec./click	10.7 sec./click	9.9 sec./click

The ratings task resulted in a 1-9 score for each of the 20 benefits. For each respondent, we chose 30 pairs (from three versions of a cyclic design) of the total number of 380 to be rated. Since MaxDiff requires two judgments per task, we chose 15 quads (from three versions of a computer-generated, balanced plan) for use in the MaxDiff task. Hence, we tried as best as we could to equalize the total number of clicks in the pairs and MaxDiff tasks. Both the paired comparison task and the MaxDiff task were analyzed using HB methods, resulting in 20 utilities for each person, typically ranging from about -4 to +4. The ratings task data suggested respondents often used a limited part of the scale. While the mean scores are very highly correlated across these two methods, the forced discrimination of the MaxDiff task should result in greater differentiation across items.

To test this hypothesis, we performed t-tests of mean benefit differences within each method. That is, we selected a reference item and compared each item's score to the reference item's score. We averaged the t-values obtained as a way to compare results, and we found that the average t-test for the rating scales was 3.3, for the paired comparison the average t-test result was 6.3, and the average for MaxDiff was 7.7. We conclude that the rating scale discriminated least among the items when comparing each one to the other, the MaxDiff results were most discriminating, and the paired comparison task was in between the other two, but closer to MaxDiff.

We then looked at the ability of each method to discover differences across pre-existing groups. We used 19 items from the survey, each with two to five categories and performed F-tests of mean differences across the 19 items. Once again, we had 19*20 = 380 tests within method. By chance, we would expect that 19% of the tests (95% significance level) would be significant. Using the raw ratings data, we found 30 significant differences. Transforming the data to a within-person standardization (an often-used method to remove response biases) only yielded 22 significant differences. The paired comparison method yielded 40 significant differences, while the MaxDiff method resulted in 37, both about twice what would be expected by chance. Once again, we conclude the rating scales are less discriminating than the other two methods, but this time the paired comparison method performed a little better than MaxDiff.

We also gave each person four sets of three of the items as a holdout task, both prior to the scaling task and after (to assess test-retest reliability). We asked the person to rank-order the three items within each of the four sets. We then used the raw ratings or the utilities at the individual level to predict the rankings. Once again, MaxDiff was the winning method. As a percent of test-retest reliability, the hit rates were 97%, 88% and 85% for MaxDiff, Paired Comparisons, and Ratings respectively. While the performance of paired comparisons and ratings is commendable, the MaxDiff performance is quite astonishing, performing at about the same level as test-retest reliability.

We conclude that MaxDiff is certainly a superior method of collecting preferences than a ratings tasks. If we compare MaxDiff to paired comparisons, the evidence is that MaxDiff is superior, but not dramatically so.

LATENT CLASS ANALYSIS

We advocate using the data from the MaxDiff task in a Latent Class (finite mixture) choice model (DeSarbo, Ramaswamy, and Cohen, 1995; Cohen and Ramaswamy, 1998) leading to easily identifiable segments with differing needs. All of this occurs in a scale-free and statistical-model-based environment. For readers not familiar with Latent Class Analysis, we present this short description of its advantages. Interested readers are referred to Wedel and Kamakura (1999) for a more detailed discussion.

Latent Class Analysis (LCA) has a great deal in common with traditional Cluster Analysis, namely the extraction of several relatively homogeneous and yet separate groups of respondents from a heterogeneous set of data. What sets LCA apart from Cluster Analysis is its ability to accommodate both categorical and continuous data, as well as descriptive or predictive models, all in a common framework. Unlike Cluster Analysis, which is data-driven and model-free, LCA is model-based, true to the measurement level of the data, and can yield results which are stronger in the explanation of buyer behavior.

The major advantages of LCA include:

- Conversion of the data to a metric scale for distances is not necessary. LCA uses the data at their original level of measurement.
- LCAs can easily handle models with items at mixed levels of measurement. In Cluster Analysis, all data must be metric.
- LCA fits a statistical model to the data, allowing the use of tests and heuristics for model fit. The tandem method, in contrast, has two objectives, which *may* contradict one another: factor the items, then group the people.
- LCA can handle easily cases with missing data.
- Diagnostic information from LCA will tell you if you have overfit the data with your segmentation model. No such diagnostics exist in Cluster Analysis.
- Respondents are assigned to segments with a probability of membership, rather than with certainty as in Cluster Analysis. This allows further assessment of model fit and the identification of outliers or troublesome respondents.

Perhaps the biggest difference between Cluster Analysis and LCA is the types of problems they can be applied to. Cluster Analysis is solely a descriptive methodology. There is no independent-dependent, or predictor-outcome relationship assumed in the analysis. Thus, while LCA can also be used for descriptive segmentation, its big advantage lies in simultaneous segmentation and prediction.

If we think of a discrete choice model as a predictor-outcome relationship, then we can apply an LCA. In this case, the outcomes or response variables are the Most and Least choices from each set and the predictors are the presence or absence of each of the items in the set, and whether the item was chosen as most (coded +1) or chosen least (coded -1). Recognizing the

need for conducting *post hoc* market segmentation with Choice-based Conjoint Analysis (CBCA), DeSarbo, Ramaswamy, and Cohen (1995) combined LCA with CBCA to introduce Latent Class CBCA, which permits the estimation of benefit segments with CBCA. LC-CBCA has been implemented commercially in a program from Sawtooth Software and from Statistical Innovations.

To summarize this and the prior section:

- We advocate the use of Maximum Difference scaling to obtain a unidimensional interval-level scale of benefit importance. The task is easy to implement, easily understood by respondents and managers alike, and travels well across countries.
- To obtain benefit segments from these data, we advocate the use of Latent Class Analysis. LCA has numerous advantages over Cluster Analysis, the chief among them being that it will group people based on their pattern of nominal-level choices in several sets, rather than by estimating distances between respondents in an unknown or fabricated metric.

The next section discusses an empirical example of the application of these techniques and compares them to results from using traditional tandem-based segmentation tools.

An Example

Our client, a multinational company offering industrial products around the globe, wished to conduct a study of its global customers. The goal of the research was to identify key leverage points for new product design and marketing messaging. Previous segmentation studies had failed to find well-differentiated segments and thus the marketing managers and the researchers were amenable to the use of the techniques described above. For the sake of disguising the product category and the client, we present the category as file servers.

The survey was administered in the client‘s three largest markets: North America, Germany, and Japan. 843 decision-makers were recruited for an in-person interview: 336 in North America, 335 in Germany, and 172 in Japan. The questionnaire contained background information on the respondent’s company, their installed base of brands and products, and a trade-off task that examined new products, features, and prices. The benefit segmentation task is described next.

A list of thirteen product benefits was identified that covered a range of needs from product reliability to service and support to price. Prior qualitative research had identified these attributes as particularly desirable to server purchasers. The benefits tested were:

1. Brand name/vendor reputation
2. Product footprint
3. Expandability
4. Ease of maintenance & repair
5. Overall performance
6. Lowest purchase price
7. Redundant design

8. Reliability
9. Security features
10. Management tools
11. Technical support
12. Upgradeability
13. Warranty policy

A glossary was included with the survey so that respondents understood the meaning of each of these.

To develop the MaxDiff task, we created thirteen sets of four attributes each. Across the sets, every possible pair of items appeared together exactly once. Each benefit appeared once in each of the four positions in a set (first, second, third, and fourth). And, each benefit appeared exactly four times across the thirteen sets. When shown a set of four items, the respondents were asked to choose the item that was the most important and the least important when deciding which server to buy.

In this study, the utilities for the benefits range from positive 3.5 to negative 3.5. We have found that looking at raw utilities may sometimes be unclear to managers. For ease of interpretation, we rescale the utilities according to the underlying choice model. Remember that the model estimated is a multinomial logit (MNL) model, where the sum of the choices after exponentiation is 100%. Hence, if we rescale the utilities according to the MNL model, we will get a "share of preference" for each benefit. If all benefits were equally preferred in this study, then each one's share of preference would be 7.7% (=1/13). If we index 7.7% to be 100, then a benefit with an index score of 200 would result from a share of preference of 15.4% (7.7% times 2). We have found that using this rescaling makes it much easier for managers and analysts to interpret the results. In this paper, we present only the index numbers and not the raw utilities.

By using the standard aggregate multinomial logit model, we obtained the results in Table 2, after rescaling.

Table 2
Overall Product Benefit Importances
from MaxDiff Task

Reliability	571
Overall Performance	277
Ease of Maintenance & Repair	84
Tech support	80
Expandability	59
Management tools	54
Upgradeability	50
Warranty policy	33
Brand name/reputation	27
Redundant design	24
Security features	27
Lowest Purchase Price	10
Product footprint	3

It is obvious that Product Reliability is the most important benefit followed by Overall Performance. In this market, Lowest Purchase Price and Product Footprint are the least important items. We then conducted a segmentation analysis of the Maximum Difference data using the Latent Class Multinomial logit model. A six segment solution was selected with the following segments emerging.

Table 3

Overall Product Benefit Importances
from MaxDiff Task by Benefit Segment

	Easy to Buy & Maintain	Never Breaks	Grows with Me	Help Me Fix It	Brand's the Clue	Managed & Safe
Reliability	264	601	373	554	623	481
Overall Performance	185	197	309	120	228	266
Ease of Maintenance & Repair	100	33	71	157	23	51
Technical support	86	34	34	305	23	58
Expandability	81	30	192	33	21	30
Management tools	53	29	38	23	26	190
Upgradeability	58	12	225	16	10	31
Warranty policy	100	14	21	45	20	29
Brand name/reputation	45	28	10	20	300	7
Redundant design	56	306	11	16	8	10
Security features	31	12	10	6	10	139
Lowest Purchase Price	213	3	5	4	7	5
Product footprint	28	1	2	1	2	3
Percent of total sample	17%	11%	19%	14%	16%	24%
Percent of expected product purchases	31%	19%	23%	9%	9%	9%

In all segments, Reliability is the most important benefit, but its importance varies greatly from a low index number of 264 in the first segment to a high of 623 in the fifth. The second most important benefit is Overall Performance, again ranging widely from 122 to 309. We would call these two "price of entry benefits" in the server category. Respondents in all segments agree, in varying intensities, that Reliability and Performance are what a server is all about. Segment differences reveal themselves in the remaining benefits.

- **Segment 1**, *Easy to Buy & Maintain* (17% of sample and 31% of future purchases), values Lowest Purchase Price (213), Ease of Maintenance & Repair (100), and Warranty Policy (100).
- **Segment 2**, *Never Breaks* (11% of sample and 19% of future purchases) values Redundant Design (306) even more than Performance (197). They have a high need for uptime.
- **Segment 3**, *Grows with Me* (19% and 23%), Values Upgradeability (225) and Expandability (192). They want to leverage their initial investment over time.
- **Segment 4**, *Help Me Fix It* (14% and 9%), values Technical Support (305) and Ease of Maintenance & Repair (157) even more than Performance.

- **Segment 5**, *Brand's the Clue* (16% and 9%), uses the Brand Name/Reputation (300) to help purchase highly reliable (623) servers. As the old saying goes, "No one ever got fired for buying IBM."
- **Segment 6**, *Managed & Safe* (24% and 9%), looks for Management Tools (190) and Security Features (139) when purchasing servers.

Note that Lowest Price, the second lowest index number overall is very important to the first segment, with an index score of 213. The benefits have very large variations across segments, indicating good between-segment differentiation. By looking at the number of servers expected to be purchased, we also provided guidance to management on which segments to target.

SUMMARY

The intent of this paper has been to present practicing researchers with an innovative use of state-of-the-art tools to solve the problems that are produced when using traditional rating scales and the tandem method of clustering. We also compared the results of the suggested method against the traditional tools of rating scales and paired comparisons and found that the new tools provide "better" results.

Therefore, we suggest that practitioners adopt Maximum Difference scaling for developing a unidimensional scale of benefit importance. The MaxDiff task is easy for a respondent to do and it is scale-free, so that it can easily be used to compare results across countries. Furthermore, the tool is easy to implement, relatively easy to analyze with standard software, and easy to explain to respondents and managers alike.

To obtain benefit segments, we suggest using Latent Class Analysis. LCA has numerous advantages over Cluster Analysis. The disadvantages of this latter method are well-known but not often heeded. The benefits of LCA have been demonstrated in many academic papers and books, so its use, while limited, is growing. We hope that this paper will spur the frequent use of these two methods.

This paper has shown that current research practice can be improved and that the traditional methods are lacking and need to be updated. By describing the use of these newer methods and comparing them to traditional methods, we have shown that the modern researcher can overcome scale use bias with Maximum Difference Scaling and can overcome the many problems of Cluster Analysis by using Latent Class models.

We conclude by quoting from Kamakura and Wedel's excellent book (1999) on Market Segmentation:

> *"The identification of market segments is highly dependent on the variables and methods used to define them."*

We hope that this paper demonstrates that the use of different scaling methods can influence the results of preference scaling and also segmentation research.

REFERENCES

Aaker, David A. (1995) *Strategic Market Management*. New York: John Wiley & Sons.

Arabie, Phipps and Lawrence Hubert (1994) "Cluster analysis in marketing research," in *Advanced Methods of Marketing Research*. Richard J. Bagozzi (Ed.). London: Blackwell Publishers, 160-189.

Chang, Wei-Chen (1983). "On using principal components before separating a mixture of two multivariate normal distributions," *Applied Statistics*, 32, 267-75.

Chrzan, Keith and Andrew Elder (1999). "Knowing when to Factor: Simulating the tandem approach to Cluster Analysis," Paper presented at the *Sawtooth Software Conference*, La Jolla, CA.

Cohen, Steven H. and Venkatram Ramaswamy. (1998) "Latent segmentation models." *Marketing Research Magazine*, Summer, 15-22.

Cohen, Steven H. and Paul Markowitz. (2002) "Renewing Market Segmentation: Some new tools to correct old problems." *ESOMAR 2002 Congress Proceedings*, 595-612, ESOMAR: Amsterdam, The Netherlands.

Cohen, Steven H. and Leopoldo Neira. (2003) "Measuring preference for product benefits across countries: Overcoming scale usage bias with Maximum Difference Scaling." *ESOMAR 2003 Latin America Conference Proceedings.* ESOMAR: Amsterdam, The Netherlands.

DeSarbo, Wayne S., Kamel Jedidi, Karen Cool, and Dan Schendel. (1990) "Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups." *Marketing Letters*, 3, 129-146.

DeSarbo, Wayne S., Venkatram Ramaswamy, and Steven H. Cohen. (1995) "Market segmentation with choice-based conjoint analysis." *Marketing Letters,* 6 (2), 137-47.

Dillon, William R., Narendra Mulani, and Donald G., Frederick. (1989) "On the use of component scores in the presence of group structure." *Journal of Consumer Research*, 16, 106-112.

Finn, Adam and Jordan J. Louviere. (1992) "Determining the appropriate response to evidence of public concern: The case of food safety." *Journal of Public Policy and Marketing*, 11:1, 19-25.

Green, Paul E. and Abba Krieger. (1995) "Alternative approaches to cluster-based market segmentation." *Journal of the Market Research Society*, 37 (3), 231-239.

Haley, Russell I. (1985) *Developing effective communications strategy: A benefit segmentation approach.* New York: John Wiley & Sons.

Louviere, Jordan J. (1991) "Best-worst scaling: A model for the largest difference judgments." Working paper. University of Alberta.

Louviere, J.J. (1992). "Maximum difference conjoint: Theory, methods and cross-task comparisons with ratings-based and yes/no full profile conjoint." Unpublished Paper, Department of Marketing, Eccles School of Business, University of Utah, Salt Lake City.

Louviere Jordan J., Adam Finn, & Harry G. Timmermans (1994). "Retail Research Methods," *Handbook of Marketing Research*, 2nd Edition, McGraw-Hill, New York.

Louviere, Jordan J., Joffre Swait, and Donald Anderson. (1995) "Best-worst Conjoint: A new preference elicitation method to simultaneously identify overall attribute importance and attribute level partworths." Working paper, University of Florida, Gainesville, FL.

Lynch, John G., Jr. (1985) "Uniqueness issues in the decompositional modeling of multiattribute overall evaluations: An information integration perspective." *Journal of Marketing Research*, 22, 1-19.

McIntosh, Emma and Jordan Louviere (2002). "Separating weight and scale value: an exploration of best-attribute scaling in health economics," Paper presented at *Health Economics Study Group*. Odense, Denmark.

Myers, James H. (1996) *Segmentation and positioning for strategic marketing decisions*. Chicago: American Marketing Association.

Paulhus, D.L. (1991). "Measurement and control of response bias," in J.P. Robinson, P. R. Shaver, and L.S. Wright (eds.), *Measures of personality and social psychological attitudes*, Academic Press, San Diego, CA.

Punj, Girish N. and David W. Stewart (1983). "Cluster Analysis in Marketing Research: Review and Suggestions for Application." *Journal of Marketing Research*, 20, 134-48.

SAS Institute (2002). *The SAS System for Windows*, SAS Institute" Cary, North Carolina.

Statistical Innovations, (2003). *Latent Gold. Statistical Innovations:* Belmont, MA.

Steenkamp, Jan-Benedict E.M. and Frenkel Ter Hofstede (2002). "International Market Segmentation: Issues and Outlook." *International Journal of Research in Marketing*, 19, 185-213.

Steenkamp, Jan-Benedict E.M. and Hans Baumgartner (1998). "Assessing measurement invariance in cross-national consumer research." *Journal of Consumer Research* 25, 78-90.

Wedel, Michel and Wagner Kamakura. (1999). *Market Segmentation: Conceptual and Methodological Foundations.* Dordrecht: Kluwer Academic Publishers.

Wedel, Michel and Wayne S. DeSarbo. (2002) "Market segment derivation and profiling via a finite mixture model framework." *Marketing Letters*, 13 (1), 17-25.

COMMENT ON COHEN

JAY MAGIDSON,
STATISTICAL INNOVATIONS INC.

Overall, I found Steve's presentation to be clear, concise, and well organized. More important is the substance of his primary message, that segmentation can be improved upon, substantially in many cases, through the use of the new methods of maximum difference (max-diff) scaling and latent class (LC) modeling.

Steve presents strong arguments in favor of LC over traditional ad-hoc clustering techniques such as K-Means, by listing many weaknesses in K-means that are remedied by LC. The fact that only 58% of the cases classified by K-means into one of 6 segments remain in these segments (the rest being reassigned to one of the other 6 clusters) when simply repeating the analysis following a reverse ordering of the cases, is a striking illustration of the inherent inconsistency of K-means clustering.

To Steve's excellent list of problems with the traditional approach to clustering, I would add that use of Euclidean distance to measure closeness between cases works *only* if within each segment, all variables have equal variances. This is an unrealistic limitation which has been shown to result in high rates of misclassification even when all variables are standardized to Z-scores prior to the analysis (Magidson and Vermunt, 2002a, 2002b). LC models on the other hand, do not make such assumptions. Moreover, LC segmentations are invariant to linear transformations made to one or more variables.

The maximum difference approach to scaling, introduced originally by Jordan Louviere, is an important contribution that may well alter the way that conjoint data is collected in the future. In Steve's example, he shows that 5 of the 6 pair-wise comparisons are captured by just 2 selections – best and worst choices. In my own research I am finding that the choice of best and worst in discrete choice studies provides extremely powerful information. When used with LC, max-diff outperforms by a significant margin the equally parsimonious design involving the first and second choice.

To obtain the final segments, Steve pointed out that he first estimated individual coefficients using HB, and then subjected these coefficients to a LC analysis. A better approach would be to perform a simultaneous (1-step) analysis using LC rather than using this *tandem* approach. While individual coefficients would then be unnecessary to obtain the segments, should such be desired for other reasons, they could still be attained – directly from LC.

The trick to obtaining individual coefficients for each case with LC is to weight the segment-level coefficients by the posterior membership probabilities obtained for each case. This approach makes use of the fact that HB may be viewed as a parametric and LC as a nonparametric method for random effects modeling (Vermunt and Magidson, 2003). In the HB case the random effects are assumed to be continuous, while for LC they are assumed to be discrete. In my current research with maximum difference scaling, I am finding that use of individual coefficients obtained from LC produce higher hit rates than those produced by HB.

Overall, Steve's paper makes an important contribution to the field. It is for good reasons that his presentation was voted best at the conference.

REFERENCES:

Magidson, J. and J.K. Vermunt (2002a) "Latent class models for clustering: A comparison with K-means" in *Canadian Journal of Marketing Research*, Vol. 20, pp. 37-44.

Magidson, J. and J.K. Vermunt (2002b) "Latent class modeling as a probabilistic extension of K-means clustering", in *Quirk's Marketing Research Review*, March issue.

Vermunt J.K. and Magidson, J. (2003, forthcoming) "Random Effects Modeling", in *Sage Encyclopedia of Social Science Research Methods*, Sage.

THE PREDICTIVE VALIDITY OF KRUSKAL'S RELATIVE IMPORTANCE ALGORITHM

KEITH CHRZAN AND JOSEPH RETZER
MARITZ RESEARCH
JON BUSBICE
IMS AMERICA

INTRODUCTION

An intuitively meaningful, computationally intensive technique for assigning relative importance to attributes in a regression framework is described by Kruskal (1987). This approach involves averaging some measure of importance over all possible attribute orderings. "Averaging over orderings" (AOO) in turn provides comparatively stable estimates, which appear less sensitive to problems with data ill-conditioning. This paper begins with an illustration of common importance analysis techniques along with a critique of their effectiveness. Next, an explanation and illustration of Kruskal's technique employing "averaging over orderings" is presented. We then examine an application of the AOO approach to prediction and compare these estimates to those obtained via standard multiple OLS regression analysis using a jackknife sampling approach. The increased stability of the parameter estimates when using "averaging over orderings" appears to offer marginal increases in predictive performance.

COMMON METHODS FOR MEASURING ATTRIBUTE IMPORTANCE

Two techniques often used to estimate attribute importance include bivariate correlations and multiple regression. Each technique attempts to measure some portion of variation in a dependent variable explained by an attribute (this common variation is also referred to as "shared information"). Usually when two or more variables are considered for their explanatory power, a certain amount of overlap occurs in that some common portion of the variation in the dependent variable is explained by more than one attribute. This information overlap is illustrated in Chart 1 where the circle labeled "Overall Satisfaction" represents the total variation in the dependent variable. The "Quality" and "Image" circles reflect corresponding quantities for two independent attributes. The regions labeled A, B and C represent portions of shared information between the attributes and dependent variable, "Overall Satisfaction" (the sum, A + B + C, is the total amount of shared information between the attributes and dependent variable). The regions can be described in terms of explained variation as:

- The variation in "Overall Satisfaction" uniquely explained by "Quality".
- The variation in "Overall Satisfaction" explained by both "Quality and "Image".
- The variation in "Overall Satisfaction" uniquely explained by "Image".

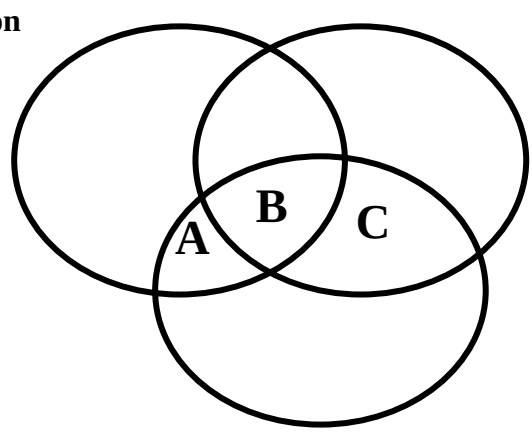


Chart 1

When assigning attribute importance, a problem arises in attempting to partition the overlapping information (region B) between the two attributes. Correlation analysis implicitly credits both attributes with the overlap. Specifically, when using correlations the importance of the "Quality" attribute is reflected in the sum of regions A and B. "Image's" importance is shown as the sum of regions B and C. This approach therefore effectively double counts region B.

The third measure, standardized multiple regression coefficients, also fail to partition region B. This approach is at the other extreme in that it credits neither of the attributes with the overlap. Using regression coefficients, the importance of the "Quality" attribute is associated with region A while "Image's" importance is given by region C. This effectively ignores region B. Neither of these techniques results in a desirable measure of importance.

ASSIGNING IMPORTANCE BY AVERAGING OVER ORDERINGS (AOO)

An alternative approach to measuring importance involves examining the ***additional*** contribution an attribute makes as it is added to a set of one or more existing attributes.

In order to estimate this additional contribution, we first need to remove the variability explained by the existing attributes. This implies an "ordering" in consideration of the attributes. For example, measuring the additional variation in "Overall Satisfaction" explained by "Image" ***after*** accounting for "Quality" implies our ordering is:

Quality Image

Since "Quality" is considered first, any variation in "Overall Satisfaction" that it explains (uniquely or otherwise) should be assigned to "Quality" (i.e. regions A + B in Chart 1). Next, since "Image" is considered ***after*** "Quality", any explained variation the two have in common should not be credited to "Image". In other words "Image" should be credited only with explained variation that is unique to it (i.e. region C). This is the ***additional*** variation explained by "Image".

The above ordering therefore credits "Quality" with the entire overlap, region B. While this eliminates the problem of double counting (correlations) and also doesn't ignore the overlap altogether (standardized regression coefficients), without strong prior information suggesting that "Quality" should be considered first, we may feel uncomfortable assigning region B entirely to one attribute. To avoid doing so, we instead consider all possible orderings with equal weight[1] and then average the importance assigned to the attributes in each ordering (see Kruskal (1987)). This technique offers numerous advantages from both intuitive and practical points of view. The main advantage from an intuitive point of view is that it allows for the overlapping information (region B), to be divided among the attributes. In the case of our two attribute example there are only two orderings to consider:

Quality Image

and,

Image Quality.

Assigning importance to each attribute and averaging, results in the following importance estimates:

Quality: A + (B/2)

Image: C + (B/2)

As we add additional attributes (i.e. "Satisfaction with (Value)") the analysis becomes more involved. Consider the ordering:

Quality Image Value

The relative importance of "Quality" is calculated with no accounting for the other variables. The contribution of "Image" is calculated ***after*** accounting for (removing) the information with respect to "Overall Satisfaction", common to both it and "Quality". Finally, the relative importance of "Value" is measured after removing any information with regard to the dependent variable "Overall Satisfaction", common to it and either of the other two attributes. The potential orderings for this set of attributes are:

(1) Quality Image Value (2) Quality Value Image (3) Image Quality Value

(4) Value Quality Image (5) Image Value Quality (6) Value Image Quality

The relative importance of any given attribute is calculated for each of the above orderings[2]. Next, an overall measure of relative importance for that attribute is constructed by averaging over the six terms[3]. These steps would be repeated for each attribute. The advantage of averaging over orderings is that it provides a more accurate picture of importance by taking into account all possible scenarios.

---

[1] Non-equal weighting may also be used in order to incorporate prior information.

[2] Kruskal suggests using squared correlations and squared partial correlations as underlying measures of importance. This approach however, yields measures of shared information that do not sum to the total amount of shared information. Theil and Chung correct this problem by applying Kruskal's "averaging over orderings" analysis to information theoretic measures (see Appendix II). These measures represent the relative amount of "information" contained in the attribute set with respect to the dependent variable. See Theil (1987 and 1988) for a complete description of "Averaging over Orderings" in an information theoretic context.

[3] Note that some of the measures of relative importance, for a particular variable, will be the same in different orderings. For example, in the case of "Quality" it's relative importance is the same in orderings (1) and (2). The measures will likewise be equal for "Quality" in orderings (5) and (6).

In addition, through numerous applications of the "averaging over ordering" technique (on both observed and simulated data), evidence indicates that its importance estimates are much less affected by highly collinear attributes. Regression coefficients, on the other hand, may be adversely impacted by such attribute correlation.

AN ILLUSTRATION

As an example of relative importance analysis, customer satisfaction data for an information call center is considered. Specifically, overall customer satisfaction with call center representative (Overall Satisfaction) is considered to be a function of the following measures attributed to the representative:

Sincerity

Courtesy

Insightful Questions (To Determine Needs)

Positive / Helpful (Attitude)

(Shows) Compassion (for Situation)

Provides (Needed) Info

Re-Direct (To Sources of Further Info)

Correlations (between the attributes and the dependent variable) range from .63 to .69 (see Table 1). This suggests a substantial overlap of information within the attribute set, regarding the dependent variable "Overall Satisfaction". The information overlap dilutes the ability of the correlations to distinguish among attributes on the basis of importance. This overlap also suggests that a good deal of information would be ignored if we were to use standardized regression coefficients (see Table 1) in place of correlations, as measures of attribute importance.

Using the "Averaging over Orderings" approach we can estimate the average partial squared correlation of each attribute with "Overall Satisfaction", and arrange them in order of importance. In addition we could use averaging over orderings to estimate information, measured in bits, for each attribute.[4]

Table 1:
Importance / Information Measures

Provide Info	Re-Direct	Insightful Questions	Sincerity	Positive Helpful	Courtesy	Compassion	
23 %	17 %	16 %	13 %	12 %	11 %	9 %	Pct. Info. / Importance
0.31	0.23	0.21	0.17	0.16	0.15	0.12	= 1.35 Info. (bits)
0.69	0.68	0.68	0.67	0.66	0.64	0.63	Correlation
0.29	0.15	0.15	0.13	0.06	0.14	-0.03	Reg. Coeff.

The total amount of information contained in the attribute set is the sum of the individual values (calculated to be 1.35 (bits)). With this in mind we can draw numerous insights based on the results. First, if we wish to know the percentage of total information, regarding "Overall

[4] See Theil and Chung 1988.

Satisfaction", accounted for by a particular attribute, we could estimate information ratios using the formula: (attribute info)/(total info) x 100. For example:

1. Average Percent accounted for by "Sincerity", .17/1.35 x 100 = 12.59 %.
2. Average Percent accounted for by "Provides Info", .31/1.35 x 100 = 22.96%. Etc.

Relative comparisons can also be made i.e., "Provides Info" is 2.58 (= .31/.12) times as important in explaining "Overall Satisfaction" as is "Compassion".

Since the information measures are additive (unlike regression and correlation coefficients), additional insight is gained by examining groups of attributes as categorized by the researcher.

AOO Prediction Using Jack-Knife Sampling

In the next section we review the steps taken for constructing "out of sample" prediction performance measures using OLS and AOO respectively. Six independent empirical data sets with varying degrees of collinearity were used to predict some dependent measure of interest. Specifically, for each data set:

- A sequential pass is made through the data set, leaving out observations one at a time.
- This results in 2 subsets of data:
  - The first contains n-1 observations
  - The second contains one observation to be predicted.
- Using the n-1 observations in the first data set, we build two predictive models using:
  - OLS and
  - Kruskal's averaged over orderings (AOO) applied to regression beta coefficient estimates.
- Using the model estimates from the previous step, we predict the single, holdout observation using first OLS and then AOO beta estimates.
- Once each observation has been predicted, correlation measures between predictions and actual values of the dependent measure for, first, OLS and next, AOO models are estimated.
- The test of difference of "correlated" correlations is performed and reported. (See Appendix I).
- An overall "global" test of significance is also performed. (See Appendix I).

Results

Results from each of the six data sets employed are summarized in Table 2 below. Specifically, the table reports, for each study,

N = study sample size,

k = number of independent attributes,

C.I. = condition index of independent data set,

R^2 = OLS coefficient of multiple determination,

r_{OLS} = correlation between OLS predictions and actual dependent variable level,

r_{AOO} = correlation between AOO predictions and actual dependent variable level,

t = Hotelling's "correlated correlations" t-score.

Table 2:
Results from Individual Studies 1 to 6

Study	N	k	C.I.	R^2	r_{OLS}	r_{AOO}	t
1.	195	10	31.52	36.84	.539	.548	0.68
2.	64	8	36.14	56.31	.586	.623	1.37
3.	157	16	55.02	42.58	.514	.534	0.65
4.	1001	7	61.54	59.87	.745	.752	2.67*
5.	56	10	167.57	68.44	.708	.710	0.96
6.	68	9	80.65	81.00	.730	.800	3.41*

The results suggest that AOO estimates provide some improvement in prediction (albeit a marginal one) over standard OLS estimates. It is important to note that intuition would suggest superior performance from the AOO estimates when the data is ill-conditioned. Ill-conditioning is related but not equivalent to conditioning (Belsley 1991). For this reason we may not expect to see the greatest differences in correlations when high collinearity is present. Data from study 6 did in fact exhibit the greatest amount of instability despite having the second highest condition index. As is shown in the table, study 6 benefited most from the AOO estimates.

In addition to the individual tests, a global test of significance was performed. The chi-square statistic for the test is given as:

$$\sum_{i=1}^{6} -2\ln(p_i) = 31.67 \sim \chi^2_{12}$$

This test is significant at a .01 level implying that, overall, the AOO estimates provide superior predictions for the 6 data sets under consideration.

CONCLUSION

Various techniques commonly used to measure attribute importance are flawed (see Kruskal and Majors (1989)). Kruskal's "averaging over orderings" approach offers an extendable framework for appropriately measuring attribute relative importance. This approach provides numerous insights into what is driving the variable of interest. The technique offers a way to partition shared information (as opposed to double count or ignore) among multiple attributes

regarding some variable of interest. Evidence also shows it is less sensitive to problems associated with highly correlated data.

The extension of Kruskal's technique to an information theoretic measure, as suggested by Theil and Chung (1988), adds intuitive meaning to the analysis. Information theoretic measurement of attribute importance can also be applied to situations where our independent variable set is categorical (ANOVA), as well as when the dependent variable itself is categorical (Logit). Soofi, Retzer and Yasai (2000) illustrate these applications in detail.

The benefits of Kruskal's method are available with no decrement in predictive validity.

In fact, Kruskal's method appears to outperform standard OLS, even for prediction.

REFERENCES

Belsley, D. A. (1991). Conditioning diagnostics: collinearity and weak data in regression. *John Wiley & Sons, Inc*. New York, NY.

Cohen and Cohen (1983), Applied multiple regression / correlation analyses for the behavioral sciences. *Second Edition. Lawrence Erlbaum*, Hinsdale, NJ.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41, 6-10.

Kruskal, W., & Majors, R. (1989). Concepts of relative importance in scientific literature. *The American Statistician*, 43, 2-6.

Soofi, E. S., Retzer, J. J. & Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences Journal*, 31, Number 3, 596-625.

Theil, H. (1987). How many bits of information does an independent variable yield in a multiple regression? *Statistics and Probability Letter*s, 6, 107-108.

Theil, H., & Chung, C. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions**.** *The American Statistici*an, *4*2, 249-252.

APPENDIX I
HYPOTHESIS TESTS

Hotelling's within population "Correlated Correlations" t-test.[5] is used to compare the correlation between "OLS predictions with actual" vs. "AOO predictions with actual". Hotelling's test is a standard test of correlated correlations (correlations that share a variable) and can be expressed as:

$$t_{N-3} = [r_{12} - r_{13}] \cdot \frac{\sqrt{[N-3]\cdot[1+r_{23}]}}{\sqrt{2[1-r_{23}^2 - r_{12}^2 - r_{13}^2 + [2\cdot r_{23}\cdot r_{12}\cdot r_{13}]]}}$$

Where:

r_{ij} = correlation between variables i and j and

N = sample size.

Also, a global test involving the sum of the p-values from the individual t-tests may be performed. Specifically,

$$\sum_{i=1}^{K} -2\ln(p_i) \sim \chi^2_{2K}$$

Where:

p_i = p-value from the i^{th} test.

[5] See Cohen and Cohen 1983.

APPENDIX II
AOO USING INFORMATION THEORETIC MEASURES

A useful extension to Kruskal's approach involves an AOO of information as opposed to partial squared correlations. [6]

Specifically, Theil represents the total information in the attribute set pertaining to the dependent variable as $I(R^2)$ where

$$I(R^2) = \log_2(1 - R^2)$$

In addition $(1\text{-}R^2)$ may be decomposed as,

$$1 - R^2 = (1 - r^2_{y,x_1})(1 - r^2_{y,x_2 \cdot x_1}) \cdots (1 - r^2_{y,x_p \cdot (x_1, x_2, x_3 \ldots x_{(p-1)})}).$$

Substituting in the RHS of the decomposition equation into Information function, I(·) gives,

$$I(R^2) = I(r^2_{y,x_1}) + I(r^2_{y,x_2 \cdot x_1}) + \cdots + I(r^2_{y,x_p \cdot (x_1, x_2, x_3 \ldots x_{(p-1)})}).$$

The above gives a unique additive decomposition of the information in each attribute for the ordering,

$$x_1, x_2, x_3, \ldots, x_p.$$

Measuring individual attribute information for all possible orderings and then averaging these measures, provides the average amount of information an attribute contributes to the explanation of the dependent variable. This measure has a number of advantages over partial squared correlations, not the least of which is additively, i.e. the sum of the measures is intuitively meaningful (the sum is the total information, which in turn is a nonlinear transformation of R^2).

[6] See Theil 1987, and Theil & Chung 1988.

SEGMENTATION

NEW DEVELOPMENTS IN LATENT CLASS CHOICE MODELS

JAY MAGIDSON
STATISTICAL INNOVATIONS INC.
TOM EAGLE
EAGLE ANALYTICS
JEROEN K. VERMUNT
TILBURG UNIVERSITY

Discrete choice models have proven to be good methods for predicting market shares for new products based on consumers' expressed preferences between choice alternatives. However, the standard aggregate model fails to take into account the fact that preferences (utilities) differ from one respondent to another (or at least from one segment to another). This failure often yields poor share predictions. The most popular remedy for this problem has been to use a mixture model. In this paper, we provide insights into this problem and illustrate the solution posed by latent class (LC) finite mixture models. We also describe several recent advances in the development of LC models for choice which have been implemented in a new computer program (Vermunt and Magidson, 2003a).

We conclude with a comparison of the LC finite mixture approach with the Hierarchical Bayes (HB) *continuous* mixture approach to choice modeling in a case study involving boots. We find that while both models provide comparable predictions, the LC models take much less time to estimate. In addition, the discrete nature of the LC model makes it more useful for identifying market segments and providing within-segment share predictions.

INTRODUCTION

The basic aggregate model as introduced by McFadden (1974) postulates that a choice of one alternative A_j is made from a set of alternatives $A = \{A_1, A_2,\ldots,A_J\}$, according to a random utility model

$U_j = V_j + e_j$

where Vj represents the systematic component of the utility and e denotes a stochastic error. The alternative Aj selected, is the one with highest utility U_j

In the simplest situation, the systematic utility component is assumed to satisfy a linear function of the choice attributes $X_1, X_2, \ldots, X_K$

$$V_j = \beta_{0j} + \beta_1 X_{j1} + \beta_2 X_{j2} + \ldots + \beta_K X_{jK}$$

$\beta_k X_{jk}$ is called the partworth utility associated with attribute k.

β_{0j} is called an alternative-specific constant, and may be omitted from the model.

Let Z denote the union of all the sets of alternatives. Then, under the assumption that e follows an 'Extreme Value Type 1 (Gumbel)' distribution, it follows that for any subset of alternatives $A' \subseteq Z$, the probability of choosing $A_j \in A'$ is given by the multinomial equation,

$$P_j = \exp(V_j) / \sum_{k \in A'} \exp(V_k) \quad ,$$

providing a probabilistic justification for this conditional logit model.

The implication of this equation is that if any alternative is excluded from the choice set A, its choice probability is allocated among the remaining alternatives proportional to their original choice probabilities. That is, it is assumed that none of the remaining alternatives is more likely (than any other) to serve as a substitute for the omitted alternative. Generally, this proportional-substitution-of-alternatives assumption, also known as IIA (Independence of Irrelevant Alternatives), does not hold in practice. McFadden (1974) recognized this as a weakness in his proposed model:

"This points out a weakness in the model that one cannot postulate a pattern of differential substitutability and complementarity between alternatives. … The primary limitation of the model is that the IIA axiom is implausible for alternative sets containing choices that are close substitutes."

THE LATENT CLASS SOLUTION TO THE IIA PROBLEM

LC Modeling assumes that IIA holds true within each of T ≥ 1 latent classes or segments:

$$P_{j.t} = \exp(V_{j.t}) / \sum_{k \in A'} \exp(V_{k.t}) \quad t = 1,2,\ldots,T$$

To illustrate the problem that occurs when the IIA assumption is violated and how the LC specification resolves this problem, consider the classic Red Bus/Blue Bus problem where the following 3 alternative means of transportation are available:

A_1 = Car, A_2 = Red Bus, A_3 = Blue Bus

For simplicity, we assume

$\exp(V_1) = .50$ and $\exp(V_2) = \exp(V_3) = .25$

In a choice between only A'_1 = Car and A'_2 = Red Bus, the aggregate model allocates the .25 Blue Bus probability between the remaining choices to preserve the 2:1 A_1: A_2 ratio of probabilities. This yields:

$P(\text{Car}) \equiv P(A'_1) = .50/(.50+.25) = .67$ which is clearly unreasonable.

In reality, the Red Bus would serve as a substitute for the Blue Bus yielding, $P(A'_1) = .5$.

With real data, LC modeling would reject the aggregate (1-class) model because it would not yield predicted choices between the Car and Red Bus (in a 2-alternative choice set) that are consistent with the *observed* choices. That is, the LC statistical criteria (discussed later) would reject the aggregate model in favor of T > 1 segments.

For simplicity, we will suppose that in reality there are 2 equal sized latent classes: those who prefer to take the bus (t=1) and those who prefer to drive (t =2).

Table 1

	exp(Vj)		
	Alternative j		
	1	2	3
Segment t	CAR	Red Bus	Blue Bus
1	0.02	0.49	0.49
2	0.98	0.01	0.01
Overall	0.50	0.25	0.25

In the case of the blue bus no longer being available, proportional allocation of its share over the 2 remaining alternatives separately within class yields:

P(Car.1) = .02/(.02 + .49) = .04, P(Car.2) =.98/(.98 + .01) = .99, and overall, P(Car) = .5(.04) +.5(.99) = .52

The Red Bus/Blue Bus problem illustrates the extreme case where there is *perfect* substitution between 2 alternatives. In practice, one alternative will not likely be a perfect substitute for another, but will be a more likely substitute than some others. Accounting for heterogeneity of preferences associated with different market segments will improve share predictions.

LATENT CLASS CHOICE MODELING

Thus far we have shown that LC choice models provide a vehicle for accounting for the fact that different segments of the population have different needs and values and thus may exhibit different choice preferences. Since it is not known apriori which respondents belong to which segments, by treating the underlying segments as hidden or *latent* classes, LC modeling provides a solution to the problem of unobserved heterogeneity. Simultaneously, LC choice modeling a) determines the number of (latent) segments and the size of each segment, and b) estimates a separate set of utility parameters for each segment. In addition to *overall* market share projections associated with various scenarios, output from LC modeling also provides separate share predictions for each latent segment in choices involving any subset of alternatives.

Recent advances in LC methodology have resolved earlier difficulties (see Sawtooth Software, 2000) in the use of LC models associated with speed of estimation, algorithmic convergence, and the prevalence of local solutions. It should be noted that despite those early difficulties, the paper still concluded with a recommendation for its use:

> "Although we think it is important to describe the difficulties presented by LCLASS, we think it is the best way currently available to find market segments with CBC-generated choice data"

ADVANCES IN LC MODELING

Several recent advances in LC choice modeling have occurred which have been implemented in a computer program called Latent GOLD Choice (Vermunt and Magidson, 2003a). These advances include the following:

- Under the general framework of LC regression modeling, a unified maximum likelihood methodology has been developed that applies to a wide variety of dependent variable scale types. These include choice, ranking (full, partial, best/worst), rating, yes/no (special case of rating or choice), constant sum (special case of choice with replication weights), and joint choices (special case of choice).
- Inclusion of covariates to better understand segments in terms of demographics and other external variables, and to help classify new cases into the appropriate segment.
- Improved estimation algorithm substantially increases speed of estimation. A hybrid algorithm switches from an enhanced EM to the Newton Raphson algorithm when close to convergence.
- Bootstrap p-value – Overcomes data sparseness problem. Can be used to confirm that the aggregate model does not fit the data and if the power of the design is sufficient, that the number of segments in the final model is adequate.
- Automated smart random start set generation – Convenient way to reduce the likelihood of local solutions.
- Imposition of zero, equality, and monotonicity restrictions on parameters to improve the efficiency of the parameter estimates.
- Use of Bayes constants – Eliminates boundary solutions and speeds convergence.
- Rescaled parameters and new graphical displays to more easily interpret results and segment differences.
- New generalized R-squared statistic for use with any multinomial logit LC regression model.
- Availability of individual HB-like coefficients.

Each of these areas is discussed in detail in Vermunt and Magidson (2003a). In the next section we will illustrate these advances using a simple brand pricing example involving 3 latent classes.

LC BRAND PRICING EXAMPLE:

This example consists of six 3-alternative choice sets where each set poses a choice among alternative #1: a new brand – Brand A (at a certain price), alternative #2: the current brand – Brand B (at a certain price) and alternative #3: a None option. In total, Z consists of 7 different alternatives.

Table 2

Alternative	Brand	Price
A1	A	Low
A2	A	Medium
A3	A	High
B1	B	Low
B2	B	Medium
B3	B	High
None	None	

The six sets are numbered 1,2,3,7,8 and 9 as follows:

Table 3

	PRICE BRAND B		
PRICE BRAND A	**Low**	**Medium**	**High**
Low	1	4	7
Medium	2	5	8
High	3	6	9

Shaded cells refer to *inactive* sets for which share estimates will also be obtained (along with the six *active* sets) following model estimation.

Response data were generated[1] to reflect 3 market segments of equal size (500 cases for each segment) that differ on brand loyalty, price sensitivity and income. One segment has higher income and tends to be loyal to the existing brand B, a second segment has lower income and is not loyal but chooses solely on the basis of price, and a 3rd segment is somewhere between these two.

Table 4

	UpperMid	Loyal to Brand B	Price Sensitives
INCOME			
lower	0.05	0.05	0.90
lower middle	0.05	0.05	0.90
upper middle	0.88	0.10	0.02
higher	0.15	0.75	0.10

[1] The data set was constructed by John Wurst of SDR.

LC choice models specifying between 1 and 4 classes were estimated with INCOME being used as an active covariate. Three attributes were included in the models – 1) BRAND (A vs. B), 2) PRICE (treated as a nominal variable), and 3) NONE (a dummy variable where 1=None option selected). The effect of PRICE was restricted to be monotonic decreasing. The results of these models are given below.

Table 5

With Income as an active covariate:

LC Segments	LL	BIC(LL)	Npar	$R^2(0)$	R^2	Hit Rate	bootstrap p-value*	std. error
1-Class Choice	-7956.0	15940.7	4	0.115	0.040	51.6%	0.00	0.0000
2-Class Choice	-7252.5	14584.4	11	0.279	0.218	63.3%	0.00	0.0000
3-Class Choice	**-7154.2**	**14445.3**	**19**	**0.287**	**0.227**	63.5%	**0.39**	**0.0488**
4-Class Choice	-7145.1	14484.8	27	0.298	0.239	64.1%	0.37	0.0483

* based on 100 samples

The 3-class solution emerges correctly as best according to the BIC statistic (lowest value). Notice that the hit rate increases from 51.6% to 63.5% as the number of classes is increased from 1 to 3 and the corresponding increase in the $R^2(0)$ statistic[2] is from .115 to .287. The bootstrap p-value shows that the aggregate model as well as the 2-class model fails to provide an adequate fit to the data.

Using the Latent GOLD Choice program to estimate these models under the technical defaults (including 10 sets of random starting values for the parameter estimates), the program converged rapidly for all 4 models. The time to estimate these models is given below:

Table 6

	Time* (# seconds) to:	
LC Segments	Fit model	Bootstrap p-value
1	3	14
2	5	18
3	7	67
4	11	115

* Models fit using a Pentium III computer running at 650Mhz

[2] The R^2 statistic represents the percentage of choice variation (computed relative to the baseline model containing only the alternative-specific constants) that is explained by the model. In this application, the effect of the alternative-specific constants is confounded with the brand and None effects, and thus we measure predictive performance instead relative to the *null* model which assigns equal choice probabilities to each of the 3 alternatives within a set. This latter R^2 statistic is denoted by $R^2(0)$.

The parameters of the model include the size and part-worth utilities in each class. For the 3-class model they are given below.

Table 7

	Upper Mid	Loyal to B	Price Sensitives	
Size	0.35	0.33	0.32	
$R^2(0)$	0.054	0.544	0.206	Overall 0.287

Attributes	Upper Mid	Loyal to B	Price Sensitives	p-value	Mean	Std.Dev.
BRAND						
A	-0.29	-1.15	0.03	2.1E-84	-0.47	0.50
B	0.29	1.15	-0.03		0.47	0.50
PRICE						
low	0.42	0.01	1.25	1.0E-53	0.55	0.51
medium	0.02	0.01	0.05		0.03	0.02
high	-0.44	-0.02	-1.30		-0.57	0.53
NOBUY	0.02	-1.04	0.62	1.4E-44	-0.14	0.68

The p-value tests the null hypothesis that the corresponding part-worth utility estimate is zero in each segment. This hypothesis is rejected ($p<.05$) showing that the effects are all significant. Notice that several within-segment utility estimates are close to zero[3]. In particular, the PRICE effect for the Loyal segment is zero[4] except for sampling variability. For this segment, the *unrestricted* PRICE effects turned out to be .00, .02, -.02 for the low, medium and high price levels respectively. The difference between the .00 and .02 reflect sampling error and are smoothed by the monotonicity restriction.

Viewing the part-worth utilities as random effects (see e.g. Vermunt and van Dijk, 2001, Vermunt and Magidson, 2002, 2003b,), we see that the brand A effect of -.29 (for the Upper Mid segment) occurs with overall probability .35, -1.15 with overall probability .33 and .03 with overall probability .32. Hence, the HB-like mean and Std Dev. Parameters can be computed as above. Similarly, individual HB-like part-worth parameters can be computed for each respondent, using that individual's posterior probability of being in each segment as weights in place of the overall probabilities.

The parameters in the model for predicting class membership as a function of INCOME are estimated simultaneously with the parameters given above. These include an intercept for the classes plus the direct relationship between INCOME and class membership. These estimates are expressed using effects coding in the following table. Identifying those estimates that are large in absolute value we see that class 1 (relative to the other classes) is more likely to have upper middle income, class 2 higher income, and class 3 lower and lower middle income. The

[3] Standard errors for these parameter estimates confirm that they do not differ significantly from zero.

[4] The PRICE effect for the Loyal to B segment could be further restricted to zero along with the BRAND effect for the Price Sensitive segment but for expediency this is not done in this paper.

low p-value shows that the relationship between INCOME and class membership is highly significant.

Table 8

Model for Classes	Upper Mid	Loyal to B	Price Sensitives	
Intercept	-0.04	-0.28	0.31	
Covariates	Class1	Class2	Class3	p-value
INCOME				
lower	-0.56	-0.94	1.50	3.1E-88
lower middle	-0.97	-0.49	1.46	
upper middle	1.79	-0.13	-1.66	
higher	-0.26	1.56	-1.30	

Let us now return to the utility parameters. These log-linear parameters can be re-expressed in a form that allows easier interpretation, and that can also be used to develop an informative graphical display. To transform the parameters to column percentages, we can use a simple formula. For the BRAND effect parameter associated with brand A in segment t (betaA.t), we transform it into a column percentage form, ProbA.t, as follows:

ProbA.t = exp(betaA.t)/[exp(betaA.t)+exp(betaB.t)]

Thus, we obtain ProbA.t + ProbB.t = 1 for t=1,2,3

Table 9

BRAND	Upper Mid	Loyal to B	Price Sens.
A	0.3605	0.0907	0.5155
B	0.6395	0.9093	0.4845

The interpretation of these numbers is as follows: In a choice between Brand A and Brand B where the other attribute (PRICE) is set at a common level, these re-scaled parameters represent the probability of choosing each Brand given latent class t. This interpretation holds regardless of the price level. Thus, for example the Price Sensitives are indifferent in their choice between brands A and B (.50 vs. .50) except for the nonsignificant difference in utilities (.03 vs -.03).

For attributes like price, these quantities indicate price sensitivities

Table 10

PRICE.A			
low	0.4769	0.3360	0.7237
medium	0.3198	0.3360	0.2192
high	0.2033	0.3281	0.0570

Using the overall latent class probabilities, these numbers can then be transformed to row %s to yield an insightful display.

Table 11

PROBMEANS Output — Row %s

	Upper Mid	Loyal to B	Price Sensitives
Overall Probability	0.35	0.33	0.32
Attributes			
BRAND			
A	0.39	0.09	0.52
B	0.33	0.44	0.23
PRICE			
low	0.33	0.22	0.46
medium	0.38	0.38	0.24
high	0.36	0.55	0.09
NOBUY			
0	0.33	0.46	0.21
1	0.37	0.18	0.44
Covariates			
INCOME			
lower	0.08	0.04	0.88
lower middle	0.05	0.07	0.88
upper middle	0.86	0.10	0.04
higher	0.16	0.76	0.08

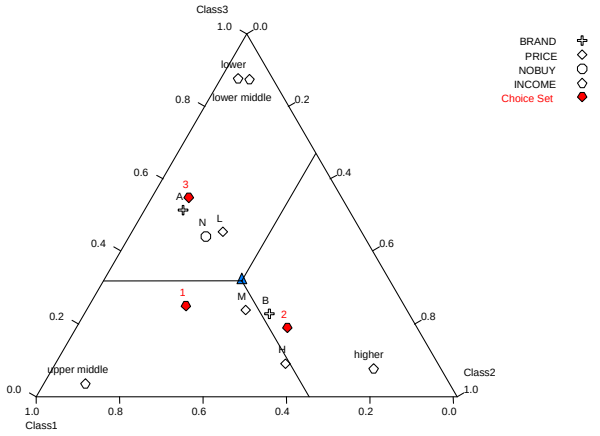
The same type of transformation from column to row percentages can be applied to the predicted choice probabilities.

Table 12

Column Percentages				Row Percentages			
Set 1 n = 1350	Upper Mid	Loyal to B	Price Sensitives	Set 1	Upper Mid	Loyal to B	Price Sensitives
Choice 1	0.27	0.08	0.41	Choice 1	0.38	0.11	0.52
2	0.48	0.83	0.38	2	0.30	0.48	0.22
3	0.24	0.09	0.21	3	0.46	0.16	0.37
Set 2 n = 1350				Set 2			
Choice 1	0.20	0.08	0.17	Choice 1	0.46	0.18	0.36
2	0.53	0.83	0.53	2	0.29	0.43	0.27
3	0.27	0.09	0.29	3	0.43	0.14	0.43
Set 3 n = 1350				Set 3			
Choice 1	0.14	0.08	0.05	Choice 1	0.53	0.29	0.18
2	0.57	0.83	0.61	2	0.30	0.41	0.29
3	0.29	0.09	0.34	3	0.42	0.13	0.45
Set 4 n = 0				Set 4			
Choice 1	0.32	0.08	0.55	Choice 1	0.36	0.09	0.56
2	0.39	0.83	0.16	2	0.29	0.59	0.11
3	0.29	0.09	0.29	3	0.45	0.13	0.41
Set 5 n = 0				Set 5			
Choice 1	0.24	0.08	0.27	Choice 1	0.42	0.14	0.44
2	0.43	0.83	0.26	2	0.30	0.54	0.16
3	0.33	0.09	0.47	3	0.39	0.10	0.51
Set 6 n = 0				Set 6			
Choice 1	0.17	0.08	0.09	Choice 1	0.52	0.23	0.25
2	0.47	0.83	0.32	2	0.31	0.50	0.19
3	0.36	0.09	0.59	3	0.36	0.09	0.55
Set 7 n = 1350				Set 7			
Choice 1	0.38	0.08	0.63	Choice 1	0.36	0.08	0.56
2	0.29	0.82	0.05	2	0.26	0.70	0.04
3	0.34	0.09	0.33	3	0.47	0.12	0.41
Set 8 n = 1350				Set 8			
Choice 1	0.29	0.08	0.34	Choice 1	0.42	0.12	0.46
2	0.33	0.82	0.08	2	0.28	0.66	0.06
3	0.39	0.09	0.58	3	0.38	0.09	0.53
Set 9 n = 1350				Set 9			
Choice 1	0.21	0.08	0.12	Choice 1	0.53	0.20	0.28
2	0.36	0.82	0.11	2	0.29	0.63	0.08
3	0.43	0.09	0.77	3	0.35	0.07	0.58

Now, these row percentages can be used to position the corresponding attribute, covariate level, and choice on a common scale in a Barycentric coordinate display. For example, the following display plots the choices associated with set #6 (1:brand A Higher price vs. 2:brand B Medium price vs 3:None) together with INCOME and attribute levels in a common plot.

Figure 1:
Barycentric Coordinate Display of 3-Segment Solution



In this plot, each segment corresponds to a vertex of the triangle. From this plot it can be seen that the lower right vertex corresponds to the Loyal to B segment. It is associated with Higher Income, Choice 2 (Brand B Higher Price) in set #3, brand B and Higher prices in general indicating the lack of price sensitivity. In contrast, the top vertex corresponds to the Price Sensitives. It is associated with Low and Lower Middle Income, a relatively higher preference for Brand A and choice 3 (None) when faced with the medium and higher priced set #3 options, Lower prices and the None option, more so than the other segments. Similarly, the lower left vertex corresponds to the UpperMid segment. It is associated with Upper Middle income and has a relatively higher likelihood of making choice #1 (Brand A Medium Price) in set #3.

COMPARISON BETWEEN LC AND HB

LC models utilize a discrete distribution of heterogeneity as opposed to a continuous distribution as assumed by HB. HB models assume that each respondent has their own unique preferences, while LC models assume that each respondent belongs to one of K (latent) segments, each of which has its own unique preferences. However, since the LC model also yields estimates for each respondent's probability of belonging to each segment, usage of these posterior probabilities as weights result in unique HB-like individual coefficients. Thus, even in the case that each individual has his/her own unique preference, an LC model containing a large number of classes can be used instead of HB to account for the heterogeneity. This approach avoids the necessity of making distributional assumptions (required by HB) which may cause poor predictions for cases with few responses (see e.g., Andrews et. al, 2002). This weighting is justified by viewing LC modeling as a non-parametric alternative to traditional HB-like random effects modeling (Vermunt and van Dijk, 2001).

To examine how LC and HB compare in practice, we used both in a CBC boot study (see Appendix).

SUMMARY

When some alternatives in a set are more similar to each other than to the others, differential substitution is likely to occur which violates the IIA assumption in the aggregate model. In this case, use of the aggregate model is inappropriate and the resulting share predictions are distorted. The recommended remedy in this situation is to use a model that accounts for respondent heterogeneity of preferences. When comparing LC with HB models, LC modeling has the advantage of finding segments more directly than HB, and also is much faster to estimate. Moreover, a general maximum likelihood framework exists for LC models that includes the ability to test various kinds of choice models, in restricted and unrestricted forms.

In our case study, prediction on hold-out sets (within-case/internal validation) of HB was found to be somewhat better than LC although there was a substantial fall-off in prediction error in the validation data, which is indicative of over-fitting. Our results appear to be consistent with Andrews et. al. (2002) who concluded:

> "... models with continuous and discrete representations of heterogeneity … predict holdout choices about equally well except when the number of purchases per household is small, in which case the models with continuous representations perform very poorly"

Inclusion of covariates is a key issue to improve prediction on hold-out cases since the LC/HB models themselves don't improve prediction over the aggregate model in our case study. Although the application to hold-out cases is not emphasized in choice experiments we believe that it is an important topic. In our case study, the covariates were poor predictors of the LC segments so this aspect of the comparison was not addressed.

Our overall conclusion is that LC modeling has emerged as an important and valuable way to model respondent preferences in ratings-based conjoint and CBC, simulate choice shares and find segments simultaneously. Several advances have been incorporated into a commercially available latent class tool called Latent GOLD Choice which substantially enhances both the speed and reliability of estimation. LC methods are now practical with traditional CBC data as well as with ranking (full, partial, best/worst), rating, yes/no (special case of rating or choice), and constant sum models, within choice sets. Compared to HB, LC is much faster, and directly provides insights regarding segments.

REFERENCES

McFadden, Daniel (1974) “Conditional logit analysis of qualitative choice behavior” I. Zerembka (ed.), Frontiers in econometrics, 105-142. New York: Academic Press.

Sawtooth Software (2000) “CBC Latent Class Analysis Technical Paper”. Sawtooth Software Technical paper Series, 2000.

Vermunt, Jeroen K. and Liesbet A. van Dijk. “A non-parametric random-coefficient approach: the latent class regression model”, in Multilevel Modelling Newsletter, 2001, 13, 6-13.

Vermunt, J.K. and Magidson (2003a) “Latent GOLD Choice 3.0 Users Guide" Belmont, MA: Statistical Innovations.

Vermunt, J.K. and Magidson “Non-parametric Random Effects Model”, in Encyclopedia of Social Science Methodology, Sage, forthcoming 2003b.

Vermunt, J.K. and Magidson J., (2002), “Latent Class Models for Classification”, Computational Statistics and Data Analysis.

Natter, Martin and Markus Feurstein 2002, “Real world performance of choice-based conjoint models”, European Journal of Operational Research, 137, 448-458.

Andrews, Rick L., Andrew Ainslie, and Imram S. Currim, 2002, “An empirical comparison of logit choice models with discrete vs. continuous representations of heterogeneity”, Journal of Marketing Research, Vol. XXXIX (November), 479-487.

APPENDIX

CASE STUDY: A COMPARISON OF MODELS

Preliminaries

As part of our discussion, we have put together a simple comparison of Latent Class choice models to other forms of choice models. The comparison is not to do a definitive test of which model is best, but rather to demonstrate the power of modeling heterogeneity among respondents in explaining choice behaviors. We can understand choice behaviors much better using very simple models by capturing respondent heterogeneity. Improvement in computer and estimation technologies makes this process easier than ever before.

We compare three forms of models: the simple, aggregate, multinomial logit model, the latent class form of the same MNL model, and the hierarchical Bayes form of the same MNL model. The aggregate MNL model is developed in its most naïve form for comparative purposes only. By naïve form, we mean that the utility function is comprised of only generic attribute characteristics. No individual characteristics, nor complex model forms (such as the nested MNL, mixed logit, or GEV form are examined). The latent class models we develop use exactly the same utility function, but, of course, latent classes are uncovered that have clear parameter differences among the classes. Finally the HB MNL model fits parameters at the individual level; segments of one, so to speak.

The case is disguised. The business problem was that of a well-known hiking boot material manufacturer wishing to determine whether they could enter the well-established hiking boot market with their own branded hiking boot. One of their concerns was whether their entry into the market would simply cannibalize their existing share of the materials market by stealing market share away from the existing brands that use their material in their own boots. If the boot material manufacturer, Brand X, enters the market, from whom would they steal market share? Indeed, if they enter the market, into which channels should they concentrate sales.

Channel and brand differentiate this market. Certain brands concentrate their sales in lower priced markets using lower priced channels. Generally, these channels offer lower priced boots catering to the more casual day hiker and outdoors enthusiast. Other brands concentrate their sales on the serious hiker, rock climber, and outdoors person. Higher priced boots are found in specialized sporting goods, outdoors stores, or even specialty shops. Brand X must determine which channel(s) they would enter if they decide to launch their own boot. We narrow our focus of analysis of respondent heterogeneity on the patterns of choices made among brand and channel.

In addition to these questions, Brand X was also interested in the pricing of the boots, and whether they should exclusively introduce any new material feature.

DESIGN AND SAMPLE INFORMATION

A stated choice modeling exercise was commissioned to answer these questions. The design attributes and their levels are described in detail in Table 1. The key attributes are: brand, channel, two different special features, and price. Channel and price posed a price conditional

design problem because price and channel are correlated with each other in the market. Specific channels could not have the highest range of prices tested shown with them. Discount stores, for example, do not offer the highest priced, almost custom made boots. Specialty stores, conversely, seldom offer the lowest priced boots we wished to test. An addition design consideration included the fact that each boot brand also produces boots using their own materials as well as offering boots using Brand X's materials, so the design had to incorporate these levels that are confounded directly with the brand into the design. Finally, one brand offered their boots at only their own specialty store.

Table A1. Design Attributes

Levels	Brand	Store	Performance Feature 1	Performance Feature 2	Price*
1	Merrell	Discount Store	Merrell Level	Standard	$50
2	Vasque	Catalog	Vasque Level	Special Upgrades	$75
3	Asolo	Sporting Goods	Asolo Level		$100
4	Salomon	Department Store	Salomon Level		$125
5	Brand Alpha	Specialty Outdoors Store	Alpha Level		$175
6			Alpha New Level		$225

The final design consisted of 72 choice sets of 6 branded alternatives plus a "None of these" alternative. Brands and channels were rotated across the columns of the design in a random fashion. The design was blocked into 4 versions of 18 choice tasks each.

The final sample consisted of 573 respondents who participated in an Internet survey. This sample was broken up into three parts for analysis purposes. First, we took the total data set and randomly selected 20% (112 respondents) and put them into a holdout sample data set. The holdout sample was not used in the model estimation. The remaining 80% (461 respondents) of the sample were used for model estimation after each respondent had two choice sets randomly withdrawn to use for holdout sets. The estimation data set finally consisted of 461 respondents who responded to 16 choice tasks.

The holdout sets data set consisted of 922 choices made by respondents in the estimation data set. We compare the predictive accuracy of the models we fit against these holdout sets as a measure of internal validity. While this is commonly done in market research, there are significant problems with using the "fit" figures to select the best model (Elrod, 2000). We show these figures strictly for comparison purposes. The holdout sample respondents were used to compare the "external validity" or transferability of the estimated models parameters. More on this comparison is discussed later.

MODEL RESULTS

MNL Model

The naïve multinomial logit model was fit with 13 parameters: a simple, generic, linear effect for price (price/100); brand specific effects coding for brand and channel specific effect coding for channel; effects coding for the special features 1 and 2; and a dummy constant for the None alternative (1=None; 0=otherwise). In all, there were 13 parameters. The parameters are listed in the column label simple MNL in Table 2. The fit statistics can be also be found in Table 2. The model took less than one minute to estimate on a 1.4 GHz Pentium 4 PC using Windows 2000 and 756 mb of high speed memory.

The major results are that price is the most important attribute. The relative impact on the utility is far above that of any other attribute. Looking at brand, we see that Brand X has a brand value almost as strong as Merrell, the strongest brand. Channel is less important than brand. The sporting goods and department stores are the strongest channels, followed by Outdoors and discount stores, with the Internet trailing last. The additional feature 2 and additional feature 1 have the second strongest impact on utility. The most immediate conclusion about brand and channel is that Merrell offered in a sporting goods store would generate the largest market share holding all else constant. Brand X would be a close second in the same channel. Notice with this coding of the MNL model we did not isolate the interactions between brand and channel, so the impact of each on utility is independent of one another.

Table A2
Models and Parameters

	Naïve MNL	8 Class Latent Class Model								HB MNL Model		
Variable	**1 Class**	**Class1**	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Class8**	**Mean***	**Lower 95%**	**Upper 95%**
Discount	0.0128	0.062	-0.084	-0.271	0.188	-0.737 *	-0.338 *	0.119	-0.271	-0.078	-1.312	0.953
	(0.045)	(0.115)	(0.307)	(0.155)	(0.172)	(0.176)	(0.164)	(0.212)	(0.285)	(0.671)		
Internet	-0.136 *	-0.135	0.264	0.041	-0.251	-0.411 *	-0.035	-0.078	0.231	-0.0709	-0.06072	0.427
	(0.036)	(0.111)	(0.280)	(0.104)	(0.147)	(0.120)	(0.111)	(0.195)	(0.207)	(0.310)		
Department	0.0436	-0.157	-0.350	0.054	-0.090	0.442 *	0.305 *	-0.005	0.351	0.076	-0.5986	0.7841
	(0.033)	(0.106)	(0.287)	(0.091)	(0.118)	(0.098)	(0.111)	(0.139)	(0.184)	(0.426)		
Sporting Goods	0.0657	0.333 *	-0.656	-0.039	-0.098	0.184	0.212	-0.015	-0.126	0.04478	-0.8387	0.9312
	(0.046)	(0.163)	(0.676)	(0.141)	(0.163)	(0.121)	(0.130)	(0.213)	(0.228)	(0.525)		
Outdoors**	0.014	-0.102	0.826	0.215	0.250	0.522	-0.144	-0.021	-0.186	0.028	-0.6941	0.8224
										(0.456)		
Merrell	0.2791 *	0.396 *	0.422	-0.195	0.721 *	-0.340 *	0.474 *	0.063	2.600 *	0.3256	-1.0447	1.85295
	(0.033)	(0.111)	(0.227)	(0.112)	(0.119)	(0.107)	(0.117)	(0.151)	(0.184)	(0.901)		
Vasque	-0.3137 *	-0.264 *	-0.883 *	-0.429 *	-0.416 *	-0.095	-0.282 *	-0.228	-0.966 *	-0.306	-1.0721	0.4637
	(0.036)	(0.117)	(0.319)	(0.115)	(0.123)	(0.095)	(0.118)	(0.144)	(0.249)	(0.486)		
Asolo	0.0611	0.103	-0.119	0.385 *	-0.049	0.329 *	0.162	-0.121	-0.114	0.2999	-0.593019	1.2507
	(0.034)	(0.109)	(0.235)	(0.113)	(0.147)	(0.097)	(0.121)	(0.182)	(0.192)	(0.591)		
NorthFace	-0.2864 *	-0.420 *	0.128	-0.364 *	-0.715 *	0.746 *	-1.337 *	-0.123	-0.897 *	-0.650	-2.1656	1.1964
	(0.045)	(0.212)	(0.497)	(0.156)	(0.222)	(0.095)	(0.189)	(0.182)	(0.281)	(1.075)		
Brand X**	0.260	0.184	0.452	0.603	0.458	-0.640	0.983	0.409	-0.623	0.33047	-0.7753	1.4024
										(0.679)		
Perf Feat 1 a	0.0884 *	0.127	-0.504 *	1.179 *	0.516 *	-0.089	0.326 *	0.211	-0.108	0.336	-0.34969	1.019
	(0.034)	(0.093)	(0.246)	(0.236)	(0.122)	(0.099)	(0.154)	(0.213)	(0.176)	(0.427)		
Perf Feat 1 a	0.3812 *	-0.072	0.781 *	2.384 *	0.445 *	0.361 *	0.814 *	0.736 *	0.273	0.762899	-0.337455	1.9377
	(0.036)	(0.125)	(0.367)	(0.260)	(0.122)	(0.092)	(0.134)	(0.227)	(0.154)	(0.721)		
Perf Feat 1 a**	-0.470	-0.055	-0.277	-3.563	-0.960	-0.272	-1.140	-0.948	-0.165	-1.099	-2.651	0.2638
										(0.880)		
Perf Feat 2	0.3388 *	0.305 *	0.832 *	0.758 *	0.271 *	0.190 *	0.202 *	1.551 *	0.311 *	0.62855	-0.1762	1.5738
	(0.018)	(0.058)	(0.155)	(0.063)	(0.066)	(0.051)	(0.080)	(0.165)	(0.096)	(0.505)		
Price	-2.4128 *	-5.670 *	-7.371 *	-3.152 *	-6.597 *	-1.039 *	-1.384 *	-3.631 *	-1.537 *	-5.401	-9.7698	-0.8767
	(0.061)	(0.284)	(0.778)	(0.210)	(0.441)	(0.139)	(0.208)	(0.372)	(0.290)	(2.766)		
None	-1.3075 *	-3.714 *	-2.260 *	-0.860 *	-7.887 *	-0.923 *	-3.222 *	-5.262 *	-1.642 *	-3.5625	-9.0885	1.9471
	(0.071)	(0.261)	(0.644)	(0.293)	(0.631)	(0.211)	(0.460)	(0.475)	(0.403)	(0.329)		
Class size		0.236	0.217	0.139	0.113	0.104	0.081	0.072	0.039			
Log Likelihood	-9,990.50		-7,219.30							-4056.100		
-2(LL0-LL)	2,966.40 13 df		8,508.80	216 df								
BIC	20,060.70		15,763.40									
R-Square	0.09		0.41							0.65		
Rho-Square	0.13		0.37							0.650		
MAE	1.35		0.90							0.64		
MSE	0.68		0.44							0.290		
CLE	0.54		0.34							0.2		
Est. Time	< 10 secs	6 min. 6 seconds on a 1.4 GHz Pentium 4 PC								2 hours 10 min. on a 1.4 GHz Pentium 4 P		

LATENT CLASS MODELS

We tested several different latent class models. One particular model included the use of covariates to assist us in predicting latent class membership. This model used the same exact 13 parameters as the MNL model to predict utilities within each latent class. We tested the demographics of Income, gender, age and education as covariates. We purposely used these simple demographics to keep the problem small and because they are measures clearly available for non-respondents.

In LC models, covariates are used to predict class membership. They directly affect the overall utility of an alternative by affecting the class membership probability. As such, however, they do not directly influence any single parameter.

The Latent Gold Choice program was used to test a range of segment solutions, 2 to 20, models on the estimation data set using the 13 parameters in the utility function and the 4 covariates. The program was set to run 100 random start values, each with 100 iterations. The EM algorithm maximum was set to 1000 and a maximum of 50 Newton-Raphson iterations were allowed for final model fit. The best fitting model according to the BIC statistic had 8 classes

(BIC=15,763). The parameters and fit statistics can be found in Table 2. The program took 6 minutes and 6 seconds running on the same machine as the MNL model.

The model is a significant improvement over the naïve MNL model. The McFadden Rho-square jumped from 0.13 to 0.37. All of the fit statistics show a remarkable improvement over the MNL model. The parameters are generally larger than those found in the MNL model. This is likely due to scale differences between the class-level results derived in the LC model as compared to the MNL. Tests for scale differences were not conducted.

The most interesting results begin to emerge when we examine the patterns of respondent heterogeneity across the latent classes. Table 3 depicts the proportion of the sample who prefer specific combinations of brand and channel. This table was constructed by examining the latent class parameters for brand and channel. We determined the specific combination of brand and channel parameter with the highest utility (brand parameter + channel parameter) within each latent class. The specific combination represents one cell in the matrix. We then placed the class size (in %) into the cell and added together all cells with more than one segment. We split class 2 equally between Brands X and Merrell because the Brand X and Merrell brand parameters are so close.

Table A3
Brand x Channel Class Preferences

	Merrell	Vasque	Asolo	NorthFace	Brand X*	Channel Preference
Discount store					7.2%	7.2%
Over the internet						
Department store	3.9%				8.1%	12.0%
Sporting goods store	23.6%					23.6%
Outdoors store*	22.3%			10.4%	24.5%	57.2%
Brand Preference	49.8%			10.4%	39.8%	100%

In the case of the naïve MNL model, this matrix would consist of a single cell. In the latent class model, seven cells are occupied indicating much more diversity in the Brand X channel preferences. While the Merrell brand is still the dominate brand, Brand X now captures more of the market than what would have been predicted with the simple MNL. We should note these figures do not represent actual market shares, rather they represent that portion of the market with specific brand by channel preferences without estimating the actual market shares. It is an examination of the parameters across respondents. The actual substantive conclusions would need to be calculated using the market simulator

There is also a good deal of price heterogeneity across the classes as well. Classes 1, 2 and 4 all have price parameters 1 unit less than the weighted mean of the price parameters. These are predominately Merrell and Brand X classes. Classes 3 and 7 are relatively close to the average parameter, while classes 5, 6 and 8 are clearly much less price sensitive than any of the other classes. These classes include the Northface/Outdoors store combination and the Merrell/Department store combination.

The analysis of the covariates reveals that the covariates selected do not assist in a meaningful manner our ability to predict latent class membership. Table 4 shows the covariate parameters and their associated standard errors. Very few are significantly different from zero, even less than the number we would expect at chance. Ass such, these selected covariates will not assist us in making better predictions to the holdout sample data set. Of course, in the actual study, many more covariates were examined, and more were significant predictors of class membership, but still fewer than we would have liked. This suggests the posterior probability of class membership is not a function of the respondents' characteristics.

Table A4
8 Class Latent Class Covariate Parameters

Covariates	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Intercept	1.7672 *	1.7308 *	-0.6307	0.7199	-0.7182	-0.0715	-0.6165	-2.181
	(0.667)	(0.666)	(2.171)	(0.707)	(2.225)	(1.604)	(1.634)	(2.655)
Income								
$100k +	-0.628 *	-0.633 *	0.376	0.025	0.098	-0.288	0.118	0.932
	(0.320)	(0.315)	(0.325)	(0.350)	(0.334)	(0.500)	(0.613)	(1.261)
$25k to < $35k	0.126	0.249	-0.971	0.540	0.128	0.612	-1.564 *	0.882
	(0.315)	(0.311)	(0.515)	(0.384)	(0.408)	(0.401)	(0.731)	(1.282)
$35k to < $50k	-0.173	-0.234	0.129	-0.010	-0.716	-0.053	-0.143	1.200
	(0.279)	(0.275)	(0.311)	(0.330)	(0.387)	(0.366)	(0.440)	(1.232)
$50k to < $75k	0.067	-0.001	-0.124	-0.260	0.011	-0.820	0.585	0.542
	(0.263)	(0.267)	(0.319)	(0.332)	(0.328)	(0.421)	(0.412)	(1.255)
$75k to < $100k	0.608	0.619	0.590	-0.294	0.480	0.548	1.005	-3.556
	(0.714)	(0.713)	(0.738)	(0.776)	(0.741)	(0.763)	(0.785)	(4.744)
Age								
18	2.359	0.876	2.171	1.002	2.334	-3.034	-3.729	-1.978
	(1.849)	(1.880)	(1.891)	(2.012)	(1.908)	(7.475)	(7.518)	(7.479)
25	-0.814	-0.319	-0.248	-0.708	0.151	-0.154	1.137	0.955
	(0.431)	(0.429)	(0.461)	(0.525)	(0.457)	(1.538)	(1.546)	(1.539)
35	-0.366	-0.266	-0.134	0.372	-0.966	0.125	1.182	0.054
	(0.430)	(0.435)	(0.451)	(0.491)	(0.521)	(1.536)	(1.541)	(1.573)
45	-0.336	-0.319	-0.211	0.727	-0.762	0.886	-0.166	0.181
	(0.451)	(0.459)	(0.489)	(0.501)	(0.542)	(1.535)	(1.610)	(1.598)
55	-0.176	-0.157	-0.921	-0.075	-0.365	0.929	0.560	0.205
	(0.435)	(0.438)	(0.510)	(0.545)	(0.491)	(1.544)	(1.636)	(1.582)
65	-0.666	0.185	-0.657	-1.317	-0.391	1.248	1.016	0.582
	(0.533)	(0.482)	(0.626)	(0.871)	(0.602)	(1.551)	(1.631)	(1.606)
Gender								
Female	0.131	0.079	-0.012	-0.090	-0.065	-0.104	0.212	-0.151
	(0.114)	(0.113)	(0.141)	(0.156)	(0.156)	(0.183)	(0.185)	(0.231)
Male	-0.131	-0.079	0.012	0.090	0.065	0.104	-0.212	0.151
	(0.114)	(0.113)	(0.141)	(0.156)	(0.156)	(0.183)	(0.185)	(0.231)
Education								
College graduate	-0.507	-0.759	1.488	-0.243	1.237	-0.082	-0.746	-0.387
	(0.566)	(0.553)	(2.143)	(0.581)	(2.200)	(0.602)	(0.628)	(1.908)
Graduate school	-0.112	0.101	1.340	-0.293	1.998	-1.392	-1.887 *	0.245
	(0.594)	(0.572)	(2.157)	(0.645)	(2.207)	(0.840)	(0.958)	(1.921)
High school or less	0.261	0.411	2.008	-0.354	-3.471	0.791	-0.483	0.839
	(1.039)	(1.026)	(2.339)	(1.142)	(6.197)	(1.071)	(1.264)	(2.171)
Some college	-0.327	-0.642	1.535	-0.656	0.836	-0.661	-0.481	0.397
	(0.556)	(0.546)	(2.140)	(0.585)	(2.202)	(0.609)	(0.622)	(1.881)
Some graduate school	0.602	-0.209	-3.611	-0.288	2.216	-0.269	0.220	1.340
	(0.962)	(0.974)	(5.669)	(1.095)	(2.339)	(1.076)	(1.050)	(2.053)
Technical school	0.084	1.099	-2.760	1.835	-2.815	1.613	3.376	-2.432
	(2.361)	(2.289)	(9.226)	(2.301)	(9.232)	(2.326)	(2.359)	(9.162)

Because the covariates did not seem to improve our membership predictability, we went back to the model and fit it without any covariates. Again, we tested the range of 2 – 20 classes, but also the 35 classes and 50 classes model. The best fitting model according to the BIC statistic was the 14 class model. The parameters and fit statistics for this model are presented in Table 5.

Table A5
14 Class Latent Class Model

Variable	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class10	Class11	Class12	Class13	Class14
Discount	-0.038	-0.019	0.671	0.331	-0.145	0.140	0.283	-0.085	-1.142 *	-0.932 *	-0.030	0.053	-6.531	-1.181
	(0.332)	(0.148)	(2.385)	(0.180)	(0.183)	(0.261)	(0.350)	(0.263)	(0.275)	(0.438)	(0.232)	(0.319)	(10.907)	(0.646)
Internet	0.264	0.008	0.301	-0.476 *	-0.146	-0.073	0.394 *	0.005	-0.478 *	-0.116	0.205	0.167	0.879	-0.265
	(0.268)	(0.133)	(2.390)	(0.162)	(0.136)	(0.203)	(0.189)	(0.214)	(0.166)	(0.262)	(0.170)	(0.235)	(2.738)	(0.454)
Department	-0.401	-0.197	0.202	-0.152	0.093	0.028	-0.075	-0.084	0.911 *	0.483 *	0.187	0.275	1.350	0.757
	(0.300)	(0.135)	(2.384)	(0.154)	(0.116)	(0.141)	(0.194)	(0.198)	(0.155)	(0.183)	(0.151)	(0.209)	(2.734)	(0.458)
Sporting Goods	-0.509	0.321	-3.130	0.119	0.420 *	-0.076	-1.022 *	-0.185	0.158	0.207	-0.045	-0.221	2.169	-0.169
	(0.611)	(0.186)	(9.406)	(0.182)	(0.133)	(0.250)	(0.341)	(0.243)	(0.174)	(0.285)	(0.174)	(0.261)	(2.735)	(0.486)
Outdoors**	0.684	-0.113	1.956	0.179	-0.222	-0.019	0.419	0.349	0.551	0.357	-0.317	-0.274	2.134	0.858
Merrell	0.586 *	0.504 *	-0.167	0.922 *	0.187	0.080	-0.327	0.450 *	0.072	-0.671 *	0.703 *	2.870 *	-0.895	0.296
	(0.240)	(0.149)	(3.158)	(0.142)	(0.136)	(0.159)	(0.259)	(0.179)	(0.148)	(0.277)	(0.143)	(0.226)	(0.593)	(2.133)
Vasque	-1.183 *	-0.211	0.683	-0.548 *	0.027	-0.267	-0.916 *	-0.453 *	0.154	-0.549 *	-0.647 *	-1.103 *	-0.433	-5.592
	(0.384)	(0.156)	(3.216)	(0.149)	(0.120)	(0.163)	(0.256)	(0.192)	(0.135)	(0.210)	(0.193)	(0.300)	(0.577)	(8.355)
Asolo	-0.118	0.327 *	0.676	-0.341 *	0.182	-0.043	0.698 *	0.025	0.008	0.000	0.331 *	-0.149	-0.771	3.916
	(0.262)	(0.151)	(3.199)	(0.161)	(0.120)	(0.148)	(0.189)	(0.194)	(0.133)	(0.210)	(0.147)	(0.235)	(0.582)	(2.118)
NorthFace	0.246	-0.793 *	-1.860	-0.487 *	-0.857 *	-0.161	0.401	-0.867 *	-0.171	0.551 *	-1.558 *	-1.141 *	3.127	1.862
	(0.479)	(0.301)	(12.721)	(0.207)	(0.186)	(0.221)	(0.322)	(0.286)	(0.190)	(0.254)	(0.286)	(0.343)	(2.188)	(2.123)
Brand X**	0.468	0.173	0.668	0.454	0.461	0.391	0.144	0.845	-0.063	0.669	1.172	-0.477	-1.028	-0.483
Perf Feat 1 a	-0.617 *	0.092	0.541	0.523 *	0.248	0.300	0.932 *	0.510 *	-0.377 *	1.161 *	0.935 *	0.038	0.045	0.542
	(0.278)	(0.120)	(0.419)	(0.134)	(0.134)	(0.270)	(0.305)	(0.222)	(0.149)	(0.553)	(0.241)	(0.201)	(0.222)	(0.492)
Perf Feat 1 b	0.936 *	-0.148	-1.624	0.471 *	0.472 *	0.742 *	1.654 *	0.431 *	0.536 *	4.379 *	1.116 *	0.225	0.676 *	1.279 *
	(0.398)	(0.152)	(1.006)	(0.146)	(0.129)	(0.214)	(0.360)	(0.186)	(0.125)	(0.674)	(0.242)	(0.171)	(0.212)	(0.507)
Perf Feat 1 c**	-0.319	0.056	1.083	-0.995	-0.720	-1.042	-2.586	-0.941	-0.159	-5.540	-2.051	-0.263	-0.721	-1.820
Perf Feat 2	0.826 *	0.273 *	0.961 *	0.232 *	0.155 *	1.410 *	1.200 *	0.170	0.345 *	1.206 *	0.208 *	0.227 *	0.151	0.290
	(0.175)	(0.078)	(0.235)	(0.077)	(0.066)	(0.142)	(0.131)	(0.103)	(0.100)	(0.131)	(0.080)	(0.110)	(0.097)	(0.176)
Price	-6.806 *	-4.739 *	-13.820 *	-5.919 *	-1.280 *	-3.371 *	-4.644 *	-7.270 *	-1.737 *	-3.477 *	-0.809 *	-1.405 *	-1.221 *	-1.668 *
	(0.913)	(0.376)	(1.686)	(0.452)	(0.192)	(0.315)	(0.583)	(0.798)	(0.246)	(0.394)	(0.226)	(0.340)	(0.236)	(0.525)
None	-1.653 *	-2.898 *	-7.241	-5.832 *	-0.356	-5.342 *	-2.783 *	-14.177 *	-4.079 *	0.212	-3.879 *	-1.330 *	0.088	0.988
	(0.747)	(0.383)	(3.837)	(0.527)	(0.249)	(0.684)	(0.681)	(5.573)	(0.529)	(0.672)	(1.092)	(0.445)	(2.211)	(2.224)
Class size	0.1936	0.1363	0.0956	0.0889	0.0766	0.072	0.0665	0.0543	0.0506	0.046	0.043	0.0333	0.0302	0.0131
Log Likelihood		-6,980.30												
-2(LL0-LL)		8,986.80	195 df											
BIC		15,156.60												
R-Square		0.46												
Rho-Square		0.39												
MAE		0.83												
MSE		0.41												
CLE		0.30												
Est. Time	8 min. 35 seconds on a 1.4 GHz Pentium 4 PC													

In addition, we repeated the brand by channel heterogeneity analysis as seen in Table 6. The pattern is similar to those we saw with the 8 class model, only now there are more cells occupied by classes than before. This demonstrates there is increased parameter heterogeneity when we extend the analysis to 14 classes without covariates.

Table A6
Brand x Channel Class Preferences

	Merrell	Vasque	Asolo	NorthFace	Brand X*	Channel Preference
Discount store	8.9%				7.2%	16.1%
Over the internet					4.3%	4.3%
Department store	3.3%	5.1%	0.7%		4.6%	13.7%
Sporting goods store	13.5%			1.5%	7.7%	22.7%
Outdoors store*	19.4%	3.2%	10.6%	1.5%	8.6%	43.3%
Brand Preference	45.1%	8.3%	11.3%	3.0%	32.4%	100%

We do not report the parameters of the 35 and 50 class models in this paper. They were run primarily to demonstrate the ability of the latent class algorithm to carry its analyses beyond the recommended fit statistics when, for whatever reason, one wishes to over fit the data. Some of the fit statistics are reported in the validation section that follows.

HIERARCHICAL BAYES MNL MODEL

Lastly, we fit a HB MNL model using the same 13 estimates as the MNL and LC models. We did include the same covariates as the 8 class LC model. In HB models, the covariates are used to predict the individual estimates of the model, not class membership as in the LC model. The results resemble 13 regression models with the regression of the individual-level estimate on the covariates. As such, the HB model covariates can have dramatically different impacts on the individual estimates.

The HB MNL model was fit using 10,000 burn-in iterations using non-informative prior estimate information. After the initial burn-in, we ran another 10,000 iterations saving every 10^{th} iteration. This recommended interval reduces the autoregressive nature of the saved mean estimates for each respondent. The model took 2 hours and 10 minutes to fit on the same PC used to fit the other models. The log-likelihood for this model is -4,056.1 (MNL=-9,990.5; 8 class LC=-7,219.3). Obviously, the HB model is an improvement over the MNL and LC models. Table 2 also shows the mean estimates across respondents and their standard deviations. Other fit statistics are also shown in the table.

The mean estimates of the HB model are similar to the weighted average of the LC model parameters, but again, larger than the MNL model. The same hypothesis of the difference being due to possible scale differences holds for the HB model.

Examining the HB estimates, we see that most have large standard deviations. In fact, for all estimates except price the 95% confidence interval ranges from positive to negative values. The price confidence interval is always negative. This suggests enough respondent heterogeneity across estimates to see sign reversals on some attributes, as we saw in the LC model results.

We now have individual-level brand and channel estimates so we can examine each respondent's estimates and assign them to a cell in the brand by channel table we have used previously. Table 7 shows the increased amount of respondent heterogeneity across these two sets of estimates. All but three cells have a non-zero proportion of the sample who prefer the specific combination of brand and channel. The Asolo brand preference stands out much more. Merrell and Brand X are still of similar brand strength spread across similar channels. This increased degree of captured respondent heterogeneity is borne out in an examination of the fit statistics and validation statistics in the next section.

Table A7
Brand x Channel Class Preferences

	Merrell	Vasque	Asolo	NorthFace	Brand X*	Channel Preference
Discount store	8.7%	0.4%	6.7%	0.4%	10.2%	26.5%
Over the internet	1.5%		4.8%		2.2%	8.5%
Department store	8.0%	0.9%	2.6%	3.9%	4.3%	19.7%
Sporting goods store	10.4%	0.2%	4.8%	1.3%	7.4%	24.1%
Outdoors store*	3.3%		7.6%	6.1%	4.3%	21.3%
Brand Preference	31.9%	1.5%	26.5%	11.7%	28.4%	100%

The covariates did not significantly assist us in estimating individual attribute estimates. While a few terms appeared to be significant, the overwhelming majority were not. As such, the demographic covariates we selected to demonstrate do not help us in fitting these models or estimates. As such, they do not help us in fitting our models to the holdout sample.

MODEL VALIDATION

While we discuss model validation, we note that this is a single study subject to a set of poorly fitting covariates with which to test model validation. We present these results simply to demonstrate the improvement in fit we can obtain by modeling respondent heterogeneity. We fully expect the HB model to outperform the LC model in this case, and both the LC and HB models to outperform the naïve MNL model. The choice of which method to use should be driven by the client and researcher needs at the end of the project. If the client desires segments, or classes, of respondents and can act upon those classes in a more tangible way than with individual level estimates, then we recommend using the LC models. If, however, the project needs require the individual level heterogeneity found in HB models, then one might prefer to use HB models. In a recent paper, Andrews and Currim (2002) suggested a study may result in poor HB model performance when the parameters are poorly identified. Other evidence suggests the HB models perform much better when you have more data per respondent than less. LC models are less subject to these issues than HB models.

Model validation is a tricky issue. Terry Elrod in his Sawtooth, 2001 paper indicated that simple measures of holdout set validation traditionally used are inappropriate measures for determining the best fitting model. He suggests using holdout sample cross-validation procedures for determining the best fitting model. Comparisons of model fit to the holdout sets/tasks of each respondent in the estimation data set are not as good as comparisons to a completely independent holdout sample. In our example, we did not perform a 4-fold cross validation holdout sample test as recommended by Elrod. We did conduct a single holdout sample test.

In addition, he recommends replacing the traditional measures of mean absolute and mean squared error (MAE and MSE) and the classification “error” rates (CLE) with the calculated log-likelihood for the models based upon the holdout sets and holdout samples. We report the mean log likelihood, multiplied by -1 to make the value positive. Table 8 displays the formulas we used to estimate these measures.

Table A8
Measures of model fit used

Mean absolute and mean squared error

$MAE = \Sigma_n ABS(prob_{actual} - prob_{pred}) / n$

$MSE = \Sigma_n (prob_{actual} - prob_{pred})2 / n$

Classification error

CLE = 1 - Hit Rate;

Where:

Hit Rate = proportion of observations correctly predicted by the model. Predicted choice is the alternative having the highest predicted probability of selection

Mean log-likelihood

-MLL = -1 * [LL(b) / n]

We have three data sets with which to compare fit measures: the estimation data set (N=461, 16 sets/respondent), the holdout sets/tasks (N=461, 2 sets/respondent randomly drawn from each respondent, and the holdout sample (N=113, 18 sets/respondent). Table 9 presents the results of our fit measures.

Table A9
Goodness of Fit Measures

	Estimation Data Set					Holdout Sets				Holdout Sample			
Model	**Time (in min.)**	**-MLL**	**MSE**	**MAE**	**CLE**	**-MLL**	**MSE**	**MAE**	**CLE**	**-MLL**	**MSE**	**MAE**	**CLE**
MNL w/o covariates	< 0.1	1.355	0.679	1.354	0.539	1.325	0.674	1.346	0.546	1.362	0.678	1.353	0.539
LC 8 class w/ covariates	6	0.857	0.440	0.895	0.335	0.959	0.483	0.932	0.370	1.415	0.709	1.363	0.568
LC 14 class w/o covariates	8.5	0.778	0.402	0.827	0.309	0.913	0.459	0.877	0.348	1.354	0.676	1.344	0.538
HB w/ covariates	130	0.550	0.291	0.641	0.204	0.861	0.435	0.788	0.318	1.744	0.796	1.253	0.579
LC 35 class w/o covariates	~17	0.665	0.350	0.710	0.256	-- Not examined --				-- Not examined --			
LC 50 class w/o covariates	~25	0.630	0.336	0.679	0.248	-- Not examined --				-- Not examined --			

The latent class models and the HB models clearly outperform the naïve MNL model in the estimation data set. The MAE and –MLL values drop quickly as we fit more parameters to this data. Estimation of the LC model is faster than the HB model, but the fit measures are clearly superior with the HB model. All three measures, the MAE, MSE and CLE, for the latent class and HB models show clear improvement over the MNL model. This is to be expected given we are fitting more parameters.

In addition to the 8 class and 14 class models, we also fit a 35 and 50 class LC model. For these models, we had to rely on the use of only the EM algorithm and the use of Bayesian constants in order for the models to converge. We only report their estimation data set fit

measures. There is discussion in the literature that HB models are over fitting the parameters to the data. As you fit more and more parameters, their ability to predict to other data sets may not be as good as their fit to the estimation data. We fit these extreme cases of the LC model to demonstrate that improved fit measures can be derived by over fitting, even with LC models. We do not attempt to test the models against the holdout tasks or holdout sample because we believe they are over fitted models.

In the holdout tasks data set, the HB model outperforms both LC models, but the level of its own performance has dropped considerably. The MSE and CLS measures dropped by over 30% and the MAE dropped by 18%. The fit measures for the LC models, however, are considerably more stable. The MSE and CLE measures dropped by approximately 12% each, while the MAE measure dropped by only 6%. This could be considered evidence of the degree of over fitting by both models.

The results of the holdout sample are very disappointing to us. Only the 14 class LC model did better than the naïve MNL model. Its improvement is so small it is virtually indistinguishable from the naïve MNL model's fit. This is a result of having poor covariate predictors in all the models. Because none of the covariates helps us predict class membership, or HB model parameters, we have no means to utilize the increased knowledge of respondent heterogeneity in making predictions to an external data set. In fact, by not having any covariate predictability, the best prediction we can make for the external data set is the aggregate, or average, predictions made by the naïve MNL model.

One implication of this poor attempt at predicting to an external data set is, that an LC or HB model fit without significant covariates which can explain the respondent heterogeneity within the estimation data set is restricted to making inferences to the sample from which we collected the data and the population it represents. Attempting to use that model to make inferences beyond the sample's domain, will be no better than the naive MNL model. Given our poor explanation of the respondent heterogeneity, we dare not attempt to draw any conclusions about the external validity of these models.

ARCHETYPAL ANALYSIS: AN ALTERNATIVE APPROACH TO FINDING AND DEFINING SEGMENTS

ANDREW ELDER
MOMENTUM RESEARCH GROUP
JON PINNELL
MARKETVISION RESEARCH

ABSTRACT

Marketing researchers commonly use cluster analysis to identify subsets, or segments, of respondents. While few would doubt the existence of individual-level differences in attitudes and sensitivities, agreement on the presence of neatly defined (and bounded) segments is less universal. An alternative approach to segmentation focuses on the extreme values in a distribution and describes people as a mixture of these extremes, or archetypes. From a marketing perspective, archetypes are interesting in that they reflect the aspirations of consumers and provide advertising or product design targets that can elicit a stronger response than those directed to a group centroid. This paper introduces the ideas of archetypes, details the mathematical calculations underlying the identification of archetypes, compares the findings from archetypal analysis to traditional cluster analysis, and provides strengths and weaknesses of the approach.

INTRODUCTION

It is human nature to confront uncertainty by attempting to quantify and encapsulate that which is unknown. When marketing professionals are faced with the uncertainty of an ill-defined target audience, we find inherent comfort in parceling the sprawling masses into identifiable sub-groups that provide a manageable structure for directing product development and delivery. With its intuitive assumptions and utilitarian applications, market segmentation has thus become a stalwart of the modern research practice.

Many researchers are quick to classify segmentation as the middle ground between a mass market and relationship (or "one-to-one") marketing. But while the Model T Ford and Amazon.com represent radically divergent approaches to product and service configuration, segmentation is more than simply a transitional perspective between the mass and the individual. In addition to describing the needs or messages for a particular customer type, it is a technique that enhances market understanding and guides strategic decisions at any level of customer specification.

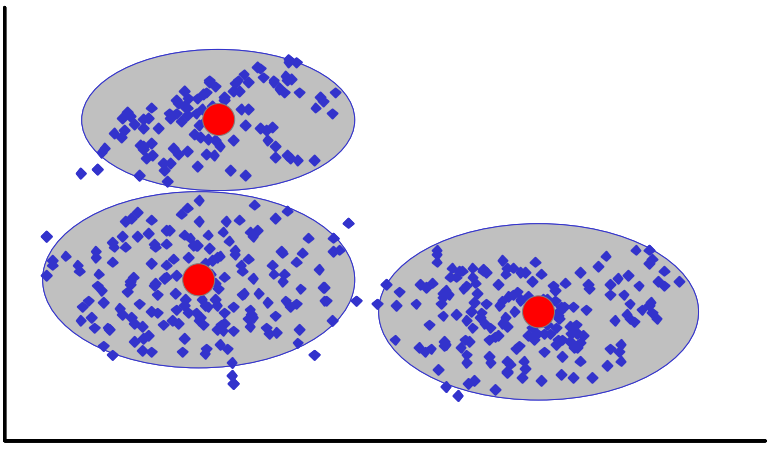
With this larger perspective, segmentation analysis clearly overlaps either extreme of the market-customization spectrum. There may only have been a single version of the Model T, but Henry Ford implemented the $5 workday for his assembly workers at least in part to develop demand among an evolving blue collar market segment. And while Ford's modern automotive market is exponentially more complex, encompassing seven million vehicles sold annually, scores of models, eight brands, online vehicle ordering, and customized marketing interaction, it

is still driven by distinct subgroups with preferences for performance, utility, economy, and social validation.

Such segmentation may appear overgeneralized in this era of technology-enabled personalization, but it provides a depth of understanding that is generally lacking from customer relationship management (CRM) applications. Customer databases, although notoriously spotty and incomplete, are populated with dry details such as log-ins, click-throughs, opt-ins, purchases, and demographics. But while a customer might opt to receive information regarding a 4x4 pickup with 6-CD changer, Ford knows very little about how to market its brand to that individual, or others like him. An efficient CRM system might inundate the prospective customer with catalogs and financing offers without any understanding of the motivation and needs driving the purchase. Segmentation, particularly when based on attitudes and perceptions, is one of the key research techniques available to explain the "who" and "why" behind the "what."

The variety of segmentation decisions available to the analyst — whether to use dependence or interdependence techniques that incorporate attitudinal, classification, or utility data — yield numerous possible routes for reaching the most meaningful grouping of customers. One of the most common approaches, particularly when dealing with attitudinal information, is to perform a k-means cluster analysis, which identifies groups that are maximally homogeneous amongst themselves and maximally heterogeneous relative to each other.

Figure 1.
An Ideal (Assumed) K-Means Cluster Analysis



A k-means approach creates discrete and potentially exhaustive clusters, defined by centroids that represent the average for all members of a given group. This approach is appealing because it matches the assumptions of what we perceive segments should be. First, we envision our data as a series of lumpy clouds. Nestled within their distributions, we expect to find unequally-distributed multivariate clusters awaiting our discovery.

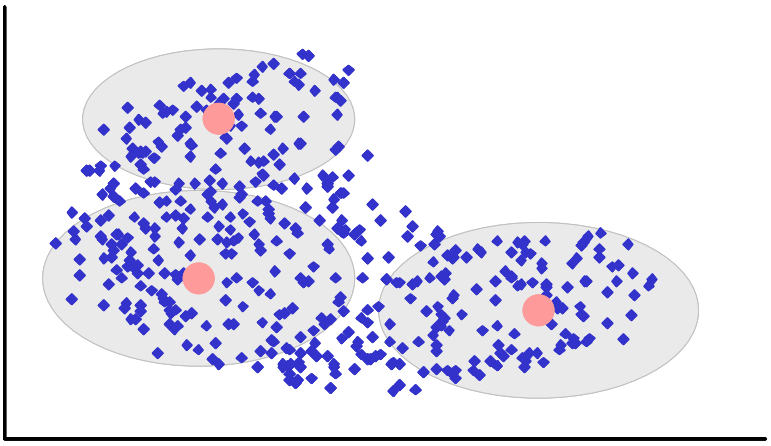
Second, we envision that there is a concentration of individuals within each of these masses. At the center of these concentrations lies an 'average' consumer who embodies a distinct combination of characteristics. This average consumer is assumed to represent the "sweet spot" in which the marketer can target the largest number of homogeneous respondents with a relevant product or message.

A third assumption is that a tidy boundary falls around each cluster, defining other individuals who are maximally similar to each other while also being maximally different from other groups. There may be a handful of individuals who fall outside these definitions, but they are assumed to be few and far between, and we can always force them into a group if we prefer to have comprehensive segment definitions.

These assumptions are seldom questioned, perhaps because they are so well understood, or perhaps because k-means segmentation has been applied successfully to a robust spectrum of scenarios. In either case, these assumptions beg the following questions:

- Is k-means segmentation the most appropriate approach when our data are not distributed according to our ideal concept of clusters?
- If we are performing segmentation to discover customer differentiation, why then do we care about the average (centroid) profile?

Figure 2.
A More Realistic Segmentation Scenario



In response to the first question, consider a scenario in which the segmentation data are not well behaved. It is hardly uncommon to analyze data that suggest a spectrum of opinions, needs, or utilities rather than the lumpy data described above. In such scenarios, there may well be segment-like groupings present, but their boundaries are obscured by a broad range of responses and opinions that fail to agglomerate into neatly defined buckets.

If we apply the k-means algorithm to a set of data, it will obtain segments and define their centroids, regardless of the data structure. It is up to the analyst to determine the robustness of the resulting segments by observing subtle indications such as the increased presence of outliers, reduced variable differentiation across segments, the lack of a notable 'kink' in the reproducibility scale, and an incoherent distribution of respondents across increasingly complex clustering solutions. The absence of cohesive groupings causes such struggles when we try to find order where structure is lacking, or perhaps not present at all.

In this scenario, the analyst has limited alternatives with which to derive k-means clusters. A tandem approach to clustering (i.e. segmenting based on factor scores rather than raw data) may be useful in cases with extreme dimensional imbalance, although this controversial technique may be detrimental if improperly applied. Otherwise, forcing an all-inclusive solution dilutes the

homogeneity of the segments, while excluding numerous outliers from the segmentation scheme denigrates the purpose of classifying respondents in the first place. The lack of an outcome or prediction from the k-means model means that we are left to interpret the significance of either solution on the basis of associations with exogenous variables. At best, our diffused segmentation scheme may be atheoretical, yet workable. But at worst, the grouping may confound conventional wisdom or defy meaningful interpretation.

At this point, the analyst may consider exploring alternative methods for segmenting the respondent pool. Hierarchical clustering provides a bit more flexibility, since the analyst can examine the segmentation "path" — either divisive or agglomerative — to determine where, if at all, the clustering begins to yield fragmented results. But this method is no less vulnerable to diffused data, and potentially exacerbates the problem by "chaining" poorly-related segments together. If a viable dependent variable is available, the segmentation could be derived from tree-based methods such as CHAID or CART, which are particularly good at exploratory analysis of descriptive variables to find unique characteristic combinations that predict the desired outcome. However, where tree-based methods are good at mining variable interactions, they are decidedly poor at providing a holistic understanding of market structure. In other words, tree-based segmentation is useful for data miners and direct marketers, but not for VPs of Marketing or corporate strategists.

One reason why tree-based methods are so attractive, however, is that they produce definitive segments that are intuitive to the end-user. By way of example, an end node that is defined as "IT professionals from companies of 5,000 or more employees in the manufacturing sector" requires no further explanation. Compare this to a hypothetical k-means segment, which might be described as "predominantly consisting of IT professionals with a tendency to come from companies of 5,000 or more employees, and a higher-than-average incidence of manufacturing." The latter example simply isn't as compelling, since k-means creates segments based on proximity to an average disposition rather an absolute classification.

Which brings us back to our earlier question: Why do we care about the average profile? Clearly there are numerous examples where it is most effective to use the average (centroid) profile, especially if the segment is characterized by a normal distribution of attitudes and opinions. If a marketer is attempting to maximize the reach of a product or message to the largest possible audience, the best strategy is to target the average profile for the intended audience segment. American electoral politics offers a basic example whereby most candidates create platforms to appeal to the influential swing voters who reside in the middle of the political spectrum.

But let's turn this perspective around to consumers and their perceptions and needs. Very few people aspire to be average, at least not in the world of marketing. Advertising exploits images and personalities we emulate, public relations campaigns trumpet the extraordinary accomplishments we admire, and viral marketing generates content that we experience and share with others. All of these channels are driven by images and messages that the consumer will strive to obtain. While the ultimate product may be delivered to a market average, the tactical and strategic marketing surrounding it will likely be positioned towards a leading edge profile.

So, for reasons of both data structure and interpretation, it may be desirable to consider an alternative paradigm for investigating segments, especially if k-means is not providing an

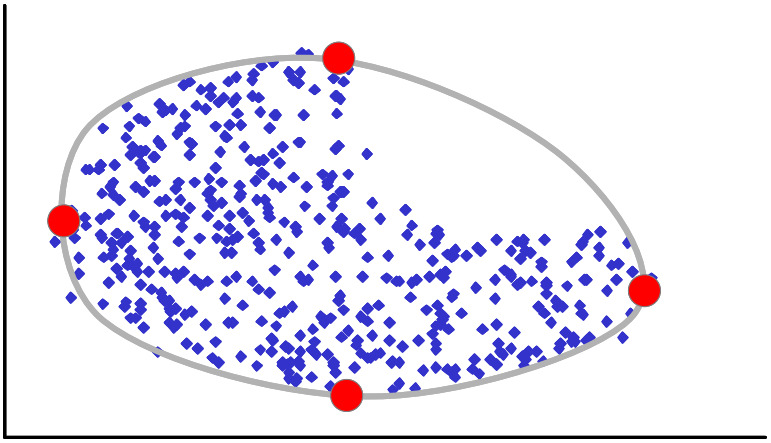
intuitive solution. What if, instead of searching for average profiles, we search for extreme definitions that exemplify the differences our segmentation is attempting to define?

ARCHETYPAL ANALYSIS

ar•che•type (*n.*) an original model from which others are copied. – *Oxford American Dictionary.*

In the context of an amorphous data cloud in perceptual space, extreme values near the fringe will be more uniformly differentiated than those on the interior. Given the right algorithm, we should be able to identify specific exterior points that best account for the shape of our perceptual cloud. Archetypal analysis offers a relatively new method of classification that searches for "pure" profiles that best define the extremities of the perceptual space.

Figure 3.
An Archetypal Approach to Segmentation



Background

Archetypal analysis has been around for roughly a decade, developed primarily in the physical sciences to identify recurring types of natural phenomenon as diverse as electrocardiogram outputs, pollution and ozone production, and physiological variations of the human head. These applications involve identifying typical (or "pure") functions around which deviations can be identified. In many such studies, archetypal analysis has been used as a data reduction method similar to principal components analysis (PCA), although without the requirement of orthogonality. Archetypal analysis has been shown to be particularly useful as a data reduction tool when PCA finds uninterpretable relationships from irregular data, in addition to its unique capability to identify respondents as a mixture of multiple "pure" profiles.

Authors Cutler and Breiman popularized archetypal analysis in their 1994 paper, and Cutler extended the application to intermittent dynamics a few years later. But aside from the efforts of a few individuals, marketing researchers have generally missed the potential application of archetypal analysis as an alternative method of market segmentation. While the full applications of archetypal analysis have yet to be realized, it is easy to envision marketing examples that reflect pure types and extremes that could be part of a segmentation exercise.

- Gatorade's advertising campaign once entreated us to "Be Like Mike." Note that there are no such campaigns asking to "Be Like Matt Harpring" of the Utah Jazz, who is also known as "the most average NBA player."
- The phrase "keeping up with the Joneses" is commonly used to describe the urge to keep pace with friends and neighbors. This phrase has meaning not because the Joneses are average, but because they typify the American Dream.
- In his book The Tipping Point, Malcolm Gladwell presents the notion that ideas and innovations are disseminated primarily through the interaction of three key groups: Connectors, Mavens, and Salesmen. Each group is an archetype — precious rather than ubiquitous — that embodies select behaviors necessary to instigate a groundswell of interest. A profiling exercise seeking to define these characteristics must look beyond homogenous groupings, and seek the defining character traits against which everyone else is to be compared.

By way of a more traditional segmentation example, consider a potential car buyer. He loves performance and wants a vehicle that evokes the image of a Steve McQueen hot rod. But other considerations creep into his purchase decision — he wants to haul lumber for his numerous home repair projects, he needs seating for his wife and two young children, he considers himself to be respectful of the environment and appreciative of good fuel economy, and he is a technophile who wants the latest and greatest in sound systems and GPS capabilities. Is there a single segment that can accommodate this "Industrious Luxu-Performance Oriented Economizing Family Man?" Perhaps, but it is more intuitive to think of this buyer according to the varying strength he is associated with pure representations of these concepts rather than a muddled combination of them all simultaneously.

The extreme profiles identified by archetypal analysis are not to be confused with outliers. A pure type is not identified simply because it lies far beyond the average distribution. Archetypes represent values that exist on the fringe, but they also explain meaningful variation across the entire sample. Each archetype is indicative of a very select type of individual, yet respondents can be described in terms of their relationships to each archetype, and the archetypes are optimally descriptive of all respondents.

Algorithmic Description

Archetypal analysis identifies extreme profiles based upon a mixture model. Archetypes are derived based on an objective function, also referred to as the archetype algorithm, which is defined below. The goal of the objective function is to create archetypes along the convex (exterior) hull of the data, such that the identified types have extreme values relative to the average profile, and that all data can be represented as a convex mixture of the archetypes.

To notate the archetype algorithm, it is necessary to start with two basic definitions.

- x_i (and x_j) refers to a data matrix of m variables for $i = 1, \ldots, n$ observations (or respondents). In other words, if we are attempting to find archetypes for 10 attributes across 250 respondents, then x_i will be a data matrix of 250*10.
- z_k refers to the archetype definitions, where each of the $k = 1, \ldots, p$ archetypes is a vector with m values (corresponding to each variable).

Our goal is to define the archetypes (z_k) as a mixture of the data values (x_i) according to the following equation:

(1) $$z_k = \sum_{j=1}^{n} \beta_{kj} x_j$$

The weights (β) identified in equation (1) are subject to two constraints that restrict the archetypes to be convex combinations of the data values. By restricting each weight to be non-negative and the summed weights for each archetype (across all individuals) to equal one, constraints (2) and (3) force the archetypes defined in equation (1) to fall on the convex (exterior) hull of the data.

(2) $$\beta_{kj} \geq 0$$

(3) $$\sum_{i=1}^{n} \beta_{kj} = 1$$

The archetypes are defined as the mixtures that minimize the residual sum of squares (RSS) found when the data values are modeled against them. This minimization equation weights the archetypal patterns and calculates their variance as follows:

(4) $$RSS(p) = \sum_{i=1}^{n} \left\| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right\|^2$$

The α weights found in equation (4) have constraints very similar to those required for the β weights, limiting each weight to be non-negative and the summed weights for each individual (across all archetypes) to equal one. Constraints (5) and (6) similarly force only convex combinations when the data values are recreated as mixtures of the archetypes.

(5) $$\alpha_{ik} \geq 0$$

(6) $$\sum_{k=1}^{p} \alpha_{ik} = 1$$

All these elements are thus combined into our overall objective function. Substituting the archetype definition from (1) into the objective function in (4) yields the following equation (7) that reveals that our archetypes must be derived by finding α's and β's that minimize the sum of squared errors, subject to the convex constraints outlined in (2), (3), (5), and (6).

(7) $$RSS(p) = \sum_{i=1}^{n} \left\| x_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} x_j \right\|^2$$

To put this in context, imagine that we are trying to model a single archetype. In this scenario, we would minimize the residual sum of squares by choosing our single archetype to be the sample mean, which leaves the RSS as simply the total sum of squares. This base RSS serves as the baseline against which subsequent solutions are evaluated.

Process Description

As previously mentioned, the objective function is solved using an iterative, or alternating, least squares solution. Prior to solving the objective function, however, this routine requires that the analyst determine the number of archetypes to derive and their accompanying starting points, or seeds, by which to define the archetypes. Conceptually, any number of methods could be used

to define these seeds, including the use of existing segments. In practice, the seeds are typically generated randomly. Cutler and Breiman advise selecting seeds that are adequately distributed to avoid problems with convergence or convergence to a local optimum.

Once starting positions are defined, the optimization algorithm alternates between solving for the α's and β's. The first stage of this process derives α's that minimize the squared error term given the seeded definition of the archetypes (z_k) for each member of the data set (x_i). Deriving the α's, also referred to as "the outer loop," involves separate non-negative least squares (NNLS) computations minimizing the (constrained) RSS for each individual, based on *m* observations (corresponding to each variable) across *p* variables (corresponding to each archetype).

Given these α's, the second stage of the algorithm derives β's that define archetypes that minimize the squared error term. Deriving the β's, also referred to as "the inner loop," can also be interpreted as deriving the best mixtures of the data for a given set of α's. As in the first stage, this involves separate NNLS computations minimizing the (constrained) RSS, but now for each archetype, based on *m* observations (corresponding to each variable) across *n* variables (corresponding to each individual).

Upon completion of each iteration, a new set of archetypes is defined according to the most recent mixture of data values. At this stage, the RSS are compared to the previous iteration to determine whether the objective function has improved notably by the iteration of α and β computations. The iterations continue until the improvement in explained variance is sufficiently small, at the discretion of the analyst.

Process Considerations

Once this process is completed, it is up to the analyst to evaluate the viability of the resulting archetypes. An integral part of this process is comparing results across multiple solutions invoking a varying number of archetypes. While there are no hard-and-fast rules for determining the ideal number of segments, there are two algorithmic considerations that should guide this decision.

The first consideration is the size of the error term, which naturally declines with the addition of more archetypes. The rate of decline is reflected in a plot of RSS(p) / RSS(1), which demonstrates a declining marginal improvement for each increasingly complex archetypal solution. As previously noted, the optimum solution with a single archetype (p=1) occurs at the mean for all variables, so the RSS for each archetype solution is computed as the ratio against the total sum of squares for the data set.

Interpreting this plot is not unlike interpreting a Scree plot for evaluating factor structures, where a flattening of the curve indicates a logical stopping point in the expansion of dimensions. When increasing the number of archetypes results in a small incremental decrease, approximately 5% as a guideline, then new archetypes are adding little explanatory power and the previous solution is likely to be sufficiently complex.

Complexity also contributes to the second consideration, in which the addition of archetypes increases the likelihood that the algorithm will produce sub-optimal results. As the iterative archetype algorithm searches for more points along the convex shell, it becomes increasingly common that it will settle upon an archetype (or archetypes) that minimize the RSS function locally, but fail to globally optimize the objective function.

To counteract sub-optimal findings, previous authors have run between 50 and 1,000 randomly-seeded trials for each solution of *p* archetypes. The trials must be compared for continuity and RSS to determine the optimal selection of archetypes. Sub-optimal solutions are virtually non-existent when computing two archetypes, but analysts must be exceedingly wary of local minima when seeking five or more archetypes. Results will vary considerably depending upon the nature of the source data, but previous studies have documented as many as three-quarters of all trials producing sub-optimal solutions.

Practitioners have recommended random seeding of the archetypal algorithm to avoid convergence difficulties, but such an approach clearly does not prevent the appearance of local minima. We suggest that the varied seeding approach taken by Sawtooth Software's CCA program could be similarly beneficial in an archetypal context. Density-based and hierarchical seeding should contribute to efficient archetype convergence since they would reflect the distributional realities of the data. The reproducibility statistic produced from CCA's trials could also be applied to archetypal solutions with similar interpretation in order to suggest the robustness of various solutions.

Interpretation

The mixture weights (α's) produced by the archetype algorithm provide a profiling measure for each respondent (*i*) on each archetype (*k*). Since the weights are constrained to a sum of unity across all archetypes, any respondent with a weight of 1.0 is in effect the archetype. Mixture weights lower than 1.0 demonstrate the relative association ("proximity") of a respondent to a given archetype, whereby lower scores indicate greater distance from a given archetypical profile.

Other than the select individuals who are uniformly congruent with a specific profile, respondents are not inherently classified into segments by the archetypal analysis. The two primary options are to assign archetype ("cluster") membership based on each individual's highest individual mixture weight, or to use an arbitrary cut-off value (e.g. greater than or equal to .5). It is up to the analyst to determine which approach is most appropriate to the analysis.

Assigning segment membership by the highest mixture score affords the benefit of comprehensive segmentation across all respondents, assuming that ties are resolved through some decision criteria. This comes at the cost of increased heterogeneity for the segment as a whole, but such standards are less of a concern since the segments are still defined solely in relation to the unchanging archetypes. Compare this to k-means, where the segment definition (centroid) fluctuates in response to the addition or removal of respondents, making the results more susceptible to marginalization in the presence of outliers.

If a cut-off is used, then various respondents will be considered outliers. Once again, this is much different than in k-means, where outliers appear throughout the fringes of the distribution, littering the perceptual space around and between the designated segment boundaries wherever the cluster spheres fail to intersect. Thus the k-means outliers have nothing in common with one another, other than the fact that their extreme perspectives might actually be quite influential.

In contrast, archetypal outliers reside in the middle of the data cloud, meaning they are inherently undifferentiated according to our extreme definitions. These outliers can also be considered to be an equal combination of multiple competing factors. Compared to our k-means outliers, archetypal outliers are less likely to be influential by nature of their undifferentiated

perspectives, but more likely to form a cohesive, and potentially influential group. In other words, archetypal outliers have the opportunity to form their own meaningful segment, whereas k-means outliers are unified only by their common lack of association to existing segments.

Case Studies

Archetypal analysis is easily applied to a wide variety of metric data, as evidenced by the physical sciences literature. This comparison focuses on archetypal analysis as applied to attitudinal ratings — a staple of k-means segmentation. The first two examples summarize "real-world" applications in which actual customer-derived data are subjected to both k-means and archetypal analysis. The latter portion of this comparison uses artificial data to evaluate each approach's ability to recover "known" segments.

Actual Data

Both "real-world" examples come from a professional services perspective. The first scenario involves a series of importance ratings for nine service characteristics across four (disguised) categories — range of services, customer support, ease of use, and cost. The nine attributes, rated on a five-point importance scale, were subjected to a k-means cluster analysis using Sawtooth Software's Convergent Cluster Analysis (CCA) package. Strong reproducibility scores and intuitive interpretation suggested a four-cluster solution.

These same ratings data were then entered into an archetypal analysis. To facilitate a direct comparison, we sought four archetypal profiles. The algorithm was run with random seeding over 50 iterations, with the most consistent, globally-optimal solution selected as the final archetype definitions. Respondents were assigned segment membership based on their proximity to the nearest archetype, regardless of weight magnitude, in order to obtain a comprehensive segmentation.

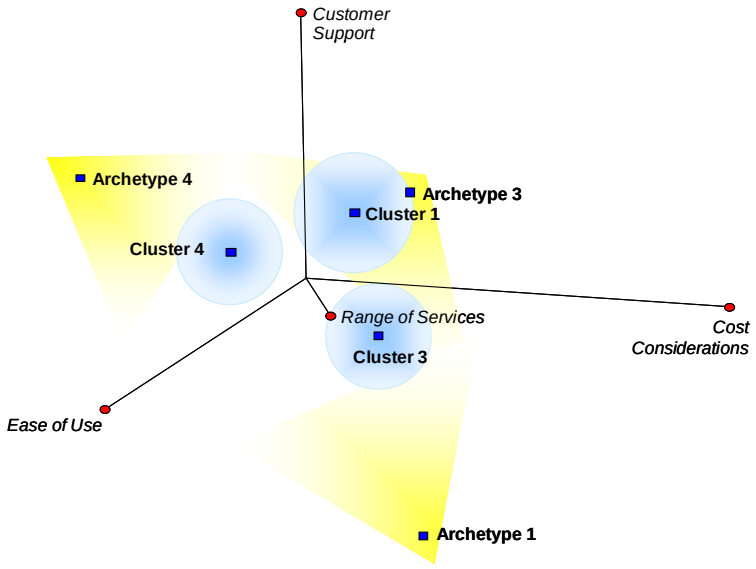
The two solutions produced substantially different segment distributions and overall segment membership. While the k-means clustering algorithm produced comparatively even distributions across segments, the archetypal algorithm lumped half of the respondents into a single archetype. Only 41% of respondents found their way from the original cluster solution into a comparable segment, with most cluster members being divided across multiple archetypes. In Table 1 (and all subsequent classification tables), optimized segment comparisons are marked in yellow, while other notable overlapping associations are marked in gray.

Table 1.
Comparison of Cluster and Archetypal Assignment (example A)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	TOTAL	(col. %)
Archetype 1	1	126	195	4	326	12%
Archetype 2	246	57	121	175	599	21%
Archetype 3	641	78	446	265	1,430	51%
Archetype 4	15	179	0	278	472	17%
TOTAL	903	440	762	722	2,827	100%
(row %)	32%	16%	27%	26%	100%	

And yet, the interpretation from cluster to archetypes is not substantially different. Cluster 1 and Archetype 3 exhibit the closest relationship, while the pairings of Cluster 3 / Archetype 1, Cluster 4 / Archetype 4, and Cluster 2 / Archetype 2 are conceptually similar but of different magnitude. A perceptual map of the two primary dimensions (excluding Cluster 2 / Archetype 2, which lie primarily in a third dimension) demonstrates the difference between the cluster centroids and archetype definitions. In all cases, the archetypes form the boundary (or extremes) of perception and extend inward, while the clusters congregate towards the center and radiate about the centroid.

Figure 4.
Perceptual Map Comparison of Clusters and Archetypes (example A)



Clusters 1, 3, and 4 are not highly differentiated from one another, as they all fall under the influence of Archetype 3 to varying degrees. This is also demonstrated by the overlap in definition between the various clusters and Archetype 3, which is also the dominant archetype. In contrast, Archetypes 1 and 4 are much more clearly defined as having preferences for services that are cost conscious and easy to use (1) or easy to use with strong customer support (4). If we are looking for a segmentation by which to define the customer base, archetypal analysis is clearly much more discriminating in the images by which we identify the relevant groups. In addition, Archetype 1's size suggests that the customer support / cost conscious consumer is even more prevalent than we might glean from the clusters.

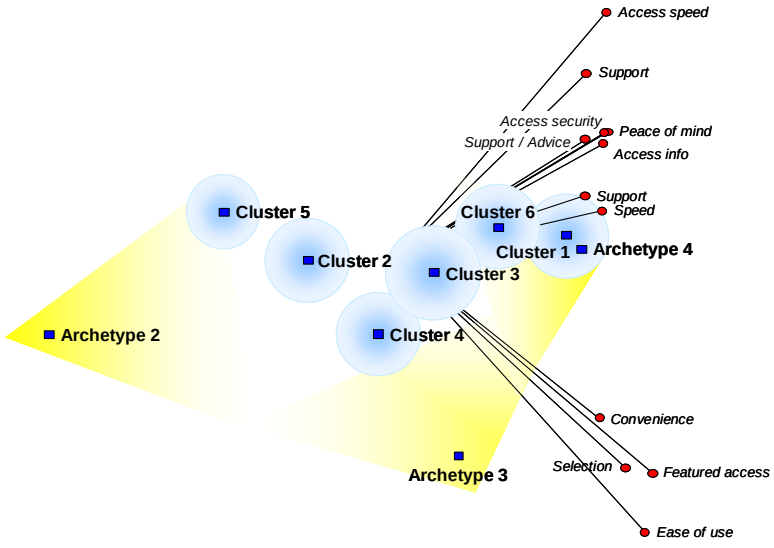
Our second example draws from a similar subject matter, beginning with a typical k-means cluster solution. A CCA-derived k-means approach suggested that a six-segment solution is a stable means to capture variation across a different set of (disguised) attributes. Unlike the previous example, in which each segment demonstrated a unique association with a variety of attributes, this time we observed clusters that were distributed across a spectrum of high / medium / low levels of association. The slight fluctuation between "security" and "luxury" attributes created a variety of clusters, but these appear to be captured primarily by three archetypes — one that values "security," another that values "luxury," and a third archetype that finds little value in either dimension.

Table 2.
Comparison of Cluster and Archetypal Assignment (example B)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	TOTAL	(col. %)
Archetype 1	0	69	36	11	4	0	120	8%
Archetype 2	0	114	0	17	99	0	230	15%
Archetype 3	0	25	75	83	1	27	211	14%
Archetype 4	233	12	241	68	0	224	778	51%
Archetype 5	0	5	2	35	3	3	48	3%
Archetype 6	0	49	62	27	8	2	148	10%
TOTAL	233	274	416	241	115	256	1,535	100%
(row %)	15%	18%	27%	16%	7%	17%	100%	

Once again, the archetypal analysis provides a sharper, albeit less nuanced, approach to defining perceptual segments. Although the archetypes were derived to match the six k-means clusters, there are three primary archetypes that account for most of the respondents. The three primary archetypes (2, 3, and 4) represent the high and low extremes, plus a well-defined mixture, while the remaining archetypes (1, 5, and 6) are comparatively small and uninteresting representations of interior clusters. For all practical purposes, the analysis produced three usable archetypes. This is especially clear when evaluated in perceptual space, which reveals that the three archetypes capture most of the variation contained within the six clusters.

Figure 5.
Perceptual Map Comparison of Clusters and Archetypes (example B)



Do we lose value by moving from six segments to three archetypes? Perhaps, if their profiles demonstrate that slight variations in perception have an impact on exogenous variables. What we clearly gain is a tighter focus on the characteristics that create extreme perceptions in the common scenario of high / medium / low associations. In this situation, it is certainly advisable to use the archetypes as a supplemental tool for understanding market structure and driving marketing messages, even if they appear to oversimplify desired segmentation characteristics.

Simulated Data

In our previous examples, the archetypes were compared against k-means clusters in an attempt to determine how well the former replicated the latter. In applying both methods against simulated data, however, we gain additional insight into the ability of each to capture "real" segments, as defined during the data construction.

To create simulated segments, we turned to the same process used to test the tandem approach to cluster analysis (Elder and Chrzan, 1999). The process defined a segment and factor structure within 35 variables, then added random variation and intentional skew to replicate the appearance of typical ratings-based data. The designation of cluster targets inherently models the perspective of k-means clustering, and provides a pre-identified anchor against which archetypes and clusters alike can be compared. As part of the tandem analysis, the synthesized variables were also condensed into their intended dimensions through principal components (factor) analysis. Both the original variables and the resulting factors were submitted to k-means and archetypal segmentation.

Table 3.
Cluster and Archetypal Assignment of Artificial Segments (raw variables)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Arch. 1	Arch. 2	Arch. 3	Arch. 4	TOTAL
Real 1	108	31	5	106	126	75	9	40	250
Real 2	7	161	25	57	19	187	23	21	250
Real 3	14	26	172	38	18	29	180	23	250
Real 4	100	24	19	107	58	34	27	131	250
TOTAL	229	242	221	308	221	325	239	215	1,000

The k-means solution performed comparably using either variables or factors. In both scenarios, the clusters correctly identified just over half of the 1,000 hypothetical respondents. In comparison, the archetypal analysis performed better with the raw data – correctly identifying 62% of the simulated segments – but slipped to 53% prediction among the factored data. The archetype performance, particularly with raw data, is interesting since the k-means clusters are located much closer to the "real" segments at the interior, as opposed to the archetypes defined on the convex hull of the data. One would assume that the proximity of the k-means centroids would produce better prediction.

Table 4.
Cluster and Archetypal Assignment of Artificial Segments (factor scores)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Arch. 1	Arch. 2	Arch. 3	Arch. 4	TOTAL
Real 1	122	87	11	30	101	111	5	33	250
Real 2	11	144	22	73	7	149	20	74	250
Real 3	13	13	186	38	23	11	149	67	250
Real 4	105	12	19	114	98	17	9	126	250
TOTAL	251	256	238	255	229	288	183	300	1,000

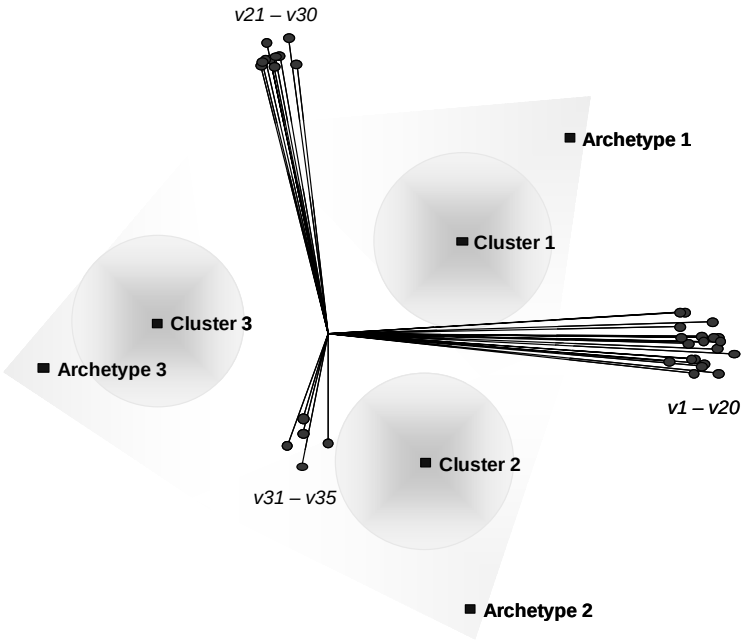
Performance aside, the archetypes provide a relatively consistent interpretation with the k-means clusters. Except for their extreme positions, the archetypes reflect the same simulated structure captured by the k-means segments, as demonstrated by the substantial overlap between the two solutions. There is a difference between raw and factored data, with the former yielding 74% segment agreement, compared against a notably stronger 89% agreement demonstrated with dimension scores. It appears that "smoothed" data denigrates the ability to discern pure archetypes, as accuracy declines across all four "actual" segments.

Table 5
Comparison of Cluster and Archetypal Assignment (artificial segments)

	Raw Variables				Factor Scores			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Arch 1	152	0	0	69	213	0	16	0
Arch 2	0	238	9	78	29	248	2	9
Arch 3	0	0	212	27	0	3	180	0
Arch 4	77	4	0	134	9	5	40	246
TOTAL	229	242	221	308	251	256	238	255

The fact that the archetypes outperform the k-means segments for raw, dimensionally-imbalanced variables suggests that there are opportunities in which classifying by extremes is more accurate than classifying by the average. Perhaps an archetypal segmentation working at the convex fringe is less susceptible to dimensional overloading than is a traditional k-means clustering, allowing a better representation of low- to moderately-dimensional data containing skewed variable representation. Our single example is insufficient for defining when archetypal opportunities are best capitalized upon, and further research in this area is needed. In the meantime, however, this analysis has demonstrated that archetypal analysis is at least a valuable supplement to k-means clustering, and could in fact be a viable replacement technique.

Figure 6.
Perceptual Comparison of Clusters and Archetypes (raw artificial variables)



CONCLUSIONS

Clustering routines such as k-means or hierarchical segmentation are reliable staples of market research, but even their best solutions can fail to provide compelling segments with clear implications for future marketing actions and messages. Centroid-based clusters are typically defined at the margins — relative distinctions that reflect tendencies and preferences rather than absolutes — which can appear more like ambiguity than certainty in the eyes of a marketing manager. Archetypal analysis offers a compelling alternative to traditional segmentation methods, because its extreme definitions are more differentiated and identifiable than the interior clusters found using hierarchical or k-means approaches. Although originally developed to typify physical phenomena, archetypal analysis appears well suited to market research applications.

Archetypal analysis is particularly suited to segmentation applications when it is particularly desirable to draw attention to segment differences and articulate clear positions. The most obvious applications occur in advertising and positioning exercises, when archetypical results are much more easily applied to creative or consulting endeavors. It could also be quite beneficial to product evaluation and conjoint utilities in scenarios when distinctive developments are required, such as when differentiating a line extension from existing products or creating an entirely new product category.

The comparisons described in this paper indicate that the results obtained from archetypal analysis are different, yet consistent with those obtained through k-means segmentation. Archetypal segments consistently diverge from k-means clusters in that their extreme profiles are more starkly drawn than those derived from centroids, and their size is consistently skewed relative to the even distribution favored by k-means clustering routines. However, both real-world and synthetic applications of archetypal analysis provide consistent insights into cluster

membership and definition, but with the added benefit of definitive characterizations that lend themselves to better segment understanding and messaging opportunities.

One of the limitations evident thus far is that archetypal analysis does not function well in high-dimensional space. The algorithm demonstrates a clear tendency to identify non-optimal solutions when searching for five or more archetypes. Our comparison also revealed a situation in which high-dimensional archetypes missed nuanced characteristics revealed through k-means clusters, although this example involved a "hi/medium/low" distribution more reminiscent of a spectrum than discrete clusters. In true multi-dimensional examples, we expect archetypal analysis to identify extreme examples with greater consistency.

Another concern springs from the relationship between archetypal definition and variable (as opposed to segment) outliers. Since archetypes are drawn according to the convex hull at the data periphery, their definitions are clearly influenced by the presence of extreme values. This phenomenon is of little concern when dealing with ratings or similarly bounded data, but could have a dramatic impact when analyzing data with a wide range of variances and distributional characteristics. For broader applications or archetypal analysis, additional research should focus on the implications of data transformation, including variable standardization, respondent centering, and weighting schemes.

With these reservations in mind, the analyst must weigh when archetypal analysis may be an appropriate substitute or supplement to k-means clustering. In most scenarios, archetypal analysis can serve as a powerful supplement to further describe existing segments in easily identifiable and differentiated terms. This is particularly relevant for generating messages or images that are intended to resonate with given segments. Archetypal analysis becomes a viable replacement option when segmentation data are distributed in a decidedly non-clustered way, such as when data values reflect a spectrum rather than true multi-dimensional agglomeration. Any k-means solution that produces a large number of outliers and low reproducibility scores should prompt consideration for replacing the centroid-based assignment with a segmentation based on archetypal analysis.

The introduction of archetypal analysis provides researchers with a newfound capability to attach distinct marketing messages and aspirational images to market segments, which makes them more meaningful as a marketing instrument. It also provides a new method of bridging the gap from the mass market to the individual, whereby segments are no longer structured as cohesive groups but rather as individuals with varying associations to multiple ideals. Perhaps more important than the effectiveness of its algorithm is the congruence of archetypal analysis with the modern marketing environment that must respect and respond to individuals while building efficient products and strategies around groups.

REFERENCES

Cui, David and James J. Cochran (2001) "Archetypal Analysis: A Consumer Description and Segmentation Tool," Unpublished working paper, Cincinnati: University of Cincinnati.

Cutler, Adele and Leo Breiman (1994) "Archetypal Analysis," *Technometrics* 36 (November): 338-347.

Elder, Andrew and Keith Chrzan (1999) "Knowing When to Factor: Simulating the Tandem Approach to Cluster Analysis," *Sawtooth Software Conference Proceedings*, Sequim: Sawtooth Software.

Fasulo, Daniel (1999) "An Analysis of Recent Work on Clustering Algorithms," *Technical Report #01-03-02*, Seattle: University of Washington.

Ortigueira, Manuel Duarte (1998) "Archetypal ECG Analysis," *Proceedings of RECPAD-98*, Lisbon, Portugal: Instituto Superior Técnico.

Pinnell, Jon (2003) "Customer Relationship – Manage, Measure or Just Understand?" *Proceedings of Technovate Conference*, Cannes: ESOMAR.

Riedesel, Paul (2003) "Archetypal Analysis in Marketing Research: A New Way of Understanding Consumer Heterogeneity," http://www.action-research.com/archtype.html.

Sawtooth Software (1995), *CCA System*. Sequim: Sawtooth Software.

Stone, Emily and Adele Cutler (1996), "Introduction to Archetypal Analysis of Spatio-Temporal Dynamics," *Physica D* 90 (February 1): 110-131.

Stone, Emily and Adele Cutler (1996), "Archetypal Analysis of Spatio-Temporal Dynamics," *Physica D* 90 (February 1): 209-224.

PERSPECTIVES ON ADVANCED METHODS

TRADE-OFF VS. SELF EXPLICATION IN CHOICE MODELING: THE CURRENT CONTROVERSY

LARRY GIBSON
ERIC MARDER ASSOCIATES, INC.

A controversy has recently emerged about an issue widely considered settled — the use of self explicated weighting in choice modeling. This paper reviews the background of the controversy including evidence of the validity of SUMM — a particular self-explicated model, the basic weaknesses of trade-off approaches, and the possibility of a major validation study to resolve the controversy.

BACKGROUND

Choice modeling is vitally important — not only to Marketing Research but to Marketing. Along with choice experimentation, choice modeling is one of the few available techniques which yields the unambiguous, quantitative predictions desperately needed by Marketing. Unfortunately, most Marketing Research is content to simply describe the market as it currently exists rather than predicting the effect of the marketers' decisions. Yet, as Alfred Politz pointed out many years ago, decisions have their effect in the future and therefore better decision-making requires better predictions, not better descriptions.

The sad, public, decision-making record of leading marketers is a major reason for the declining status of Marketing. Coca Cola's "New" Coke, P & G's Olestra, Miller's "Dick" advertising campaign ($100,000,000) have dramatically weakened Marketing's claim to corporate leadership.

As a result, Marketing is in trouble. Marketing Education is losing class hours and its best students; Marketing Management is losing corporate clout; and Marketing Research is being decentralized, downsized, and ignored. Of course, these are generalities and there are exceptions, but they are robust generalities.

EVOLUTION OF CONJOINT TRADE-OFF APPROACHES

Conjoint analysis evolved from our inability to answer a very old, very basic marketing question, "What's important to the customer?" We tried simply asking customers "What's important..?" but this direct question usually failed us so we cross-tabbed preference and perception to see what perceptions were associated with preference. Then we correlated and we simplified with factor analysis. We 'mapped', we segmented, and we experimented with different scales.

The idea of trade-off questions was much more promising. Even if our customers did not know or could not tell us what was important, we could infer "What's important…?" by asking them to choose between alternative combinations of attributes. We started with trade-off matrices; moved to concept ratings and rankings; and to choice-based questions today. Choice simulators were added changing the output from the frequently ambiguous, "What's

important…?" to the larger and more important question, "What if I…?" but the old focus still shows through.

Unfortunately, there is a technical cost to the use of trade-off questions. As the number of attributes and levels rises, the required number of questions also rises — exponentially and conjoint interview capacity becomes a problem.

Over the years, extraordinary intellectual and financial resources have been invested in conjoint analysis. How many papers have been written, how many conferences have been held, how many PhD. dissertations have been awarded on this subject? And a great deal of progress has been made. Self-explicated values were incorporated in the ACA and hybrid models. Partial factorial designs improved efficiency. More sophisticated simulators were developed.

Most recently, Hierarchical Bayesian analysis has made much more efficient use of the data but here too, there is a price. Conjoint analysis has become so complex that explaining it is a problem. One professional instructor fends off questions on Hierarchical Bayesian analysis by saying, "Don't worry about how it works. The software takes care of everything!"

Despite enormous effort, increasing complexity, and significant progress, the basic interviewing capacity of conjoint analysis remains limited and the implications of this limitation are seldom discussed and largely ignored.

EVOLUTION OF SUMM

In contrast, Eric Marder Associates has always focused on predicting the effect of marketing decisions. Since its founding in 1960, nearly all its work has ended with a statement of the form, "If you do 'A', your share of choice will rise 3.2% or fall 5.1%" For years, it relied primarily on a proprietary method called STEP to make these statements; only later did it develop SUMM, its proprietary choice modeling procedure.

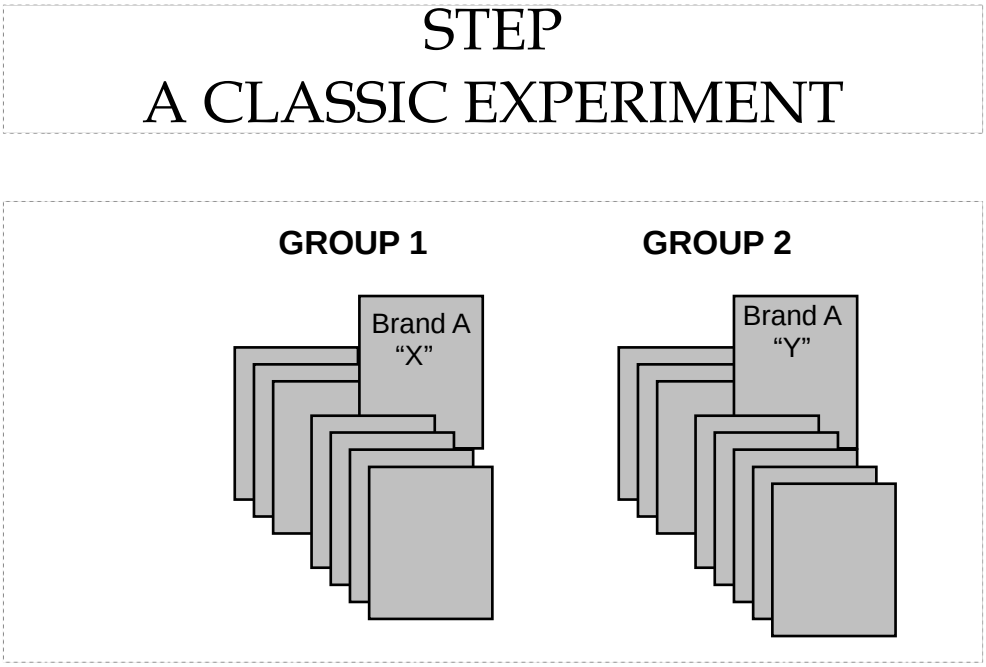
STEP: STRATEGY EVALUATION PROGRAM

STEP is a classic controlled experiment, traditionally the 'gold standard' — the ultimate methodological arbiter — of science. Over the years, literally thousands of STEP studies have been used to evaluate new product concepts, brand names, product features, brand positioning, products, advertisements, and prices. Notice the use of STEP for price testing. Nearly 200 STEP studies have been concerned with price where its ratio data is vital.

The primary characteristics of STEP are self-administration — there are no interviewers to explain the questions and add variance to the data; an isomorphic, competitive frame — test choices are the same as those in the 'real world'; equal treatment of all competitors — each competitor has a similar 'brand' page showing its price, its package, a statement about the brand; and a single sticker-allocation criterion question.

STEP respondents can only react to the choices as they are presented. They cannot identify the test brand; they cannot learn the test issues. They cannot analyze or second guess the test or its sponsors.

Conjoint users seem comfortable asking several choice questions of each respondent even though this inevitably reveals the test issues to many respondents. Perhaps they should be reminded that medical researchers have found that only ‘double-blind’ experiments with placebos produce valid findings on the *physical* outcomes which they study. Meanwhile we marketing researchers study much less stable *psychological* outcomes.



In its simplest form — most STEP studies are more complex — two randomly equivalent groups of respondents, are given test booklets and asked to choose between these brands by pasting ten stickers on the brand pages. The booklets are identical except that Group 1 sees the brand ‘A’ page with price ‘X’ or claim ‘X’ or picture ’X’ while Group 2 sees the brand ‘A’ page with price ‘Y’ or claim ‘Y’ or picture ‘Y’.

Any difference in brand ‘A’s share of choice between the groups must be *caused* by sampling error or by the difference between ‘X’ and ‘Y’. There are no other explanations.

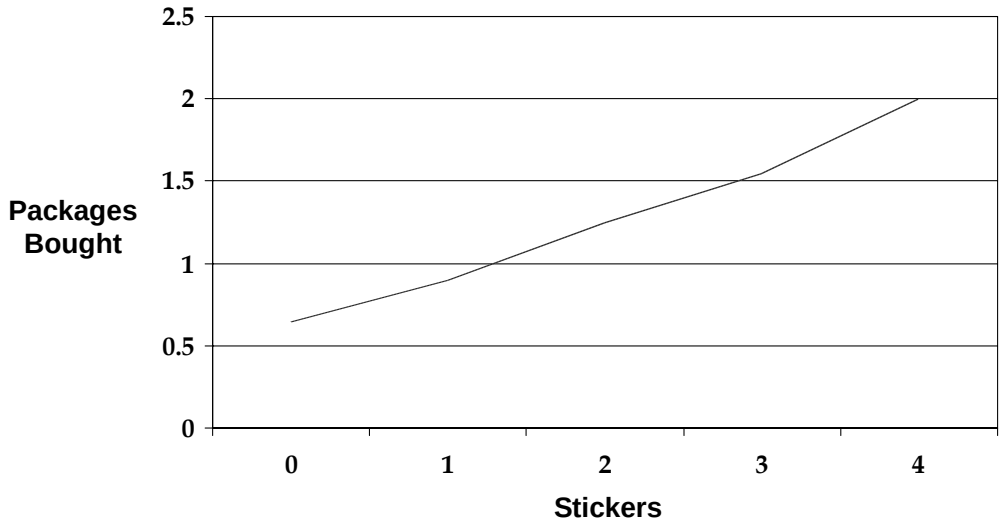
Notice that STEP findings are not the result of simulation or interpolation or inference. The different shares are — given measurable sampling error — literally made to happen by the different stimuli. We actually observe differences in customer choice caused by the different prices, or words, or pictures.

STEP has demonstrated extraordinary validity which should not be surprising given its rigorous test conditions. Of course, STEP control-cell shares conform to known market shares. Further, there are dozens of anecdotes of successful prediction of the outcome of marketing decisions — many based on the ratio properties of the data.

Even more powerful evidence of validity is provided by two cases which directly address the question, "What about individual respondents? Are those who give the client brand five STEP stickers rather than four more likely to buy the client's brand in the 'real world'?"

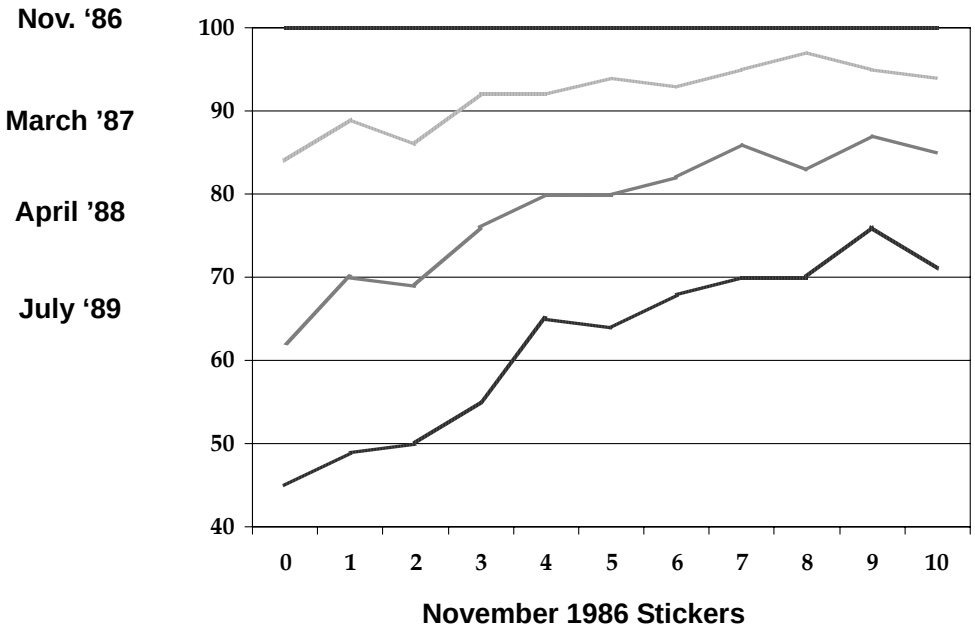
Some years ago, a STEP test was conducted to measure the concept appeal of a new baby product prior to its placement in test stores. After the new product was introduced and after product samples had been given to all new mothers in the market, test respondents were called to determine their purchases of the new product.

STEP VALIDITY
STICKERS vs. LATER BUYING



After 20 test weeks, the number of packages purchased correlated with the number of stickers previously allocated by the same respondents at .99 — in spite of the product sampling.
Even more dramatic results were found in a subscriber retention study conducted among 11,000 subscribers to a certain service in November 1986. Some current subscribers gave the service 10 stickers; others gave it 5 stickers; some gave it no stickers. Subsequently, the client rechecked his subscriber lists to learn how many of the 11,000 were still subscribing.

STEP VALIDITY
STEP STICKERS vs. RETENTION



In March 1987, 3 months later, about 95% of those who gave the service 10 stickers were still subscribers compared to less than 90% among those who allocated 0, 1, or 2 stickers.

In April 1988, 16 months later, the pattern was more distinct. Among those who gave the service no stickers, only 62% were still subscribers; among those who gave all 10 stickers, 86% were still subscribers.

In July 1989 — over 2 ½ years after the original STEP test — those little STEP stickers were even more predictive of subsequent behavior! Less than half those respondents who allocated 2 or fewer stickers were still subscribers; over 70% of those allocating 8 or more stickers were still subscribers.

Accordingly, we believe that STEP is the most valid way to evaluate marketing decisions when a relatively small number of reasonably well-defined alternatives are to be evaluated. We know of no other method which is so clearly in the tradition of scientific method or has shared such powerful evidence of validity; few others even claim to produce ratio data.

Of course, there are other, somewhat similar, methods which adopt various short cuts to lower costs. However, formal value-of-information analyses routinely show that the additional research cost is totally swamped by the profitability of even a modest improvement in marketing's decision-making batting average.

SUMM: SINGLE UNIT MARKETING MODEL

SUMM was developed for those decision-making situations for which STEP is not appropriate — when a large number of poorly-defined alternatives must be evaluated; it was not designed to replace STEP.

The primary characteristics of SUMM are the unique 'unbounded', self-explicated liking scale to measure value; the total model of choice which includes individual, subjective brand perceptions as well as values; the complete 'map' of those attributes and levels which could affect choice; and the flexible analysis program to answer "What if…?" questions.

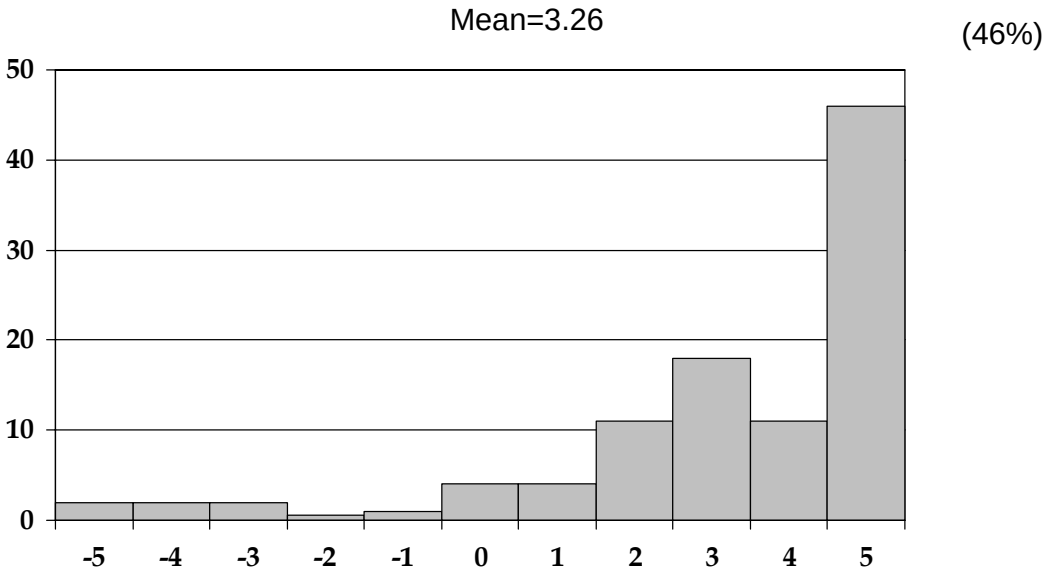
SUMM MEASUREMENT
'L & D' UNBOUNDED SCALE

"Write one or more letters
Into each box.
Like = L, LL, LLL,
or as many L's as you want
Dislike = D, DD, DDD,
Or as many D's as you want
Neutral = N

Here is the wording of the 'unbounded' scale question. Notice the natural '0' point, the absence of numbers, and the respondent's unlimited ability to express himself. Notice also that respondents are not asked to parse or analyze their decision-making process. They are simply asked how much they like or dislike the various attribute levels.

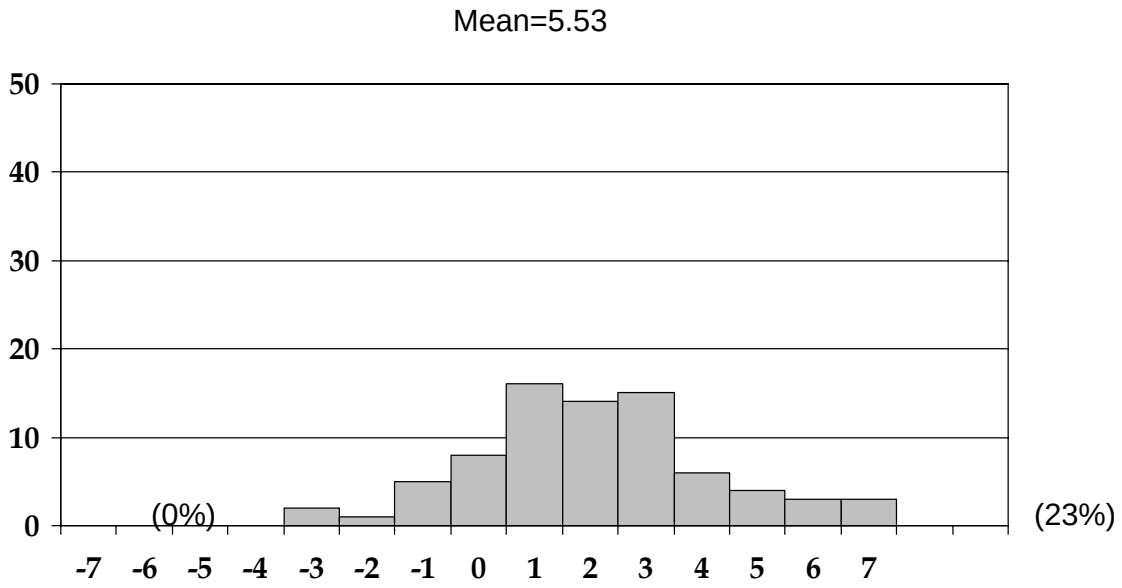
Before looking at the kind of data this scale produces, let's remind ourselves of the type of data produced by conventional 'bounded' scales.

JOHN F. KENNEDY
+5/-5 RATINGS



Here are President Kennedy's ratings on a conventional +5/-5 'bounded' scale. The data show a familiar 'J' shaped distribution with a meaningless 3.26 average. Nearly half the respondents rate President Kennedy at the end-points of the scale. Notice that the 46% at +5 cannot rate him any higher; the 2% at -5 cannot rate him any lower. Other self-explicated methods based on conventional 'bounded' scales also suffer from the resulting insensitivity.

JOHN F. KENNEDY
"L & D" RATINGS



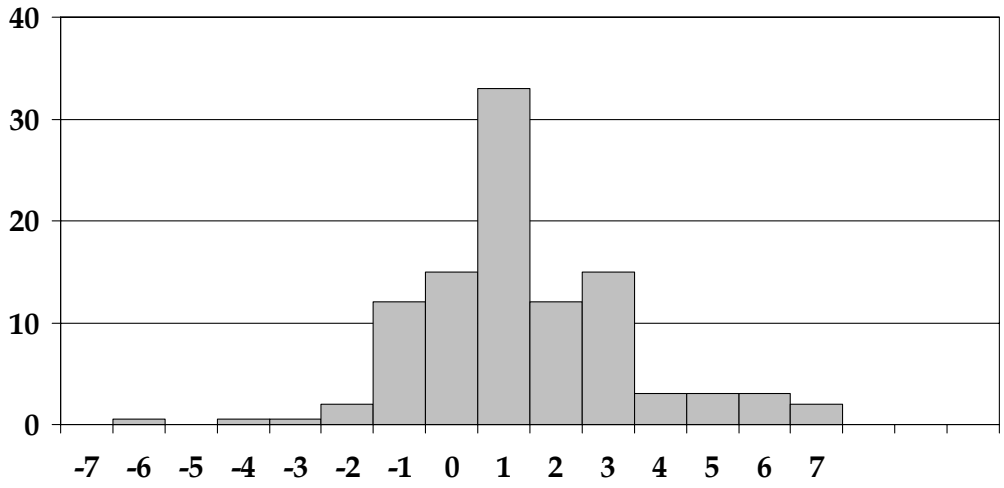
In contrast, the ratings of President Kennedy based on the 'unbounded' scale show central tendency and a higher 5.53 average. More than 7 'L's' were used by 23% of the respondents — one respondent actually wrote down 30 'L's'! Meanwhile, 8% of the respondents rated President Kennedy neutral and 8% rated him negatively.

ANALYSIS

The calculation of both observed and simulated shares of choice is straightforward. The 'L's and 'D's' are converted to numbers. (L=+1, LL=+2, etc — N=0, D=-1, DD=-2) A brand's utility for any respondent is the sum of the 'L's' minus the sum of the 'D's' for that brand's perceived attribute levels. Each respondent 'chooses' his highest scoring, his most-liked brand — winner take all.

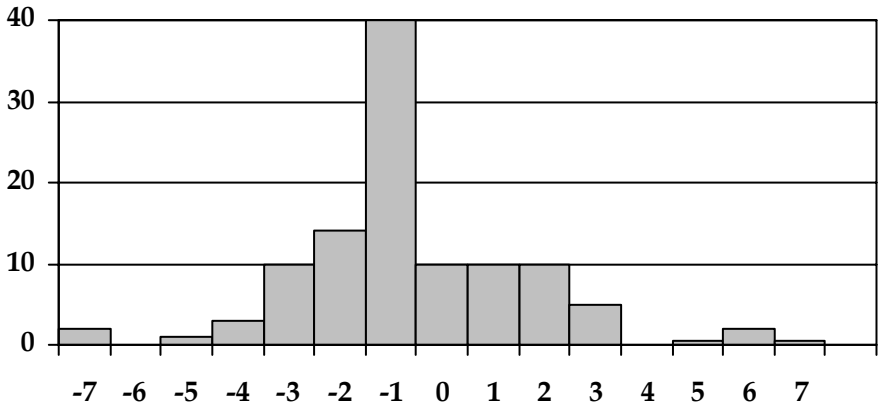
For simulation, the new perceptions to be evaluated are substituted for the observed perceptions. Individual choices and overall shares are then recalculated.

PEANUT BUTTER ATTRIBUTES
"Smooth"



Here is how respondents used the 'unbounded' scale to rate the attribute level "Smooth" in a Peanut Butter choice modeling study. Notice there is a central tendency with a mode of +1. However some respondents rated "Smooth" +6 and +7 probably making it a vital characteristic for them while others rated "Smooth" -4 and -7 probably making it a 'killing' characteristic for them.

PEANUT BUTTER ATTRIBUTES
"Chunky"



The opposing attribute level, “Chunky” received a modal rating of -1 but it was a vital characteristic for some respondents and a killing characteristic for others.

The ‘unbounded’ scale was given an unexpected test by Paul Riedesel of Action Marketing Research. In a commercial conjoint study — not a SUMM study, Riedesel conducted a split run of the alternative methods for determining utilities. Half the sample was asked conventional conjoint trade-off questions; half was asked the ‘unbounded’ scale questions.

Overall Riedesel found a similar ‘hit-rate’ in the two samples, a somewhat better share estimate with trade-off questions and HB analysis; and a significant time saving with the ‘unbounded’ scale.

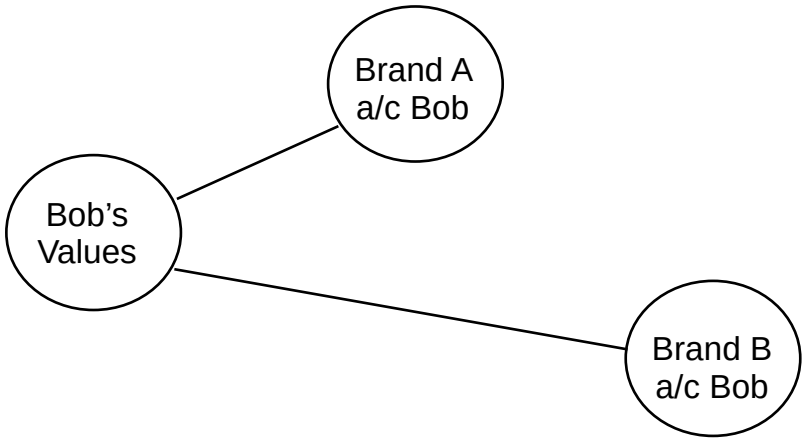
Riedesel will present the full results of his experiment at the Advanced Research Techniques Conference this summer.

CHOICE MODEL

The SUMM model of choice starts with an individual respondent’s values or desires as determined by the ‘unbounded’ liking scale. Using the interview capacity provided by this scale, the SUMM model also includes the individual respondent’s subjective brand perceptions. This addition is, we believe, critical. Customers make decisions on the basis of these subjective perceptions rather than physical reality as measured in a laboratory.

Subjective perceptions vary from individual to individual. Some customers see the Honda Accord as large, luxurious, and inexpensive; others see it as small, plain, and costly. Some Minnesota voters saw Hubert Humphrey as a liberal; others saw him as a conservative. The SUMM model handles these inconsistencies without difficulty.

SUMM
TOTAL MODEL OF CHOICE

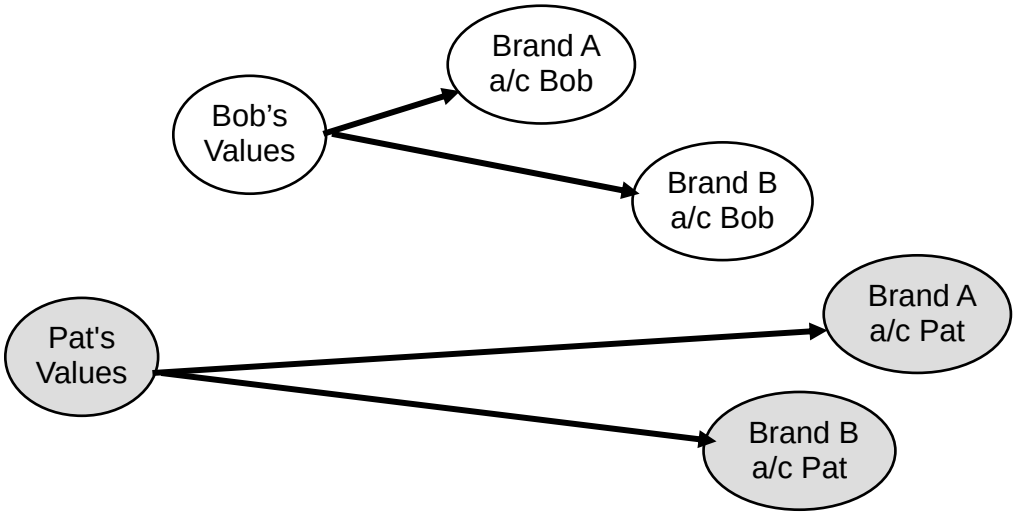


This customer 'Bob' sees Brand A as closer to his values than Brand B and in SUMM chooses Brand A. Notice that the critical issue in SUMM is not the conventional, "What's important to Bob?" but rather "What's the net (across the various attribute/levels) *relative* distance between Bob and Brand A versus the distance between Bob and Brand B?"

The critical research assumption becomes, "Does Bob respond to the liking questions and the perception questions in the same way? Are his answers internally consistent?"

Each respondent is treated as a completely separate analytical unit. Bob's values and perceptions are never added to or averaged with those of any other respondent. Bob's and Pat's data are kept totally separate. Only their final brand choices are added to compute overall share of choice.

SUMM
IDIOSYNCRATIC MODEL OF CHOICE

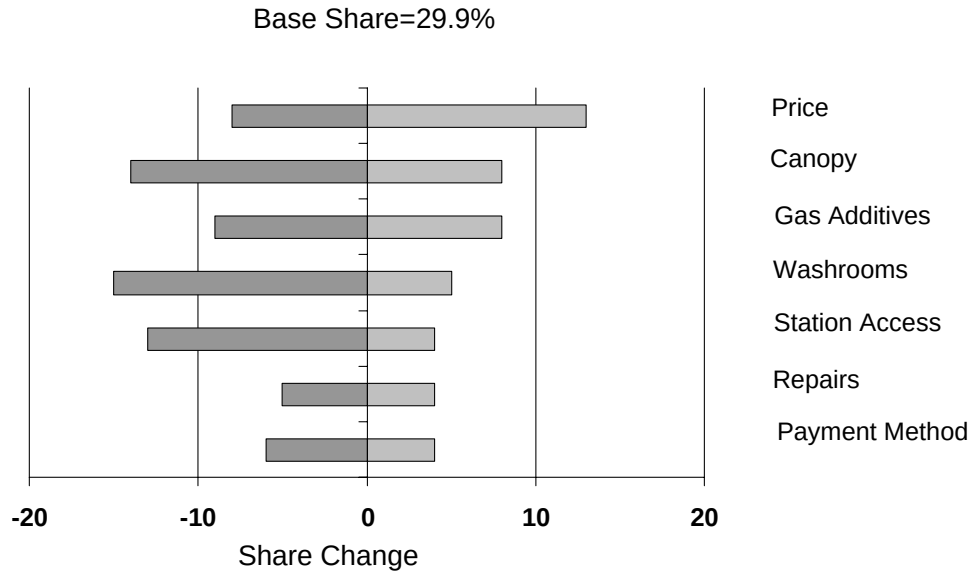


ATTRIBUTE/LEVEL "MAP"

The greater interviewing capacity of the 'unbounded' scale also enables a different approach to the design of the attribute/level 'map'. Guided by the classic rules of taxonomy, a SUMM 'map' is designed to be all-inclusive, covering all the attribute levels which could affect choice; mutually exclusive, with interactions anticipated and built in; at a common level of generality; and useful to client and respondent.

For example, the 'map' for a gasoline service station study conducted over 20 years ago would be little different today. Covering 30 attributes with 112 levels, it started with visibility, brand, and station age/condition; it ended with payment method, ATM machine, and price. Included were rest rooms — clean and safe, clean but not safe, etc. — and tire air topics irrelevant to most drivers but vital to others.

SUMM SENSITIVITY ANALYSIS SERVICE STATION



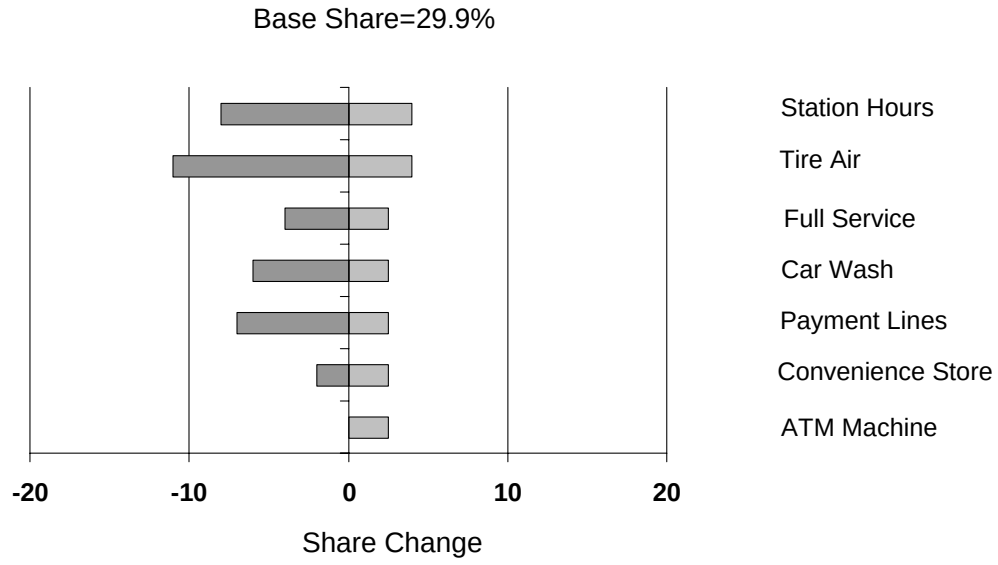
The first SUMM simulation is usually a Sensitivity Analysis in which the client's share of choice is simulated if he were seen as strongest and weakest levels on each attribute.

Price changes naturally had the largest effect on the client's share. However, notice that the client's share would rise only 13 points, to a 43% share if he were seen as having the lowest price studied — 57% still chose another brand. The client's share would fall only 8 points if he was seen as having the highest price studied — 22% still chose the client.

Accordingly, most drivers were irrelevant to the client's pricing decisions; they remained loyal to their favorite brand regardless of the client's pricing. The subset of customers who are relevant is different for each competitor and for each decision. They can only be identified by simulation.

Attributes can be important to customers positively and/or negatively. With free tire air, the client gains only 4 points but without it, the client loses 11 points.

SUMM SENSITIVITY ANALYSIS



Over the years, hundreds of SUMM studies have explored an extraordinary variety of markets from the Performing Arts to Cake Mix, from Computers to Wireless Communication, from Airlines to Toilet Paper. SUMM studies have been conducted in person, by mail, and on-line — currently about one half is conducted on line. SUMM has been used in the Orient, Australia, South America, and Europe as well as in the United States.

VALIDITY

Evidence of the validity of self-explicated methods in general and of SUMM in particular has grown over the years.

In their classic 1990 review paper, Green and Srinivasan expressed a variety of concerns about the validity of self-explication but they concluded, “The empirical results to date indicate that the self-explicated approach is likely to yield predictive validities roughly comparable to traditional conjoint analysis.” Little has since been published to modify this conclusion.

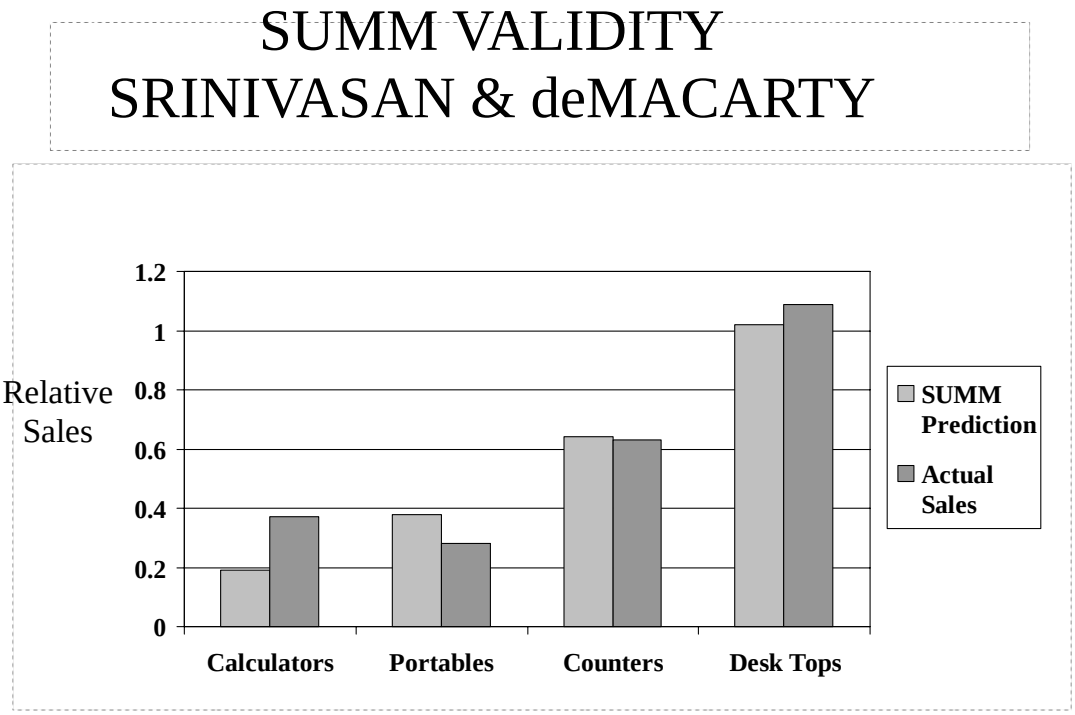
In 1991, Bondurant reported on SUMM to the Advanced Research Techniques Conference, “The level of internal consistency (hit rate) is about the same as that reported for conjoint analysis.” And “On those occasions when predictions could be verified, predictions have been remarkably accurate.”

In 1996 on receiving the Parlin award, Srinivasan noted, “I have given you examples from my own experience in conjoint analysis to show that simpler models and methods (self-explicated) often perform just as well as more elaborate models and methods.”

In The Laws of Choice (1997), Marder showed that observed shares of choice in SUMM correlate very highly with independent market data supplied by clients. Across nine widely different markets, the correlation averaged .942.

In the same book, Marder showed that findings from 32 SUMM *simulations* correlated with STEP findings — the 'Gold Standard' — at .87.

In 1999, Srinivasan and deMacarty in a prize winning Journal of Marketing Research paper presented this evidence of validity



Four pairs of Hewlett-Packard products — two calculators, two portable computers, etc. — were identified for which both SUMM share predictions and subsequent sales data were available. The analysis was based on the predicted ratio of sales for each pair versus the actual ratio of sales.

In all four cases, SUMM correctly predicted which of the pair would sell best. In three of the four cases, the predicted ratios were remarkably close.

SUMMARY

We believe that SUMM is the best choice model currently available for evaluating a large number of poorly defined alternatives. Its validity has been publicly demonstrated; it is based on a realistic model of choice; it can accommodate a complete 'map' of attribute levels; and, on the traditional criteria of the philosophy of science, it is simpler and more general than conjoint models.

Of course, like all choice models — including conjoint analysis, SUMM is limited to answering, "What if…?" questions. All choice models simply assume that the simulated changes

can actually be accomplished in customers' minds. They do not consider the difficulty of making these changes or whether, in fact, these changes *can* be made.

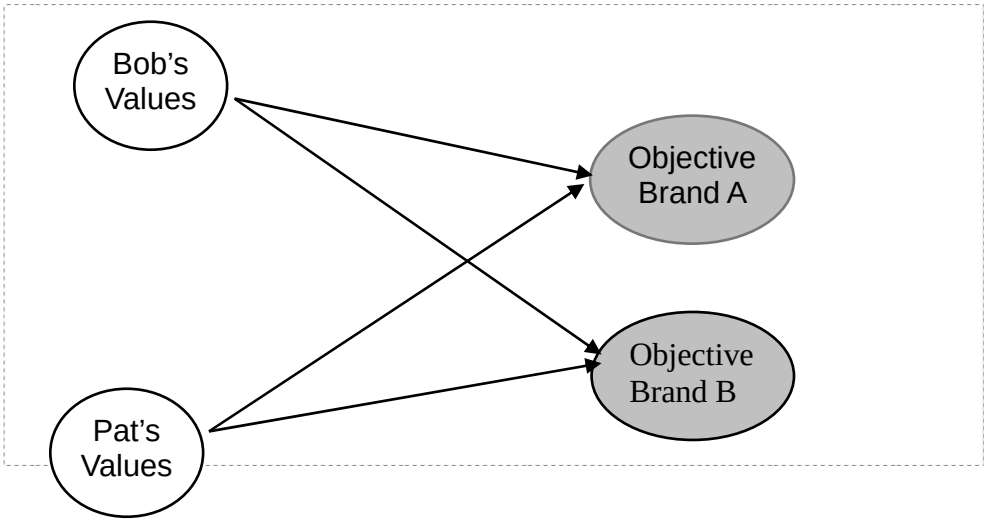
Accordingly, directly observed evidence from experiments like STEP is intrinsically superior to computer simulated evidence from a choice model.

THE CONTROVERSY

For someone accustomed to SUMM, the interview 'capacity' of conjoint analysis seems severely limited and this limited 'capacity' seems to dictate the use an inadequate model of choice and a restricted sub-set of the attribute levels which could affect choice.

Conjoint Model of Choice

CONJOINT ANALYSIS
UNREALISTIC MODEL OF CHOICE



The conjoint analysis model of choice starts with the measurement of individual, subjective utilities but it assumes that these individual, subjective utilities should be compared to the same objective, brand attributes. This assumption is clearly unrealistic; human beings choose on the basis of how they *see* the brands — not the way they really are. In addition, conjoint analysis assumes that all respondents know these objective attributes.

In some product categories — e.g. cereals, soap, and cars — *subjective* perceptions so clearly dominate choice that conjoint analysis is obviously inappropriate.

Since subjective perceptions are not even measured, conjoint analysis is blind to marketing opportunities based on misperception.

The absolute validity of the utility measurement becomes critical. The relative distance concept discussed under SUMM does not apply to conjoint analysis.

CONJOINT "MAP"

The conjoint analysis 'map' of attributes and levels is necessarily incomplete and the very concept of an all-inclusive 'map' is not even considered.

Who decides and how do they decide which attributes and levels to study? The task is clearly circular. "How can one decide which attributes and levels are important enough to include in a study designed to determine which attributes and levels are important?" Of course the decision may already have been made by the client. However, such weaker techniques as group interviews or traditional surveys of stated importance are obviously inadequate.

Attributes and levels important to some customers are inevitably omitted. How many gasoline station studies would have included 'free tire air' or the 'canopy' or' station hours'?

Since attributes and levels important to some drivers are omitted, the effect of those attributes and levels which are studied are likely to be overestimated.

Findings from the restricted 'map' are limited in both content and in time. They can only confirm or refute the significance of those attributes and levels which were selected — somehow thought to be important *at that time*. There is little opportunity for discovery. While these findings may be extremely valuable tactically, they are unlikely to serve any longer-term strategic purpose.

THE CHARGES

Despite the extraordinary intellectual and financial investment — and progress — in conjoint analysis, the basic interview 'capacity' problem remains and the ensuing weaknesses of conjoint analysis are largely ignored and seldom even discussed in public.

If self-explicated questioning is routinely incorporated into ACA and other hybrid models with evidence of its validity, why put up with the limitations of trade-off questioning?

Why does conjoint analysis continue to attract such overwhelming academic support? Is it simply its complexity and its mathematical challenge? What would a similar intellectual investment in self-explication have yielded?

THE RESPONSE

The response from prestigious conjoint analysis authorities was immediate and vociferous. Considering the direct attack on their methods, this was not at all surprising.

Several authorities pointed out that the higher validity of using both conjoint analysis and self-explicated data was well established. *Granted, assuming that only the currently popular self-explicated questions are used.*

Others claimed that the 'capacity' of conjoint analysis is not limited. Orme said, "I have personally measured over 100 levels and items." McCullough noted, "…partial profile designs with 50 attributes." Johnson asserted, "One can simulate product choice using individuals' own

perceptions." *Then why aren't complete 'maps' and more realistic models more widely considered and used?*

Some argued that conjoint analysis's limited 'capacity' is not a problem. McCullough observed, "The client doesn't always want to examine more than six attributes." Johnson said, "…many problems can be well handled by a more intensive study of a small number of attributes." *True, but why not use much more powerful experiments like STEP for these problems?*

Several shared their belief that self-explicated methods cannot be trusted. Johnson argued that this approach, "…assumes one can simply ask respondents how desirable or important various features are." Green & Krieger went further, "While Gibson may still argue that self-explicated methods show higher predictive validity than hybrid modeling, we remain unconvinced." They added, "But I wouldn't want to bet the farm on it." *Such faith-based opinions cannot be argued; they are simply unanswerable.*

Finally, they suggested that its widespread use validates conjoint analysis. Green & Krieger predicted that, "…the chances of most researchers returning to self-explicated preference modeling are little short of zero." Orme noted that, "Gibson's position swims against the prevailing tide of practice and opinion in our industry — which by itself doesn't make it wrong." (Emphasis added.) *I thought that as researchers we could agree that "truth is not determined by a show of hands.' We don't let our children get away with an 'everybody's doing it' argument.*

While these are largely thoughtful comments by distinguished authorities, it is fascinating to notice that not one commented directly on either of the two issues I raised — the necessity of including subjective perceptions in modeling customer choice and the value of a complete 'map' of attributes and levels.

So I continue to ask conjoint analysis users —

Why not use a more realistic choice model?

Why not use a complete 'map' of attributes and levels?

How can a *subset* of attributes validly predict choice?

Why the total rejection of the simpler, more general self-explicated approach?

WHAT NOW?

We could leave the discussion where it is. We could continue to talk past each other, occasionally in public but usually in private, contributing some heat but little light to this important issue. However this would deprive Marketing Research of the vigorous professional debate so necessary to the growth of a body of knowledge.

We could open up the subject in our professional meetings, encouraging more discussion and more evidence. This would be better than silence but I doubt that this controversy can be resolved by debate, argument, or isolated bits of evidence. We each *know* — from our personal experience — that we are right.

Or, we could adopt a third strategy as suggested by Paul Green, "..the only way to settle this argument is through experimentation." We could, in the grand tradition of Science, undertake a major validation experiment.

A variety of difficult issues would have to be resolved before such a study could be undertaken. Here are some of the issues with some possible answers.

> Which methods should be included? Paul Green has suggested full profile conjoint, choice based conjoint, adaptive conjoint, other hybrid methods, and SUMM.
>
> Who should execute the different methods? Presumably, the proponents of each should at least oversee their execution.
>
> How can the integrity of the study be insured? An independent, professional referee is probably required.
>
> What cases and kinds of markets should be studied? Cases requiring a large and a small number of attributes, product and service examples, markets characterized by subjective and by objective attributes.
>
> What should be the criterion measure? A real world outcome known only to the referee, a priori.
>
> How would the study be financed?

CONCLUSIONS

Choice modeling is a vitally important subject to Marketing as well as Marketing Research. At present, trade-off methods dominate choice modeling practice and academic interest. However SUMM, a particular self-explicated choice method, is simpler, more general, and has demonstrated validity. Comparisons between the two approaches highlight the significant limitations of trade-off methods. A major validation study is needed to resolve the controversy.

We invite your input to the design and execution of such a study.

REFERENCES

BonDurant, W. R. (1991), "Marketing With Choice Models," Second Annual Advanced Research Techniques Forum, American Marketing Association, Beaver Creek, Colo.

BonDurant, W. R. (1998) "Simulate Mistakes, Introduce Winners", Pre-Testing & Market Probing Summit, Institute for International Research

Gibson, L. D. (2001) "What's Wrong With Conjoint Analysis?", *Marketing Research*, Winter 2001, 16-19.

Gibson, L. D. (2002) "Conjoint Analysis Still Misses the Mark", Marketing Research, Spring 2002, 49-50.

Green, P. E. and Srinivasan, V. (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, (54), 3-19.

Green, P. E. and Krieger, A. M. (2002), "What's Right with Conjoint Analysis?", *Marketing Research*, Spring 2002, 24-27.

Hardy, H. S. (1990) *The Politz Papers*, Chicago: American Marketing Association.

Johnson, R. M. (2002) "Gibson Errs on Points of Fact," *Marketing Research*, Spring 2002, 47-48.

Marder, E. (1997) *The Laws of Choice: Predicting Customer Behavior*. New York: The Free Press.

Marder, E. (1999) "The Assumptions of Choice Modeling: Conjoint Analysis and SUMM," *Canadian Journal of Marketing Research*, (18) 3-14.

McCullough, D. (2002) "What's Wrong with Gibson's Arguments?", *Marketing Research*, Spring 2002, 48-49.

Orme, B. (2002) "A Controversial Article Inspires Debate, Perspectives on Recent Debate over Conjoint Analysis and Modeling Preferences with ACA", Sawtooth Software Research Paper Series 2002, 1-2.

Srinivasan, V. and deMcCarty, P. (1999) "Predictive Validation of Multi-Attribute Choice Models," *Marketing Research*, Winter1999/Spring 2000, 1-6.

Srinivasan, V. (1996) "Conjoint Analysis and the Robust Performance of Simpler Models and Methods", Acceptance speech on receiving the Charles Coolidge Parlin Award, Annual Marketing Research Conference, American Marketing Association.

COMMENT ON GIBSON

BRYAN ORME
SAWTOOTH SOFTWARE

We appreciate Larry coming to the Sawtooth Software Conference to challenge what might be characterized as tenets among the research community related to conjoint analysis. Our industry is so dominated by researchers that generally have very little negative to say about conjoint analysis that it is refreshing to be reminded that there are other intriguing approaches and energetic individuals willing to challenge the status quo.

Larry and I have spoken by phone and sparred in print on a few occasions over the last year. We have found much common ground, and we have agreed to disagree on a number of points. We share a common belief that much market research conducted today is not very useful. We also agree that some of the more valuable market research projects involve powerful choice simulators that yield strategic insights for management. But we tend to disagree on the best methods for developing those choice simulators.

I believe the main thrust of Larry's arguments against conjoint analysis can be boiled down to a few main points.

1. The repeated choice tasks/cards in conjoint analysis bias respondents' answers.
2. Conjoint analysis is limited to studying fewer attributes than are often needed by managers to understand the complete picture affecting the business decision.
3. If self-explicated models can produce results on par with more sophisticated conjoint/choice models, why not use the simpler approach?
4. Without information regarding how respondents perceive different brands to perform on the various attributes, the choice simulator cannot be complete.

I'll speak to each objection in turn.

REPEATED TASKS PROBLEMATIC?

In addition to SUMM, Larry's firm uses a form of discrete choice modeling (STEP) in which respondents only evaluate a single choice scenario. He has argued that to ask repeated conjoint/choice tasks with the levels varying across the profiles tips the researcher's hand regarding what is being studied and biases the subsequent answers.

In a 1996 paper delivered at A/R/T ("How Many Questions Should You Ask in Choice-Based Conjoint Studies") Rich Johnson and I examined 8 commercially-collected CBC data sets. Because each data set used randomized choice tasks, we could estimate a full set of part worths using each choice task. We compared the estimates using just the first choice task with those using all choice tasks and on average found that part worths developed only from the first task correlated 0.91 with those derived from all choice tasks. However, we also noted that as the interview progressed, derived importance for brand decreased, whereas the derived importance for price increased.

I think there are two main points to take away from this: 1) responses to repeated choice tasks produce very similar information as those from just the first task, 2) learning effects exist and can affect the part worths for later tasks relative to earlier. In practice, researchers have felt that the benefits of collecting additional information from each respondent have outweighed the drawbacks of the learning effects that can be observed from using repeated tasks. This is especially the case when one considers the benefits of obtaining individual-level preferences for a list of attributes and levels (with the resulting ability to test a huge variety of scenarios with choice simulators) as opposed to relying on aggregate analysis of a few specific choice scenarios (collected at sometimes great expense, due to sample size requirements).

LIMITED IN SCOPE?

First, there are very many studies that really only need deal with a few issues to fully answer the scope of the business problem at hand. For example, many packaged goods companies field hundreds of studies each year focused solely on varying brand, package, and price.

ACA was born in the 1980s in response to the limitations some researchers were facing when using full-profile conjoint analysis. ACA used a clever approach that combined self-explicated scaling with customized partial-profile conjoint tasks. With ACA, researchers have commonly estimated models including from 20 to 30 attributes, with as many as a dozen or so levels per attribute. Some of these models include from 100 to 150 total levels in the study. Those married to choice have discovered that they can use partial profiles in CBC, and also build very large models. There are a variety of other hybrid techniques and trade-off based approaches. Advances in computing power and in estimation techniques (for example, Latent Class and HB) have made it possible to obtain more stable estimates at the segment or individual level, for increasingly large models. To say that researchers using conjoint analysis today are limited in terms of how many attributes can effectively be measured is to ignore the advances and ingenuity of so many academics and top practitioners.

WHY NOT SIMPLER MODELS?

Larry asks why we don't simply use self-explicated models (with his improved "unbounded scale measurement"), because they have been shown to perform very well in predicting market choice. Larry commented on some research to be presented at A/R/T this year by Paul Riedesel. This research showed the self-explicated model to perform almost as well as CBC in terms of hit rates, but less well in terms of share predictions for holdout choices. More work is needed, especially with real world purchases as the criterion for success (using choice holdouts naturally favored CBC). However, this recent research illustrates a point I've seen validated a number of times: self-explicated data seem to do quite well in classifying individuals and predicting individuals' choices (hit rates); however, choice simulators based on conjoint or CBC analysis generally do better in predicting shares of preference for segments and markets. I would add that most researchers are more concerned with achieving accurate share simulators rather than accurate hit rates.

In a response to Larry printed in a recent issue of *Marketing News*, I wrote:

"I wonder how the SUMM technique might be used to model interaction effects, alternative-specific effects (e.g. designs where some attributes only apply to certain brands/alternatives), conditional pricing, or cross effects, which are important for many choice studies."

I stand by those concerns.

PREFERENCES WITHOUT PERCEPTIONS INCOMPLETE?

On this final point, I have less personal experience and conviction. Combining perceptual information and preference part worths is not new. My colleagues Rich Johnson and Chris King developed choice simulators that used part worths mapped to each respondent's perceptions of brand performance quite a bit when at John Morton Company in the late 70s and early 80s. One of their colleagues, Harla Hutchinson, delivered a paper on this topic in the 1989 Sawtooth Software Conference entitled "Gaining a Competitive Advantage by Combining Perceptual Mapping and Conjoint Analysis."

Based on conversations with Rich and Chris, combining perceptual information and preference part worths was not without problems. The perceptual information often seemed to dominate the overall sensitivity of the simulator. And, working with a model in which attributes did not necessarily have specific objective meaning, but that were mapped to subjective perceptions for each individual, made it difficult to assess how concrete changes to product specifications might affect demand.

THE UNBOUNDED SCALE

I think the "unbounded scale" Larry advocates is a very interesting idea and potentially valuable contribution, about which I truly would like to see more research, hopefully at this conference. Again, we appreciate Larry's contribution to these proceedings.

PERSPECTIVES BASED ON 10 YEARS OF HB IN MARKETING RESEARCH

GREG ALLENBY
OHIO STATE UNIVERSITY
PETER E. ROSSI
UNIVERSITY OF CHICAGO

ABSTRACT

Bayes theorem, and the Bayesian perspective to analysis, has been around for hundreds of years. It has just recently experienced a tremendous increase in popularity. In this paper, we describe the Bayesian approach, the recent revolution in marketing research that has occurred, and the revolution that is about to occur.

1. INTRODUCTION

Hierarchical Bayes (HB) methods were first applied to marketing problems in the early 90s. Since, this time, over 50 HB papers have been published in marketing journals. The popularity of the HB approach stems from several unique advantages afforded this approach over conventional methods. The Bayesian approach offers more accurate solutions to a wider class of research problems than previously possible, facilitating the integration of data from multiple sources, dealing directly with the discreteness (i.e., lumpiness) of marketing data, and tracking the uncertainty present in marketing analysis, which is characterized by many units of analysis (e.g., respondents), but relatively sparse data per unit.

The purpose of this paper is to provide an introduction to, and perspective on, Bayesian methods in marketing research. This paper begins with a conceptual discussion of Bayesian analysis and the challenge of conducting Bayesian analysis on marketing data. Hierarchical Bayes (HB) models and Markov chain Monte Carlo (MCMC) estimators are introduced as a means of dealing with these challenges. The MCMC estimator replaces difficult analytic calculations with an iterative, simple, computational procedure. When coupled with HB models, the combination offers researchers the dual capability of analyzing larger problems with more accuracy. Some challenges of implementing and estimating HB models are then discussed, followed by a perspective on where marketing research analysis is headed with this new tool. Readers interested in a formal and more complete account of these methods are referred to Rossi and Allenby (2003).

2. BAYES THEOREM AND MARKETING DATA

Bayes theorem was originally published in 1764 as "An Essay toward Solving a Problem in the Doctrine of Chances" by the Royal Society of London, England. In his essay, Bayes proposed a formal rule for accounting for uncertainty. In its simplest form, if H denotes a hypothesis and D denotes data, the theorem states that

$$\Pr(H|D) = \Pr(D|H) \times \Pr(H) / \Pr(D) \quad (1)$$

Where Pr(.) denotes a probabilistic statement. That is, Pr(H) is a probabilistic statement about the hypothesis before seeing the data, and Pr(D|H) is the probability of the data given (i.e., conditional on) the hypothesis. Bayes rule can be derived from usual operations with probabilities:

$$\begin{aligned} \Pr(H,D) &= \Pr(H) \times \Pr(D|H) \\ &= \Pr(D) \times \Pr(H|D) \end{aligned} \tag{2}$$

which, after equating the two expressions on the right of the equal signs and dividing both sides by Pr(D), yields the expression in equation (1).

Bayes theorem is useful in problems involving what has historically been called "inverse probability." In these problems, an analyst is given the data and, from that information, attempts to infer the random process that generated them by using equation (1). For example, the data (D) could be the yes-no response to the question "do you camp?" and the hypothesis (H) might be a set of factors (e.g., product ownership or enjoying other outdoor activities) that are hypothesized to be camping-related. The probability of the data given the hypothesized model can take many forms, including regression models, hazard models, and discrete choice models. Bayes theorem is used to derive probability statements about the unobserved data generating process by multiplying the probability of the data given the model, Pr(D|H), by the prior probability of the hypothesized model, Pr(H), and dividing by Pr(D).

Bayes theorem offers a formula that accounts for uncertainty in moving from the hypothesized data-generating process, Pr(D|H), to inferences about the hypothesis given the data, Pr(H|D). To a Bayesian, there is little difference between a hypothesis (H) and any other unobserved (latent) aspect of a model, including model parameters (β). Bayes rule is applied to any and all unobservables, with the goal of making inferences based on the rules of probability. A critical difference between Bayesian and non-Bayesian analysis is that Bayesians condition on the observed data (D) while non-Bayesians condition on the hypothesis (H) and model parameters (β). Non-Bayesian analysis proceeds by conditioning on the hypothesized model – i.e., assuming that the hypothesis is known with certainty – and searching for the best fitting model that maximizes Pr(D|H). The hypothesis, in reality, is never known, and such an assumption destroys the ability of the analyst to account for the uncertainty present in an analysis.

Accounting for uncertainty is important in the analysis of marketing research data. Marketing data contains large amounts of uncertainty, typically comprising many heterogeneous "units" (e.g., households, respondents, panel members, activity occasions) with limited information on each unit. These units may differ in their preferences, sensitivities, beliefs and motivations. In a conjoint analysis, for example, it is rare to have more than 20 or so evaluations, or choices, of product descriptions per respondent. In the analysis of direct marketing data, it is rare to have more than a few dozen orders for a customer within a given product category.

Researchers often report that predictions in marketing research analyses are too aggressive and unrealistic. A reduction in price of 10 or 20%, for example, results in an increase in market share that is known to be too large. Overly optimistic predictions can easily result if estimates of price sensitivity are incorrectly assumed known with certainty. When uncertainty is accounted

for, market share predictions become more realistic and tend to agree with current and past experience. The inability to accurately account for data uncertainty also creates problems when integrating data from multiple sources, when conducting an analysis in which the output from one procedure is used as input to another, and in conducting analysis of latent processes, such as the use of brand beliefs in forming consideration sets.

While Bayes theorem is a conceptually simple method of accounting for uncertainty, it has been difficult to implement in all but the simplest problems. It is typically the case that the data, D, are assumed to arise from hypothesized models where Pr(D|H) and Pr(H), when multiplied together, take a form that leads to difficulty in constructing inferences about model parameters and making predictions. This occurs, for example, in the analysis of choice data where there is assumed to exist a latent, utility maximizing process. Accordingly, until recently, researchers in marketing and other fields have tended not to use Bayesian methods, and have instead conducted analysis based entirely on Pr(D|H).

3. HIERARCHICAL BAYES (HB) MODELS

Recent developments in statistical computing have made Bayesian analysis accessible to researchers in marketing and other fields. The innovation, known as Markov chain Monte Carlo (MCMC), has facilitated the estimation of complex models of behavior that can be infeasible to estimate with alternative methods. These models are written in a hierarchical form, and are often referred to as hierarchical Bayes models. Discrete choice models, for example, assume that revealed choices reflect an underlying process where consumers have preferences for alternatives and select the one that offers greatest utility. Utility is assumed related to specific attribute levels that are valued by the consumer, and consumers are assumed to be heterogeneous in their preference for the attributes. The model is written as a series of hierarchical algebraic statements, where model parameters in one level of the hierarchy are unpacked, or explained, in subsequent levels.

$$Pr(y_{ih} = 1) = Pr(V_{ih} + \varepsilon_{ih} > V_{jh} + \varepsilon_{jh} \quad \text{for all j}) \tag{3}$$

$$V_i = x_i'\beta_h \tag{4}$$

$$\beta_h \sim \text{Normal}(\bar{\beta}, \Sigma_\beta) \tag{5}$$

where "i" and "j" denote different choice alternatives, y_{ih} is the choice outcome for respondent h, V_{ih} is the utility of choice alternative i to respondent h, x_i denotes the attributes of the i^{th} alternative, β_h are the weights given to the attributes by respondent h, and equation (5) is a "random-effects" model that assumes that the respondent weights are normally distributed in the population.

The bottom of the hierarchy specified by equations (3) – (5) is the model for the observed choice data. Equation (3) specifies that alternative j is chosen if the latent or unobserved utility is the largest among all of the alternatives. Latent utility is not observed directly and is linked to characteristics of the choice alternative and a random error in equation (4). Each respondent's part-worths or attribute weights are linked by a common distribution in equation (5). Equation (5) allows for heterogeneity among the units of analysis by specifying a probabilistic model of how the units are related. The model of the data-generating process, $Pr(D_h|\beta_h)$, is augmented with a second equation $Pr(\beta_h|\ \bar{\beta}, \Sigma_\beta)$ where $\bar{\beta}$ and Σ_β are what are known as "hyper-

parameters" of the model, i.e., parameters that describe variation in other parameters rather than variation in the data. At the top of the hierarchy are the common parameters. As we move down the hierarchy we get to more and more finely partitioned information. First are the part worths which vary from respondent to respondent. Finally, at the bottom of the hierarchy are the observed data which vary by respondent and by choice occasion.

In theory, Bayes rule can be applied to this model to obtain estimates of unit-level parameters given all the available data, $Pr(\beta_k|D)$, by first obtaining the joint probability of all model parameters given the data:

$$Pr(\{\beta_h\}, \bar{\beta}, \Sigma_\beta|D) = [\ \Pi_h\ Pr(D_h|\beta_h) \times Pr(\beta_h|\ \bar{\beta}, \Sigma_\beta)\ ] \times Pr(\bar{\beta}, \Sigma_\beta) / Pr(D) \quad (6)$$

and then integrating out the parameters not of interest:

$$Pr(\beta_k \mid D) = \int Pr(\{\beta_h\}, \bar{\beta}, \Sigma_\beta|D)\ d\beta_{-k}\ d\bar{\beta}\ d\Sigma_\beta \quad (7)$$

where "-k" denotes "except k" and $D=\{D_h\}$ denotes all the data. Equations (6) and (7) provide an operational procedure for estimating a specific respondent's coefficients (β_k) given all the data in the study (D), instead of just her data (D_k). Bayes theorem therefore provides a method of "bridging" the analysis across respondents while providing an exact accounting of all the uncertainty present.

Unfortunately, the integration specified in equation (7) is typically of high dimension and impossible to solve analytically. A conjoint analysis involving part-worths in the tens (e.g., 15) with respondents in the hundreds (e.g., 500) leads to an integration of dimension in the thousands. This partly explains why the conceptual appeal of Bayes theorem, and its ability to account for uncertainty, has had popularity problems – its implementation was difficult except in the simplest of problems. Moreover, in simple problems, one obtained essentially the same result as a conventional (classical) analysis unless the analyst was willing to make informative probabilistic statements about hypotheses and parameter values prior to seeing the data, Pr(H). Marketing researchers have historically felt that Bayes theorem was intellectually interesting but not worth the bother.

4. The MCMC Revolution

The Markov chain Monte Carlo (MCMC) revolution in statistical computing occurred in the 1980s with the publication of papers by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), eventually reaching the field of marketing with papers by Allenby and Lenk (1994) and Rossi and McCulloch (1994). The essence of the approach involves replacing the analytical integration in equation (4) with a Monte Carlo simulator involving a Markov chain. The Markov chain is a mathematical device that locates the simulator in an optimal region of the parameter space so that the integration is carried out efficiently, yielding random draws of all the model parameters. It generates random draws of the joint distribution $Pr(\{\beta_h\}, \bar{\beta}, \Sigma_\beta|D)$ and all marginal distributions (e.g., $Pr(\beta_h|D)$) instead of attempting to derive the analytical formula of the distribution. Properties of the distribution are obtained by computing appropriate sample statistics of the random draws, such as the mean, variance, and probability (i.e., confidence) intervals.

A remarkable fact of these methods is that the Monte Carlo simulator can replace integrals of any dimension (e.g., 10,000), with the only limitation being that higher dimensional integrals take longer to evaluate than integrals of only a few dimensions. A critical part of analysis is setting up the Markov chain so that it can efficiently explore the parameter space. An effective method of doing this involves writing down a model in a hierarchy, similar to that done above in equations (3) – (5).

MCMC methods have also been developed to handle the discreteness (i.e., lumpiness) of marketing choice data, using the technique of data augmentation. If we think of the data as arising from a latent continuous variable, then is it a relatively simple matter to construct an MCMC algorithm to sample from the posterior. For example, we can think of ratings scale data as arising from a censored normal random variable that is observed to be in one of k-1 "bins" or intervals for a k element scale. The resulting computational flexibility, when coupled with the exact inference provided by Bayes theorem, has lead to widespread acceptance of Bayesian methods within the academic field of marketing and statistics.

Diffusion of the HB+MCMC innovation into the practitioner community was accelerated by the existence of key conferences and the individuals that attended them. The American Marketing Association's Advanced Research Techniques (ART) Forum, the Sawtooth Software Conference, and the Bayesian Applications and Methods in Marketing Conference (BAMMCONF) at Ohio State University all played important roles in training researchers and stimulating use of the methods. The conferences brought together leading academics and practitioners to discuss new developments in marketing research methods, and the individuals attending these conferences were quick to realize the practical significance of HB methods.

The predictive superiority of HB methods has been due to the freedom afforded by MCMC to specify more realistic models, and the ability to conduct disaggregate analysis. Consider, for example, the distribution of heterogeneity, $\Pr(\beta_h| \bar{\beta}, \Sigma_\beta)$ in a discrete choice conjoint model. While it has long been recognized that respondents differ in the value they attach to attributes and benefits, models of heterogeneity were previously limited to the use of demographic covariates to explain differences, or the use of finite mixture models. Neither model is realistic – demographic variables are too broad-scoped to be related to attributes in a specific product category, and the assumption that heterogeneity is well approximated by a small number of customer types is more a hope than a reality. Much of the predictive superiority of HB methods is due to avoiding the restrictive analytic assumptions that alternative methods impose.

The disaggregate analysis afforded by MCMC methods has revolutionized analysis in marketing. By being able to obtain individual-level estimates, analysis can avoid many of the procedures imposed by analysts to avoid computational complexities. Consider, for example, analysis associated with segmentation analysis, target selection and positioning. Prior to the ability to obtain individual–level parameter estimates, analysis typically proceeded in a series of steps, beginning with the formation of segments using some form of grouping tool (e.g., cluster analysis). Subsequent steps then involved describing the groups, including their level of satisfaction with existing offerings, and assessing management's ability to attract customers by reformulating and repositioning the offering.

The availability of individual-level estimates has streamlined this analysis with the construction of choice simulators that take the individual-level parameter estimates as input, and allow the analyst to explore alternative positioning scenarios to directly assess the expected

increase in sales. It is no longer necessary to conduct piece-meal analysis that is patched together in a Rube Goldberg-like fashion. Hierarchical Bayes models, coupled with MCMC estimation, facilitates an integrated analysis that properly accounts for uncertainty using the laws of probability. While these methods have revolutionized the practice of marketing research over the last 10 years, they require some expertise to implement successfully.

5. CHALLENGES IN IMPLEMENTING HB

In addition to its widespread use in conjoint analysis because of Sawtooth Software, Bayesian models are being used by companies such as DemandTec to estimate price sensitivity for over 20,000 individual sku's in retail outlets using weekly sales data. These estimates of price sensitivity are used to identify profit maximizing retail prices. A challenge in carrying out this analysis is to estimate consumer price sensitivity given the basic assumption that rising prices are associated with declining sales for any offering, and that an increase in competitor prices will lead to an increase in own sales. Obtaining estimates of price sensitivity with the right algebraic signs is sensitive to the level of precision, or uncertainty, of the price-sales relationship.

One of the major challenges of implementing HB models is to understand the effect of model assumptions at each level of the hierarchy. In a conventional analysis, parameter estimates from a unit are obtained from the unit's data, $Pr(\beta_h|D_h)$. However, because of the scarcity of unit-level data in marketing, some form of data pooling is required to obtain stable parameter estimates. Rather than assume no heterogeneity ($\beta_h = \beta$ for all h) or that heterogeneity in response parameters follow a deterministic relationship to a common set of covariates ($\beta_h = z_h'\gamma$) such as demographics, HB models often assume that the unit-level parameters follow a random-effects model ($Pr(\beta_h | \bar{\beta}, \Sigma_\beta)$). As noted above, this part of the model "bridges" analysis across respondents, allowing the estimation of unit-level estimates using all the data $Pr(\beta_h|D)$, not just the unit's data $Pr(\beta_h|D_h)$.

The influence of the random-effect specification can be large. The parameters for a particular unit of analysis (h) now appear in two places in the model: 1) in the description of the model for the unit, $Pr(D_h|\beta_h)$ and 2) in the random-effects specification, $Pr(\beta_h| \bar{\beta}, \Sigma_\beta)$, and estimates of unit h's parameters must therefore employ both equations. The random-effects specification adds much information to the analysis of β_h, shoring up the information deficit that exists at the unit level with information from the population. This difference between HB models and conventional analysis based solely on $Pr(D_h|\beta_h)$ can be confusing to an analyst and lead to doubt in the decision to use these new methods.

The influence of the unit's data, D_h relative to the random-effects distribution on the estimate of β_h depends on the amount of noise, or error, in the unit's data $Pr(D_h|\beta_h)$ relative to the extent of heterogeneity in $Pr(\beta_h| \bar{\beta}, \Sigma_\beta)$. If the amount of noise is large or the extent of heterogeneity is small, then estimates of β_h will be similar across units (h=1,2,…). As the data become less noisy and/or as the distribution of heterogeneity becomes more dispersed, then estimates of β_h will more closely reflect the unit's data, D_h. The balance between these two forces is determined automatically by Bayes theorem. Give the model specification, no additional input from the analyst is required because Bayes theorem provides an exact accounting for uncertainty and the information contained in each source.

Finally, the MCMC estimator replaces difficult analytic calculations with simple calculations that are imbedded in an iterative process. The process involves generating draws from various distributions based on the model and data, and using these draws to explore the joint distribution of model parameters in equation (6). This can be a time consuming process for large models, and a drawback is that HB models take longer to estimate than simpler models, which attempt only to identify parameter values that fit the data best. However, as computational speed increases, this drawback becomes less important.

6. New Developments in Marketing Research

In addition to improvements in prediction, HB methods have been used to develop new marketing research methods and insights, including new models of consumer behavior, new models of heterogeneity, and new decision tools. Discrete choice models have been developed to include carry-over effects (Allenby and Lenk 1994), quantity (Arora, Allenby and Ginter 1999, Allenby, Shively, Yang and Garratt 2003), satiation (Kim, Allenby and Rossi 2002), screening rules (Gilbride and Allenby 2003) and simultaneous effects (Manchanda, Chintagunta and Rossi, 2003 and Yang, Chen and Allenby 2003). Models of satiation facilitate identifying product characteristics that are responsible for consumers tiring of an offering, and have implications for product and product line formation. Screening rules are to simplify consumer decision making, and point to the features that are needed for a brand to be considered. These features are of strategic importance to a firm because they define the relevant competition for an offering. Finally, simultaneous models deal with the fact that marketing mix variables are chosen strategically by managers with some partial knowledge of aspects of demand not observed by the market researcher. For example, the sensitivity of prospects to price changes is used by producers to design promotions and by the prospects themselves when making their purchase decisions. Prices are therefore set from within the system of study, and are not independently determined. Incorrectly assuming that variables are independent can lead to biased estimates of the effectiveness of marketing programs.

Experience with alternative forms of the distribution of heterogeneity reveals that assuming a multivariate normal distribution leads to large improvements in parameter estimates and predictive fit (see, for example, Allenby, Arora and Ginter 1998). More specifically, assuming a normal distribution typically leads to large improvements relative to assuming that the distribution of heterogeneity follows a finite mixture model. Moreover, additional benefit is gained from using truncated distributions that constrain parameter estimates to sensible regions of support. For example, negative price coefficients are needed to solve for profit maximizing prices that are realistic.

Progress has been made in understanding the nature of heterogeneity. Consumer preferences can be interdependent and related within social and informational networks (Yang and Allenby, 2003). Moreover, heterogeneity exists at a more micro-level than the individual respondent. People act and use offerings in individual instances of behavior. Across instances, the objective environment may change with implications for consumer motivations and brand preferences (Yang, Allenby and Fennell 2002). Motivating conditions, expressed as the concerns and interests that lead to action, have been found to be predictive of relative brand preference, and are a promising basis variable for market segmentation (Allenby, et.al. 2002).

The new decision tools offered by HB methods exploit availability of the random draws from the MCMC chain. As mentioned above, these draws are used to simulate scenarios related to management's actions and to explore non-standard aspects of an analysis. Allenby and Ginter (1995) discuss the importance of exploring extremes of the distribution of heterogeneity to identify likely brand switchers. Other uses include targeted coupon delivery (Rossi, McCulloch and Allenby 1996) and constructing market share simulators discussed above.

7. A Perspective on Where We're Going

Freedom from computational constraints allows researchers and practitioners to work more realistically on marketing problems. Human behavior is complex, and, unfortunately, many of the models in use have not been. Consider, for example, the dummy-variable regression model used in nearly all realms of marketing research. This model has been used extensively in advising management what to offer consumers, at what price and through what channels. It is flexible and predicts well. But does it contain the right variables, and does it provide a good representation of the process that generated the data? Let's consider the case of survey response data.

Survey respondents are often confronted with descriptions of product offerings that they encode with regard to their meaning. In a conjoint analysis, respondents likely assess the product description for correspondence with the problem that it potentially solves, construct preferences, and provide responses. The part-worth estimates that conjoint analysis makes available reveal the important attribute-levels that provide benefit, but they cannot reveal the conditions that give rise to this demand in the first place. Such information is useful in guiding product formulation and gaining the attention of consumers in broadcast media. For example, simply knowing that high horsepower is a desirable property of automobiles does not reveal that consumers may be concerned about acceleration into high-speed traffic on the highway, stop and go driving in the city, or the ability to haul heavy loads in hilly terrain. These conditions exist up-stream from (i.e., prior to) benefits that are available from product attributes. The study of such upstream drivers of brand preference will likely see increased attention as researchers expand the size of their models with HB methods.

The dummy variable regression model used in the analysis of marketing research data is too flexible and lacks the structure present in human behavior, both when representing real world conditions and when describing actual marketplace decisions. For example, the level-effect phenomena described by Wittink et al. (1992) can be interpreted as evidence of model mis-specification in applying a linear model to represent a process of encoding, interpreting and responding to stimuli. More generally, we understand a small part of how people fill out questionnaires, form and use brand beliefs, employ screening rules when making actual choices, and why individuals display high commitment to some brands but not others. None of these processes is well represented by a dummy variable regression model, and all are fruitful areas of future research.

Hierarchical Bayes methods provide the freedom to study what should be studied in marketing, including the drivers of consumer behavior. It facilitates the study of problems characterized by a large number of variables related to each other in a non-linear manner, allowing us accurately to account for model uncertainty, and to employ an "inverse probability" approach to infer the process that generated the data. HB will be the methodological cornerstone

for further development of the science of marketing, helping us to move beyond simple connections between a small set of variables.

REFERENCES

Allenby, Greg M. and Peter J. Lenk (1994) "Modeling Household Purchase Behavior with Logistic Normal Regression," *Journal of the American Statistical Association*, 89, 1218-1231.

Allenby, Greg M. and James L. Ginter (1995) "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, 392-403.

Allenby, Greg M., Neeraj Arora and James L. Ginter (1998) "On The Heterogeneity of Demand," *Journal of Marketing Research*, 35, 384-389.

Allenby, Greg, Geraldine Fennell, Albert Bemmaor, Vijay Bhargava, Francois Christen, Jackie Dawley, Peter Dickson, Yancy Edwards, Mark Garratt, Jim Ginter, Alan Sawyer, Rick Staelin, and Sha Yang (2002) "Market Segmentation Research: Beyond Within and Across Group Differences," *Marketing Letters*, 13, 3, 233-244.

Allenby, Greg M., Thomas S. Shively, Sha Yang and Mark J. Garratt (2003) "A Choice Model for Packaged Goods: Dealing with Discrete Quantities and Quantity Discounts," *Marketing Science*, forthcoming.

Arora, N. and Greg M. Allenby and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science*, 17, 29-44.

Bayes, T. (1763) "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philo. Trans. R. Soc London*, 53, 370-418. Reprinted in *Biometrika*, 1958, 45, 293-315.

Gelfand A.E. and A.F.M. Smith (1990) "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

Geman, S. and D. Geman (1984) "Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gilbride, Timothy J. and Greg M. Allenby, "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," working paper, Ohio State University.

Kim, Jaehwan, Greg M. Allenby, and Peter E. Rossi (2002) "Modeling Consumer Demand for Variety," *Marketing Science*, 21, 3, 229-250.

Manchanda, P., P. K. Chintagunta and P. E. Rossi (2003), "Response Modeling with Non-random Marketing Mix Variables," working paper, Graduate School of Business, University of Chicago.

McCulloch, R. and P.E. Rossi (1984) "An Exact Likelihood Approach to Analysis of the MNP Model," *Journal of Econometrics*, 64, 207-240.

Rossi, Peter E., Robert E. McCulloch and Greg M. Allenby (1996) "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15, 321-340.

Rossi, Peter E. and Greg M. Allenby (2003) "Bayesian Methods and Marketing," working paper, Ohio State University.

Wittink, Dick, Joel Huber, Peter Zandan and Rich Johnson (1992) "The Number of Levels Effects in Conjoint: Where Does it Come From and Can It Be Eliminated?" *Sawtooth Software Research Paper Series*.

Yang, Sha, Greg M. Allenby and Geraldine Fennell (2002) "Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions," *Marketing Science*, 21, 1, 14-31.

Yang, Sha and Greg M. Allenby (2003) "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, forthcoming.

Yang, Sha, Yuxin Chen and Greg M. Allenby (2003) "Bayesian Analysis of Simultaneous Demand and Supply," working paper, Ohio State University.

BIBLIOGRAPHY OF BAYESIAN STUDIES IN MARKETING

Ainslie, A. and Peter Rossi (1998), "Similarities in Choice Behavior across Product Categories," *Marketing Science,* 17, 91-106.

Allenby, Greg, Neeraj Arora, Chris Diener, Jaehwan Kim, Mike Lotti and Paul Markowitz (2002) "Distinguishing Likelihoods, Loss Functions and Heterogeneity in the Evaluation of Marketing Models," *Canadian Journal of Marketing Research*, 20.1, 44-59.

Allenby, Greg M., Robert P. Leone, and Lichung Jen (1999), "A Dynamic Model of Purchase Timing with Application to Direct Marketing," *Journal of the American Statistical Association*, 94, 365-374.

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research,* 35, 384-389.

Allenby, Greg M., Lichung Jen and Robert P. Leone (1996), "Economic Trends and Being Trendy: The Influence of Consumer Confidence on Retail Fashion Sales," *Journal of Business & Economic Statistics,* 14, 103-111.

Allenby, Greg M. and Peter J. Lenk (1995), "Reassessing Brand Loyalty, Price Sensitivity, and Merchandising Effects on Consumer Brand Choice," *Journal of Business & Economic Statistics*, 13, 281-289.

Allenby, Greg M. and James L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research,* 32, 392-403.

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1995), "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," *Journal of Marketing Research,* 32, 152-162.

Allenby, Greg M. and Peter J. Lenk (1994), "Modeling Household Purchase Behavior with Logistic Normal Regression," *Journal of American Statistical Association*, 89, 1218-1231.

Allenby, Greg M. (1990) "Hypothesis Testing with Scanner Data: The Advantage of Bayesian Methods," *Journal of Marketing Research*, 27, 379-389.

Allenby, Greg M. (1990), "Cross-Validation, the Bayes Theorem, and Small-Sample Bias," *Journal of Business & Economic Statistics,* 8, 171-178.

Andrews, Rick , Asim Ansari, and Imran Currim (2002) "Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery," *Journal of Marketing Research*, 87-98.

Ansari, A., Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research,* 37, 363-375.

Ansari, A., Kamel Jedidi and Sharan Jagpal (2000), "A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models," *Marketing Science,* 19, 328-347.

Arora, N. and Greg M. Allenby (1999) "Measuring the Influence of Individual Preference Structures in Group Decision Making," *Journal of Marketing Research,* 36, 476-487.

Arora, N. and Greg M. Allenby and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science,* 17, 29-44.

Blattberg, Robert C. and Edward I. George (1991), "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations," *Journal of the American Statistical Association,* 86, 304-315.

Boatwright, Peter, Robert McCulloch and Peter E. Rossi (1999),"Account-Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model", *Journal of the American Statistical Association*, 94, 1063-1073.

Bradlow, Eric T. and David Schmittlein (1999), "The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science* 19, 43-62.

Bradlow, Eric T. and Peter S. Fader (2001), "A Bayesian Lifetime Model for the "Hot 100" *Billboard* Songs," *Journal of the American Statistical Association,* 96, 368-381.

Bradlow, Eric T. and Vithala R. Rao (2000), "A Hierarchical Bayes Model for Assortment Choice," *Journal of Marketing Research,* 37, 259-268.

Chaing, Jeongwen, Siddartha Chib and Chakrvarthi Narasimhan (1999), "Markov Chain Monte Carol and Models of Consideration Set and Parameter Heterogeneity," *Journal of Econometrics* 89, 223-248.

Chang, K., S. Siddarth and Charles B. Weinberg (1999), "The Impact of Heterogeneity in Purchase Timing and Price Responsiveness on Estimates of Sticker Shock Effects," *Marketing Science,* 18, 178-192.

DeSarbo, Wayne, Youngchan Kim and Duncan Fong (1999), "A Bayesian Multidimensional Scaling Procedure for the Spatial Analysis of Revealed Choice Data," *Journal of Econometrics* 89, 79-108.

Edwards, Yancy and Greg M. Allenby (2002) "Multivariate Analysis of Multiple Response Data," *Journal of Marketing Research*, forthcoming.

Huber, J. and Kenneth Train (2001), "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths," *Marketing Letters*, 12, 259-269.

Jen, Lichung, Chien-Heng Chou and Greg M. Allenby (2003) "A Bayesian Approach to Modeling Purchase Frequency," *Marketing Letters*, forthcoming.

Kalyanam, K. and Thomas S. Shiveley (1998), "Estimating Irregular Pricing Effects: A Stochastic Spline Regression Approach," *Journal of Marketing Research*, 35, 16-29.

Kalyanam, K. (1996), "Pricing Decision under Demand Uncertainty: A Bayesian Mixture Model Approach," *Marketing Science*, 15, 207-221.

Kamakura, Wagner A. and Michel Wedel (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34, 485-498.

Kim, Jaehwan, Greg M. Allenby and Peter E. Rossi (2002), "Modeling Consumer Demand for Variety," *Marketing Science*, forthcoming.

Leichty, John, Venkatram Ramaswamy, and Steven H. Cohen (2001), "Choice Menus for Mass Customization," *Journal of Marketing Research* 38, 183-196.

Lenk, Peter and Ambar Rao (1990), "New Models from Old: Forecasting Product Adoption by Hierarchical Bayes Procedures," *Marketing Science* 9, 42-53.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green and Martin R. Young, (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15, 173-191.

Manchanda, P., Asim Ansari and Sunil Gupta (1999), "The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions," *Marketing Science*, 18, 95-114.

Marshall, P. and Eric T. Bradlow (2002), "A Unified Approach to Conjoint Analysis Models," *Journal of the American Statistical Association*, forthcoming.

McCulloch, Robert E. and Peter E. Rossi (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 217-228.

Montgomery, Alan L. (1997), "Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data," *Marketing Science* 16, 315-337.

Montgomery, Alan L. and Eric T. Bradlow (1999), "Why Analyst Overconfidence About the Functional Form of Demand Models Can Lead to Overpricing," *Marketing Science*, 18, 569-583.

Montgomery, Alan L. and Peter E. Rossi (1999), "Estimating Price Elasticities with Theory-Based Priors," *Journal of Marketing Research*, 36, 413-423.

Neelamegham, R. and Pradeep Chintagunta (1999), "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, 18, 115-136.

Otter, Thomas, Sylvia Frühwirth-Schnatter and Regina Tüchler (2002) "Unobserved Preference Changes in Conjoint Analysis," Vienna University of Economics and Business Administration, working paper.

Putler, Daniel S., Kirthi Kalyanam and James S. Hodges (1996), "A Bayesian Approach for Estimating Target Market Potential with Limited Geodemographic Information," *Journal of Marketing Research,* 33, 134-149.

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

Rossi, Peter E., Robert E. McColloch and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science,* 15, 321-340.

Rossi, Peter E. and Greg M. Allenby (1993), "A Bayesian Approach to Estimating Household Parameters," *Journal of the Marketing Research,* 30, 171-182.

Sandor, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research* 28, 430-444.

Seetharaman, P. B., Andrew Ainslie, and Pradeep Chintagunta (1999), "Investigating Household State Dependence Effects Across Categories," *Journal of Marketing Research* 36, 488-500.

Shively, Thomas A., Greg M. Allenby and Robert Kohn (2000), "A Nonparametric Approach to Identifying Latent Relationships in Hierarchical Models," *Marketing Science,* 19, 149-162.

Steenburgh, Thomas J., Andrew Ainslie, and Peder H. Engebretson (2002), "Massively Categorical Variables: Revealing the Information in Zipcodes," *Marketing Science*, forthcoming.

Talukdar, Debobrata, K. Sudhir, and Andrew Ainslie (2002), "Investing New Production Diffusion Across Products and Countries," *Marketing Science* 21, 97-116.

Ter Hofstede, Frenkel, Michel Wedel and Jan-Benedict E.M. Steenkamp (2002), "Identifying Spatial Segments in International Markets," Marketing Science, 21, 160-177.

Ter Hofstede, Frenkel, Youingchan Kim and Michel Wedel (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," Journal of Marketing Research, 34, 253-261.

Wedel, M. and Rik Pieters (2000), "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science*, 19, 297-312.

Yang, Sha and Greg M. Allenby (2000), "A Model for Observation, Structural, and Household Heterogeneity in Panel Data," *Marketing Letters,* 11, 137-149.

Yang, Sha, Greg M. Allenby, and Geraldine Fennell (2002a), "Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions," *Marketing Science*, 21, 14-31.

Yang, Sha and Greg M. Allenby (2002b), "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, forthcoming.

EXPERIMENTS WITH CBC

PARTIAL PROFILE DISCRETE CHOICE: WHAT'S THE OPTIMAL NUMBER OF ATTRIBUTES

MICHAEL PATTERSON
PROBIT RESEARCH
KEITH CHRZAN
MARITZ RESEARCH

INTRODUCTION

Over the past few years, Partial Profile Choice Experiments (PPCE) have been successfully used in a number of commercial applications. For example, the authors have used PPCE designs across many different product categories including personal computers, pharmaceuticals, software, servers, etc. In addition, PPCE designs have been the subject of numerous research projects that have both tested and extended these designs (e.g., Chrzan 2002, Chrzan and Patterson, 1999). To date, their use in applied settings and as the subject of research-on-research has shown the PPCE designs are often a superior alternative to traditional discrete choice designs particularly when a large number of attributes is being investigated. While PPCE designs are often used, no systematic research has been conducted to determine the "optimal" number of attributes to present within PPCE choice sets.

PARTIAL PROFILE CHOICE EXPERIMENTS

PPCE are a specialized type of choice-based conjoint design. Rather than presenting respondents with all attributes at once (i.e., full-profile), PPCE designs expose respondents to a subset of attributes (typically 5 or so) in each choice task. PPCE designs are particularly valuable when a study includes a large number of attributes since exposing respondents to too many attributes (e.g., more than 15) may cause information/cognitive overload causing adoption of strategies that oversimplify the decision making processes (e.g., making choices based on only 2 or 3 attributes).

Examples of choice sets for full and partial profile choice experiments for a PC study with 10 attributes might look like the following:

Full Profile:

ALTERNATIVE 1	ALTERNATIVE 2
Brand A	Brand B
2.2 GHz	3.0 GHz
256 MB RAM	512 MB RAM
40 GB hard drive	80 GB hard drive
CDRW drive	CD ROM drive
Secondary CD ROM	No Secondary CD ROM
17 " Monitor	19 " Monitor
External Speakers	No External Speakers
Standard Warranty	Extended Warranty
$1,299	$1,499

Partial Profile:

ALTERNATIVE 1	ALTERNATIVE 2
Brand A	Brand B
40 GB hard drive	80 GB hard drive
CDRW drive	CD ROM drive
Secondary CD ROM	No Secondary CD ROM
$1,299	$1,499

Like full-profile choice designs, PPCE designs are constructed according to experimental design principles. In the case of PPCE designs, the specific attributes and attribute levels that are shown are determined according to the experimental design. There are three broad categories of experimental design approaches that can be used to develop PPCE designs (Chrzan & Orme, 2001): a) manual, b) computer optimized, and c) computer randomized.

As mentioned, PPCE designs require respondents to trade off fewer attributes within a choice task compared with full-profile designs. This reduces the cognitive burden on respondents and makes the task easier resulting in less respondent error (i.e., greater consistency) during the decision making process. In other words, "respondent efficiency" is greater with PPCE designs than with full-profile designs. This has been shown across numerous studies (e.g., Chrzan & Patterson, 1999; Chrzan, Bunch, and Lockhart, 1996; Chrzan & Elrod, 1995).

One measure of Respondent Efficiency is the multinomial logit scale parameter μ (Severin, 2000). Numerous studies have shown that the ratio of scale parameters can be used to infer differences in choice consistency between different experimental conditions. Essentially, as the amount of unexplained error increases, respondents' choices become less consistent and the scale parameter decreases (the converse is also true). Thus the scale parameter measures the extent to which respondents make choices consistent with their preferences. Louviere, Hensher, and Swait (2000) outline two different approaches for identifying a model's relative scale parameter. It should be noted that although the scale parameter measures many factors that contribute to inconsistency both within and between respondents, when we randomly assign respondents to experimental conditions, we control for many of these extraneous factors. Thus, we argue that the variability that is not accounted for is a measure of the effect of task complexity on within-respondent consistency which we label respondent efficiency.

Partially offsetting the increase in Respondent Efficiency with PPCE designs is a decrease in Statistical Efficiency with these same designs. Statistical Efficiency provides an indication of the relative estimation error a given design will have compared to alternative designs. One measure of efficiency, called D-efficiency, is a function of the multinomial logit variance-covariance matrix (Bunch, Louviere, and Anderson, 1996, Kuhfeld, Tobias, and Garratt, 1994). The design efficiency of PPCE relative to full profile designs is calculated using the formula:

$$\left[\frac{\det(I(\beta))_{PPCE}}{\det(I(\beta))_{FP}}\right]^{1/p}$$

where I(β) is the "normalized" information matrix and p is the number of parameters in the model (Bunch *et al*. 1996).

There are four primary factors that influence the Statistical Efficiency of a given design (Huber and Zwerina, 1996):

- orthogonality – the greater the orthogonality the greater the efficiency
- level balance – the greater the balance among attribute levels the greater the efficiency
- overlap between alternatives – the less overlap between alternatives' levels the greater the efficiency
- degree of utility balance between alternatives – the greater the balance between alternatives the greater the efficiency

PPCE designs exhibit lower statistical efficiency relative to full-profile designs for three primary reasons. PPCE designs collect less information on a per choice set basis since fewer attributes are shown to respondents. Additionally from a conceptual basis, PPCE designs have greater overlap between alternatives (i.e., fewer attribute differences) since the attributes that are not shown can be considered to be constant or non-varying.

Research has revealed that there is a trade off between statistical and respondent efficiency (Mazzota and Opaluch, 1995; DeShazo and Fermo, 1999, Severin, 2000). With difficult tasks, as statistical efficiency increases, respondent efficiency initially increases to a point and then decreases (inverted U shape). This trade off can be expressed in terms of overall efficiency which essentially looks at D-efficiency by taking into account the estimated β parameters and relative scale. The overall efficiency of two different designs (called A and B) can be compared using the formula:

$$\left[\frac{\det(I(\beta))_{A}}{\det(I(\beta))_{B}}\right]^{1/p} \bullet \frac{\mu_A^2}{\mu_B^2}$$

Values greater than 1.0 indicate that Design A is more efficient overall than Design B. Naturally, this approach can be used to compare the efficiency of full profile designs to PPCE designs and PPCE designs relative to another.

These three efficiency metrics (statistical, respondent and overall) will be used to determine if there is an "optimal" number of attributes to present within PPCE studies. In addition, other

indices will be used to determine the number of attributes that should typically be presented within PPCE studies.

EMPIRICAL STUDY

A primary research project was designed and executed for the sole purpose of comparing PPCE designs that differ in terms of the number of attributes presented to respondents. In this study a number of comparisons are investigated including:

- efficiency levels (statistical, respondent, and overall)
- equivalency of parameter estimates
- prevalence of None
- out of sample predictive validity
- completion rates
- task perceptions

Research Design

A web-based study was conducted to address the primary research question. The sample for the research was derived from an internal customer database and individuals were sent an email inviting them to participate. To increase the response rates, the sponsor of the research was revealed and respondents who completed the survey were entered into a drawing for Personal Digital Assistants (PDAs).

A total of 714 usable completed interviews were received (the response rate was similar to other research projects using the sample same/data collection methodology). At the beginning of the survey, respondents were randomly assigned to 1 of 5 experimental conditions. Within each condition, respondents were always shown the same number of attributes during the choice tasks (see below) however across sets they were exposed to all of the attributes. For example, individuals in the 3-attribute condition always saw choice sets that contained only three attributes. Across all of the choice tasks they were given they ended up being exposed to all of the attributes. The experimental conditions and number of completed interviews within each was as follows:

Experimental Condition	Number of Respondents
3 attributes	147
5 attributes	163
7 attributes	138
9 attributes	142
15 attributes (full profile)	124
Total	714

The product category for the research was a high-end computer system. Fifteen attributes, each with three levels, were included in the experimental design (i.e., the design was 3^{15}). Each

choice set contained three alternatives plus the option of "None." Experimental designs were developed for each of the experimental conditions in a two-step process. First for each condition, a computer optimized design was constructed using SAS/QC (Kuhfeld, 2002). This design was used as the first alternative in each choice set. Then, two other alternatives were developed by using the shifting strategy discussed by Bunch et al. (1994) and Louviere, et al. (2000). For each of the conditions other than the 15-attribute condition, a total of 47 choice sets were developed (48 choice sets were constructed for the 15-attribute condition). Respondents were then randomly presented with 12 choice sets specific to their experimental condition. An additional ten choice questions were also developed for each condition and used as holdout questions to assess the predictive validity of the models. Each respondent was randomly presented with three of the holdout choice questions.

RESULTS

Efficiency metrics

We examined three measures of efficiency: statistical, respondent, and total. As we previously discussed, one common measure of statistical efficiency is D-efficiency. When examining statistical efficiency, we used the 15 attribute (full profile) condition as the baseline condition by setting its D-efficiency equal to 1.0. The other conditions were then examined relative to it and yielded the following results:

Condition	D-efficiency
3 attributes	0.23
5 attributes	0.45
7 attributes	0.53
9 attributes	0.71
15 attributes (full profile)	1.00

These results reveal that design efficiency increases as the number of attributes presented increases. For example, the full profile condition is 77% and 55% more efficient than the 3- and 5-attribute conditions, respectively.

Based only on this efficiency metric, one would conclude that full profile designs are superior. However, as we mentioned previously, we also believe that respondent and overall efficiency should be evaluated when evaluating designs.

To examine respondent efficiency, we computed the relative scale factor for each condition where the 15-attribute condition was again used as the baseline condition (i.e., its relative scale value was set equal to 1.0). The results revealed that the 3-attribute condition had the least amount of unexplained error (i.e., greatest respondent efficiency), followed by the 5-attribute condition:

Condition	Relative scale factor
3 attributes	3.39
5 attributes	2.32
7 attributes	1.78
9 attributes	1.24
15 attributes (full profile)	1.00

Combining these two metrics yields an overall efficiency metric, with the following results:

Condition	Overall efficiency
3 attributes	2.69
5 attributes	2.42
7 attributes	1.69
9 attributes	1.08
15 attributes (full profile)	1.00

These results reveal that the 3- and 5-attribute conditions have the greatest overall efficiency. From a practical perspective, these results suggest that the 3-attribute condition's greater efficiency means that using it will produce results that are as precise as those from a full profile (15-attribute) design with 63% fewer respondents ($.63 = 1 - 2.69^{-1}$). In terms of 5 attributes, one could estimate utilities to the same degree of precision as a full-profile model with 59% fewer respondents. Obviously, these findings have significant practical implications in terms of study design and sample size.

Equivalence of model parameters

Utilities were estimated for each of the five conditions and are shown below:

Attribute & Level	3 attributes	5 attributes	7 attributes	9 attributes	15 attributes
None	-0.52	-0.38	-0.08	0.01	0.22
Att1L1	-0.70	-0.36	-0.51	-0.33	-0.37
Att1L2	0.57	0.29	0.25	0.27	0.32
Att2L1	-0.52	-0.34	-0.16	-0.27	-0.02
Att2L2	0.01	0.03	0.07	0.13	0.03
Att3L1	0.67	0.53	0.55	0.26	0.35
Att3L2	-1.16	-0.62	-0.68	-0.39	-0.34
Att4L1	0.02	0.00	0.02	-0.06	-0.03
Att4L2	0.17	0.17	-0.01	0.05	-0.04
Att5L1	-0.61	-0.25	-0.39	-0.18	-0.07
Att5L2	0.19	0.11	0.10	0.13	0.07
Att6L1	-0.17	-0.43	-0.15	-0.10	0.05
Att6L2	0.11	0.19	0.13	0.15	-0.07
Att7L1	-0.39	-0.14	-0.24	-0.03	-0.19
Att7L2	-0.03	0.04	0.02	-0.02	-0.05
Att8L1	-0.55	-0.21	-0.23	-0.10	0.01
Att8L2	0.43	0.22	0.12	0.02	-0.03
Att9L1	0.22	0.28	0.19	0.08	0.05
Att9L2	0.87	0.69	0.62	0.51	0.42
Att10L1	-0.45	-0.26	-0.15	-0.14	-0.11
Att10L2	-0.20	0.02	-0.07	0.02	0.03
Att11L1	-0.02	0.12	0.09	-0.01	-0.04
Att11L2	-0.11	-0.15	-0.16	-0.04	0.12
Att12L1	-0.53	-0.57	-0.35	-0.33	-0.20
Att12L2	0.06	0.32	0.10	0.19	0.08
Att13L1	-1.37	-0.91	-0.61	-0.57	-0.44
Att13L2	0.25	0.14	0.05	0.19	0.15
Att14L1	-0.43	-0.17	-0.15	-0.13	-0.14
Att14L2	0.43	0.13	0.18	0.05	0.01
Att15L1	0.27	0.41	0.22	0.24	0.23
Att15L2	0.18	0.12	0.17	0.11	0.14

Looking at the utilities it is evident that as the number of attributes shown increases, the absolute magnitude of the coefficients decreases which suggests that respondent error increases (i.e., the scale factor decreases). This was confirmed above.

To test whether the utility vectors differed across the conditions, we used the approach outlined by Swait and Louviere (1993). The test showed that there were differences in four out the ten possible paired comparisons. Specifically, the 3-, 5-, and 7-attribute conditions were significantly different from the 15-attribute condition, and the utilities from the 3-attribute condition were significantly different from those in the 9-attribute condition ($p < .01$)

Examining the utilities, it appeared that the primary difference across the conditions was related to the None parameter: respondents appeared to be selecting none more often as the number of attributes shown increased (i.e., the utility of None increases as the number of attributes increases). We decided to constrain the None level to have the same utility in each of the model comparisons and again tested the utility vectors. The results of this comparison showed that there were still significant differences between the 3- and 5-attribute conditions versus the full-profile condition when we used the Benjamini-Hochberg (1995) method to control the α level. In the 3- vs. 15-attribute comparisons, two utilities differed whereas in 5 vs. 15, one difference was found. In both cases the utilities associated with the two PPCE conditions made more conceptual sense. The 15-attribute condition had sign reversals; specifically, the utilities should have been negative (rather than positive) based on previous research findings as well as our knowledge of the subject matter. Thus, we would argue that the utility estimates from the full profile condition were inferior in comparison to those from the PPCE conditions for those specific parameters.

In addition to testing the equivalence of the utilities, we also tested to see if the relative scale parameters (shown previously) differed across the conditions using the Swait and Louviere (1993) procedure. The results revealed that nine of the ten comparisons were significant with the 9-attribute condition versus full profile being the only test that was not significant.

Prevalence of None

As we indicated above, the primary difference between the utility vectors across the conditions was related to the "None" utility value. The table below shows the mean percentage of time within each condition that respondents selected None across the choice sets.

Condition	Mean Percentage Selecting None
3 attributes	12.8%
5 attributes	14.4%
7 attributes	20.0%
9 attributes	22.0%
15 attributes (full profile)	24.8%

Looking at the table, it is clear the selection of None increases as the number of attributes shown increases. An ANOVA revealed that there was a significant difference between the groups, $F(4,706) = 5.53$, $p < .001$, and post-hoc tests showed that usage of the None alternative was significantly lower in the 3- and 5-attribute conditions compared to the other three conditions. No other differences were significant.

Predictive Validity

To test the predictive validity of each of the conditions, within each of the conditions, we used one group of respondents to estimate each model (i.e., estimation respondents) and then another group to test the model (holdout respondents). In other words, within each of the experimental conditions, we randomly assigned respondents to one of three groups. Two of the groups were then combined, the model was estimated, and we used the third group to assess the predictive validity. This sequence was repeated a total of three times (i.e., groups one and two were used to predict group three; groups one and three predict two; groups two and three predict one). This approach minimizes the problem of "overfitting" the model since we are estimating utilities using one set of respondents and then testing the model's accuracy with a different set of respondents Elrod (1999).

In addition, we used three hold-out questions within each condition that were not used in the estimation of the models (i.e., we had 12 estimation choice sets and 3 holdout choice sets for each respondent). We looked at two measures of predictive validity, mean absolute error (MAE) and Holdout Loglikelihood criterion.

Mean Absolute Error is calculated at the aggregate (i.e,. total sample) level and involves computing the absolute value of the difference between the actual share and model predicted share for each holdout alternative. The MAE values for each of the five conditions is below:

Condition	MAE
3 attributes	0.109
5 attributes	0.100
7 attributes	0.115
9 attributes	0.093
15 attributes (full profile)	0.100

An Analysis of Variance (ANOVA) showed that there was not a significant difference between the conditions, $F(4,595) = 1.26$, $p > .05$. Thus we cannot conclude that there are differences between the conditions with respect to MAE.

We also investigated Holdout Loglikelihood criterion (LL) to assess whether this metric varied by experimental condition. The steps in calculating LL are as follows:

1. Within each condition, use multinomial logit to calculate utilities using the estimation respondents and the estimation choice sets.
2. These utilities are used to predict the probability that each alternative was selected (i.e., choice probabilities) in each of the holdout choice sets for each of the holdout respondents.
3. Among these choice probabilities, one of them was actually selected by the respondent within each choice set and is called the 'poc'.
4. For each respondent, sum the poc values across the holdout choice sets and then take the natural log of this sum in order to calculate the loglikelihood value.

5. Test whether there are differences using an ANOVA.

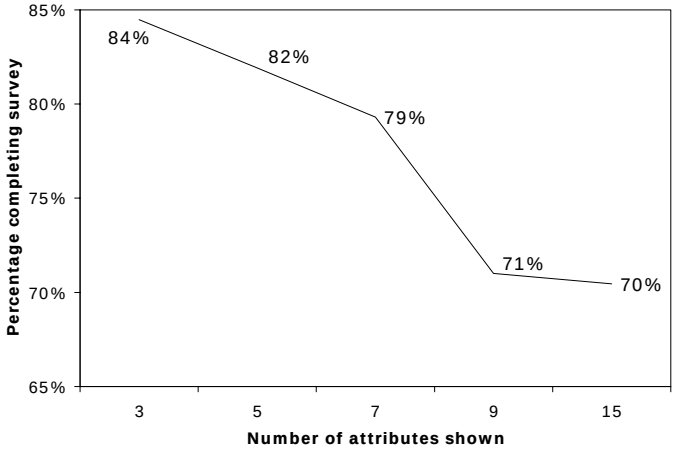
The mean LL values for each condition are shown below:

Condition	Loglikelihood Values
3 attributes	-3.48
5 attributes	-3.53
7 attributes	-3.86
9 attributes	-4.05
15 attributes (full profile)	-3.91

The ANOVA revealed that there was a statistically significant difference, $F(4,683) = 6.91$, $p < .05$. A Tukey HSD post-hoc test revealed that the 3- and 5-attribute conditions had significantly lower LL values than did the 9- and 15-attribute conditions and that the 3-attribute condition had lower LL values than did the 7-attribute condition. Since lower LL values indicate better predictive validity, we can conclude that the 3- and 5-attribute conditions are better able to predict "same format" (i.e., 3 attributes predicting 3 attributes) holdout choices than are the other conditions.

Completion Rates

Another metric we examined was the percentage of respondents who completed the survey once they had begun. We find that the percentage completing the survey monotonically decreases as the number of attributes shown increases. This can be seen in the following graph:



Task Perceptions

In the survey, individuals were asked to rate how easy and interesting the survey was and they were asked to indicate how accurately their answers in the survey would match the decisions they would make if they were making an actual purchase.

Three ANOVAs were conducted to determine if there were differences between these three metrics (i.e., easy, interesting, accuracy) and the respective results were $F(4,708) = 3.14$, $p < .05$; $F(4,709) = .74$, $p > .05$; $F(4,709) = .39$, $p > .05$. Post-hoc tests conducted using the easy metric

(the only one with a significant F value) showed that respondents in the 3-attribute condition rated the survey as being easier than respondents in the 9-attribute condition. No other differences were significant.

DISCUSSION

This research confirmed the results found in numerous other studies, namely, PPCE designs reduce overall error in comparison to full-profile discrete choice designs with large numbers of attributes. However, this research extends previous studies by demonstrating that in the present case, it appears that displaying 3 and perhaps 5 attributes is "optimal" when conducting PPCE. For general discrete choice studies, PPCE with 3-5 attributes appears to offer a superior alternative to full-profile choice studies not only because of the reduced error but also because of the higher completion rate of PPCE versus full-profile. Both of these factors suggest that PPCE studies can be conducted with fewer respondents than can comparable full-profile studies when there are a large number of attributes. Obviously this is an important benefit given constrained resources (both in terms of budgets and hard-to-reach respondents).

It should be noted that these results should be further confirmed by additional research before definitively concluding the 3-5 attributes is "optimal". In addition, at this point, PPCE studies are best suited to "generic" attributes and non-complex studies. While there is ongoing research into more complex designs (e.g., interactions, alternative-specific effects), at this point, researchers are advised to use full-profile designs when they need to estimate alternative-specific effects/interactions and/or they want to estimate cross effects (although PPCE with Hierarchical Bayes can be used to estimate individual level utilities). Moreover, it is not advised to include the None option in PPCE studies since this research has demonstrated that its usage is problematic due to its lower prevalence in these designs. Note however that the paper presented by Kraus, Lien, and Orme (2003) at this conference offers a potential solution to problems with None in PPCE studies.

Future Research

The present research investigated a study with 15 attributes. Future research should evaluate designs containing fewer attributes to assess whether the results of this research generalize to studies containing fewer attributes. We know for instance that even with as few as 7 attributes, PPCE studies offer advantages over full-profile (Chrzan, Bunch, and Lockhart, 1996). However, in these cases is 3 to 5 still optimal? Perhaps with fewer attributes, a 3-attribute PPCE study becomes much better than a 5-attribute study.

Moreover, this study was conducted via the internet using a web-based survey. Would the results of this research generalize to other data collection methodologies (e.g., paper and pencil)? Finally, if we had used a different method to assess predictive validity, would the results associated with the MAE and LL metrics have changed? For example, if all respondents were given an 8 attribute, full profile task containing two alternatives, would we have received comparable results?

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 57, 289-300.

Bunch, D.S., Louviere, J.J., Anderson, D. (1996) "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes." Working paper UCD-GSM-WP# 01-94. Graduate School of Management, University of California, Davis.

Chrzan, K (1998). "Design Efficiency of Partial Profile Choice Experiments," paper presented at the 1998 INFORMS Marketing Science Conference, Paris.

Chrzan, K., Bunch, D., Lockhart, D.C. (1996) "Testing a Multinomial Extension of Partial Profile Choice Experiments: Empirical Comparisons to Full Profile Experiments," paper presented at the 1996 INFORMS Marketing Science Conference, Gainesville, Florida.

Chrzan, K., Elrod, T. (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes." Paper presented at the AMA's 1995 Advanced Research Techniques Forum, Chicago, Illinois.

Chrzan, K., Patterson, M. (1999). "Comparing the ability of full and partial profile choice experiments to predict holdout choices." Paper presented at the AMA's 1999 Advanced Research Techniques Forum, Santa Fe, New Mexico.

DeShazo, J.R., Fermo, G. (1999). "Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency," School of Public Policy and Social Research, UCLA, California.

Elrod, Terry (2001). "Recommendations for Validation of Choice Models." Paper presented at the 2001 Sawtooth Software Conference, Victoria, BC.

Huber, J., Zwerina, K.B. (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research,* 33, 307-17.

Kraus, A., Lien, D, Orme, B (2003). "Combining Self-Explicated and Experimental Choice Data." Paper presented at the 2003 Sawtooth Software Conference, San Antonio, TX.

Kuhfeld, W. (2002). "Multinomial Logit, Discrete Choice Modeling." SAS Institute.

Kuhfeld, W., Tobias, R.D., Garratt, M. (1994) "Efficient Experimental Designs with Marketing Research Applications," *Journal of Marketing Research* 31 (November) 545-57.

Louviere, J.J., Hensher, D.A., Swait, J.D. (2000). Stated Choice Methods: Analysis and Applications. Cambridge: Cambridge University Press.

Mazzota, M., Opaluch, J. (1995). "Decision Making When Choices are Complex: A Test of Heiner's Hypothesis," *Land Economics*, 71, 500-515.

Severin, Valerie (2000). "Comparing Statistical Efficiency and Respondent Efficiency in Choice Experiments," unpublished PhD thesis, Faculty of Economics and Business, University of Sydney, Australia.

Swait, J., Louviere, J. (1993) "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research*, 30, 305-14.

DISCRETE CHOICE EXPERIMENTS WITH AN ONLINE CONSUMER PANEL

CHRIS GOGLIA
CRITICAL MIX, INC.

INTRODUCTION

The Internet has matured into a powerful medium for marketing research. Large, robust samples can quickly be recruited to participate in online studies. Online studies can include advanced experimental designs — like conjoint analysis and discrete choice modeling — that were not possible in phone or mail surveys. This paper presents the results of a controlled online research study that investigated the effect of visual stimuli and a form of utility balance on respondent participation, satisfaction, and responses to a discrete choice modeling exercise.

BRAND LOGOS AS A VISUAL STIMULI

Does the use of brand logos in an online discrete choice modeling exercise affect respondent choices? Do they encourage respondents to use brand as a heuristic — an unconscious shortcut to help them choose the best product? Or do they make the experience more visually enjoyable to the respondent and encourage them to take more time and provide carefully considered responses?

These were the questions that were addressed. Respondents were randomly assigned to one of two sample cells. One cell participated in a discrete choice modeling exercise in which all attributes — including brand — were represented by textual descriptions. The other cell participated in the same discrete choice modeling exercise except that brand logos were used in place of the brand name.

CORNER PROHIBITIONS AS A FORM OF UTILITY BALANCE

How can you gather more or better information from your choice sets? You can present more choice sets. You can also ask respondents to do more than choose the best choice through a chip allocation exercise. Another alternative would be to use prior information about the attributes to present concepts in each choice set that are more balanced and from which more information about trade-offs respondents make can be learned.

The latter approach was the one tested. Two pairs of numeric attributes were selected based on the fact that their levels were assumed to have an a priori order — a level order that all respondents were assumed to agree went from low to high. Corner prohibitions were specified such that the lowest (worst) level of one attribute in a pair could never appear with the lowest (worst) level of the other attribute in the pair. The same approach applied to the best levels.

In specifying these corner prohibitions, it was assumed that they would reduce the chance of obvious best choices and obvious worst choices from appearing in the choice sets. It also assumed that choice sets would contain concepts with greater utility balance than had these

prohibitions not been specified. Respondents participated in a discrete choice modeling exercise and were randomly assigned to one of three sample cells: (a) no prohibitions, (b) mild corner prohibitions, and (c) severe corner prohibitions.

RESEARCH METHODOLOGY

Respondents were recruited from Survey Sampling, Inc.'s online consumer panel. 2,000 respondents completed the online questionnaire and discrete choice modeling experiment, most of whom came through within the first week. An encrypted demographic string was appended to each respondent's invitation and stored in the final dataset that contained detailed demographic information about each respondent.

Sawtooth Software's CBC System for Choice-Based Conjoint was used to design the discrete choice experiment. There were six different designs (graphics or text, crossed by no prohibitions, mild prohibitions, or severe prohibitions) containing 16 random tasks (Sawtooth Software's complete enumeration design strategy) and 4 holdout tasks which were shown twice to enable measures of test-retest reliability. Each task included four full-profile concepts and there was no none option. Each respondent completed one of 100 pencil & paper versions which had been pre-generated for each design.

Critical Mix programmed and hosted the questionnaire. While there were 14,400 total choice sets (6 design cells x 100 versions x 24 tasks), all that was needed was a program that could extract and then display the appropriate choice set from the design files. The Sawtooth Software Hierarchical Bayes module was used to generate the final set of individual utilities with a priori constraints specified.

In addition to the discrete choice exercise, respondents were required to answer several segmentation questions, rate each brand on a 0 to 10 scale, and allocate 100 points among brand and the other four numeric attributes. At the end of the interview, respondents were asked about their satisfaction with the online survey experience, the time required of them, and the incentive being offered. Respondents were also given the opportunity to leave verbatim comments regarding the survey itself and the topics and issues covered.

After screening respondents, we were left with the following sample size by cell:

Total Respondents After Screening 1491	Text	Graphics
No Corner Prohibitions	T1 n=253	G1 n=251
Mild Corner Prohibitions	T2 n=248	G2 n=255
Severe Corner Prohibitions	T3 n=245	G3 n=239

METRICS

We looked for statistically significant differences between sample cells according to the following metrics:

- Time-to-complete DCM exercise by cell

- Holdout task consistency by cell
- Respondent satisfaction by cell
- MAEs and Hit Rates by cell
- Attribute Importance by cell
- Brand Preference by cell
- Utilities and Standard Errors by cell

There were no significant differences by cell in the time it took respondents to complete the discrete choice questions, no significant differences in the consistency with which respondents answered the repeat holdout tasks, and no significant differences in satisfaction ratings.

Exhibits 1 and 2 show the results of the Mean Absolute Error and Hit Rate tests by cell. Individual-level utility models performed better than the aggregate model. Constrained HB utilities — constrained to match the a priori assumptions made about the levels of each numeric — were used in the final model.

Exhibit 1
MAE (Share Prediction Error)

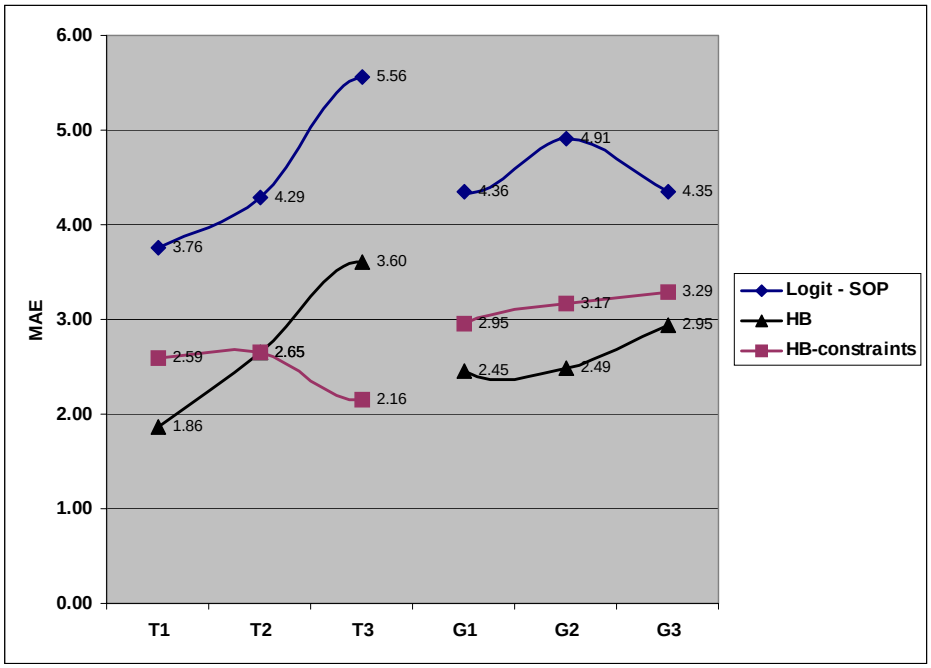


Exhibit 2
Hit Rate (Individual Prediction Accuracy)

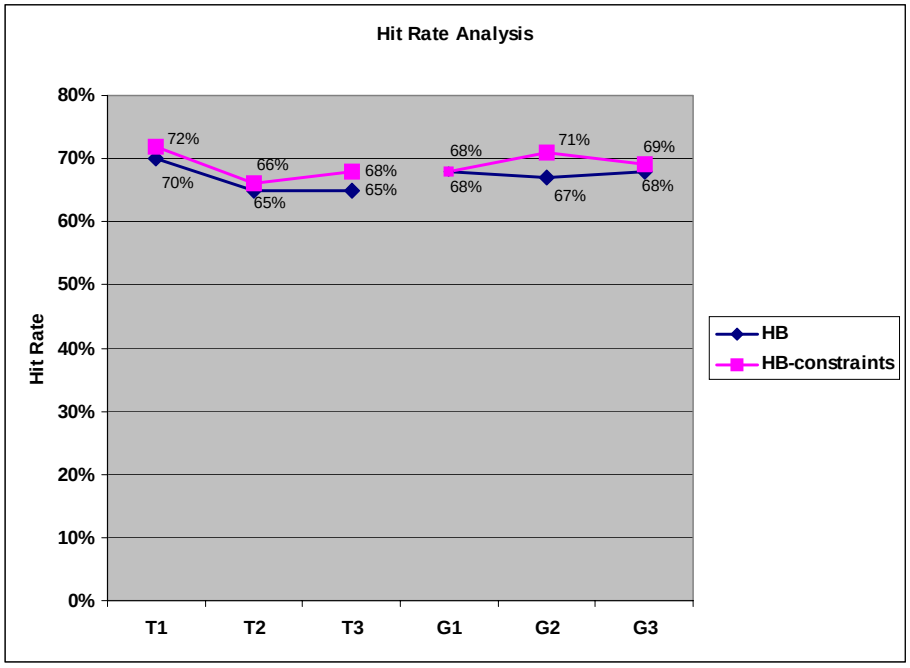


Exhibit 3 shows attribute importances by cell. 95% confidence intervals are represented by the vertical lines. While there are no significant differences in attribute importance by cell, there are significant differences in the attribute importances revealed by the discrete choice analysis and those indicated by respondents in the self-explicated 100-point allocation question.

Exhibit 3

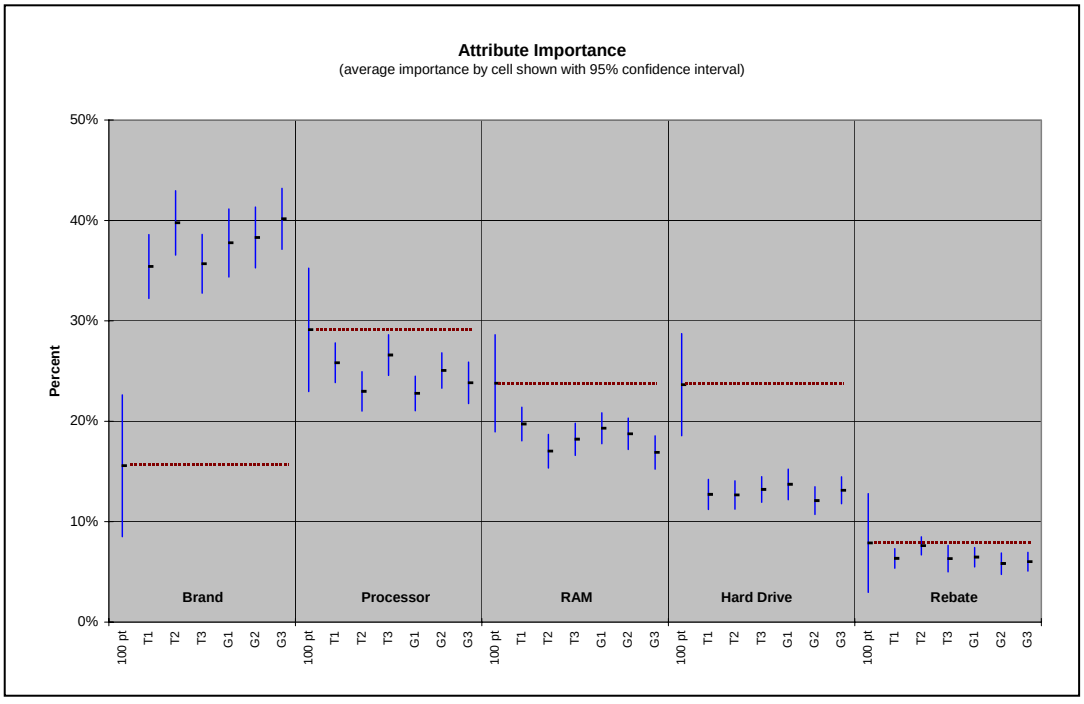
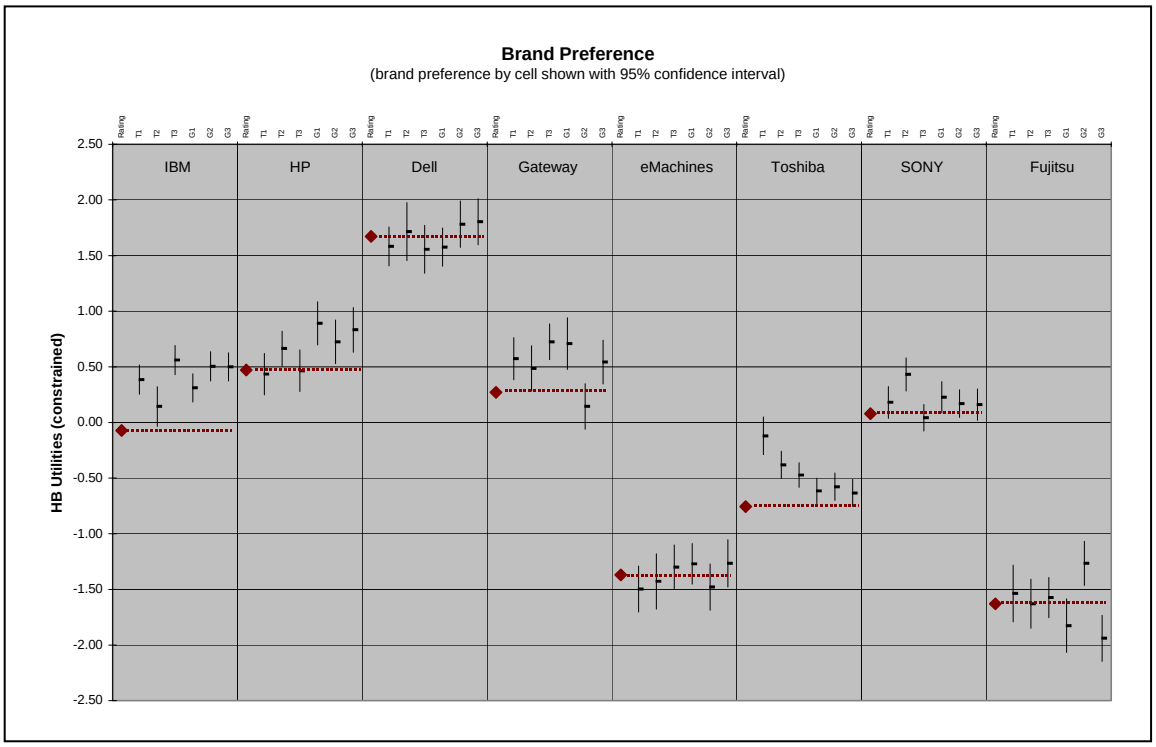


Exhibit 4 shows brand preferences by cell. While there are some significant differences between design cells, they are few and inconsistent. What's more interesting is to compare the brand preferences revealed by the discrete choice analysis to those indicated in the self-explicated 10-point rating scale question. The rescaled brand ratings are almost exactly the same as those revealed from the discrete choice analysis!

Exhibit 4



A review of the part worth utilities and standard errors for all levels of all attributes reveal occasional differences between cells, but they are few and inconsistent. A chart containing this information can be found in the Appendix.

VERBATIM RESPONSES

At the end of the online questionnaire, approximately 20% of respondents took the extra time to leave comments about the survey experience and the topics and issues covered. Following are selected verbatim responses representing the recurring themes found in these comments:

- What I realized in doing this survey is that brand name is more important than I originally realized and that very much influenced my choices
- It surprised me that I didn't always choose my first choice for brand and rather chose the configuration of the computer without being overly observant of the brand.
- TOO many screens! Half would have been sufficient! It was nice to give the bit of encouragement in the middle, though.
- The tables are a bear... Nice move to break them up with a thank you, though.

- I found the experience somewhat satisfying. However, I found myself trying to understand the rationale behind the questions and was concerned this might adversely affect my participation.

CONCLUSIONS

The techniques used in this experiment proved very robust: the use of graphics and forced utility balance (corner prohibitions) had very little or no consistent or significant effect on respondent participation, attribute importance, relative preference for levels, or predictive ability of models.

There were very different attribute importances, however, between the self-explicated 100-point allocation question and those revealed by the discrete choice experiment. The discrete choice exercise revealed that brand was much more important to respondents and that there was more stratification among the remaining numeric attributes. But within the brand attribute, the preference for brand levels was strikingly similar when comparing the results from a self-explicated rating scale to the brand utilities revealed by the discrete choice experiment.

Finally, verbatim responses indicated that 24 tasks were too many but that respondents appreciated the "rest stop" in between. Overall, the online consumer panel proved both efficient and effective at recruiting a large number of respondents to complete a complex marketing research exercise.

APPENDIX

Utilities, Standard Errors, and t-values among experimental groups:

	Utilities-Text	Utilities-Graphics	SE-Text	SE-Graphics	t-value	Utilities-No Prohibitions	Utilities-Severe Prohibitions	SE-No Prohibitions	SE-Severe Prohibitions	t-value	Utilities-Mild Prohibitions	Utilities-Severe Prohibitions	SE-Mild Prohibitions	SE-Severe Prohibitions	t-value	Utilities-No Prohibitions	Utilities-Mild Prohibitions	SE-No Prohibitions	SE-Mild Prohibitions	t-value
IBM	**15**	**17**	1.73	1.60	-0.63	**15**	**17**	1.84	1.95	-0.90	**15**	**17**	2.29	1.95	-0.74	**15**	**15**	1.84	2.29	-0.07
HP	**22**	**33**	2.15	2.47	**-3.38**	**26**	**27**	2.70	2.98	-0.41	**30**	**27**	2.85	2.98	0.53	**26**	**30**	2.70	2.85	-0.98
Dell	**59**	**68**	2.70	2.65	**-2.22**	**60**	**66**	2.92	3.41	-1.48	**64**	**66**	3.48	3.41	-0.34	**60**	**64**	2.92	3.48	-1.09
Gateway	**21**	**18**	2.13	2.44	0.87	**24**	**22**	2.85	2.61	0.38	**12**	**22**	2.90	2.61	**-2.69**	**24**	**12**	2.85	2.90	**2.95**
eMachines	**-47**	**-52**	2.54	2.44	1.38	**-46**	**-47**	2.66	3.07	0.13	**-56**	**-47**	3.37	3.07	-1.94	**-46**	**-56**	2.66	3.37	**2.18**
Toshiba	**-14**	**-23**	1.58	1.46	**4.36**	**-16**	**-21**	2.01	1.73	1.77	**-19**	**-21**	1.85	1.73	0.71	**-16**	**-19**	2.01	1.85	1.06
SONY	**7**	**7**	1.60	1.62	-0.04	**5**	**3**	1.94	1.88	0.96	**11**	**3**	2.07	1.88	**3.07**	**5**	**11**	1.94	2.07	**-2.11**
Fujitsu	**-63**	**-66**	2.71	2.76	1.02	**-67**	**-69**	3.63	3.14	0.29	**-58**	**-69**	3.22	3.14	**2.37**	**-67**	**-58**	3.63	3.22	-1.91
1.0 GHz Processor	**-70**	**-67**	1.52	1.46	-1.53	**-67**	**-70**	1.70	1.95	1.08	**-68**	**-70**	1.82	1.95	0.70	**-67**	**-68**	1.70	1.82	0.37
1.4 GHz Processor	**-14**	**-13**	0.72	0.72	-0.48	**-16**	**-12**	0.81	0.93	**-3.09**	**-12**	**-12**	0.89	0.93	0.01	**-16**	**-12**	0.81	0.89	**-3.17**
1.8 GHz Processor	**27**	**27**	0.75	0.84	0.20	**27**	**26**	0.87	1.02	1.13	**27**	**26**	1.03	1.02	1.22	**27**	**27**	0.87	1.03	-0.18
2.2 GHz Processor	**57**	**53**	1.49	1.36	1.73	**56**	**56**	1.72	1.90	-0.20	**53**	**56**	1.64	1.90	-1.45	**56**	**53**	1.72	1.64	1.32
128 MB RAM	**-56**	**-56**	1.38	1.33	-0.02	**-59**	**-54**	1.59	1.70	-1.90	**-54**	**-54**	1.69	1.70	-0.08	**-59**	**-54**	1.59	1.69	-1.83
256 MB RAM	**1**	**1**	0.49	0.55	0.99	**0**	**1**	0.57	0.67	-1.47	**1**	**1**	0.66	0.67	-0.37	**0**	**1**	0.57	0.66	-1.09
384 MB RAM	**19**	**19**	0.61	0.60	-0.55	**19**	**19**	0.69	0.77	-0.35	**19**	**19**	0.76	0.77	-0.72	**19**	**19**	0.69	0.76	0.41
512 MB RAM	**36**	**36**	1.13	1.09	-0.14	**39**	**33**	1.36	1.39	**3.13**	**35**	**33**	1.32	1.39	0.69	**39**	**35**	1.36	1.32	**2.51**
20 GB Hard Drive	**-39**	**-39**	1.22	1.19	-0.35	**-41**	**-39**	1.59	1.40	-0.76	**-37**	**-39**	1.42	1.40	1.19	**-41**	**-37**	1.59	1.42	-1.86
30 GB Hard Drive	**-1**	**-3**	0.36	0.50	**2.51**	**0**	**-3**	0.47	0.58	**3.56**	**-4**	**-3**	0.54	0.58	-1.20	**0**	**-4**	0.47	0.54	**5.07**
40 GB Hard Drive	**16**	**16**	0.55	0.58	-0.63	**16**	**16**	0.71	0.68	0.00	**16**	**16**	0.67	0.68	-0.15	**16**	**16**	0.71	0.67	0.14
50 GB Hard Drive	**25**	**25**	0.87	0.93	-0.35	**25**	**26**	1.16	1.04	-0.69	**25**	**26**	1.11	1.04	-0.84	**25**	**25**	1.16	1.11	0.13
No rebate	**-20**	**-19**	0.87	0.86	-1.16	**-20**	**-18**	1.02	1.14	-1.37	**-20**	**-18**	1.02	1.14	-1.40	**-20**	**-20**	1.02	1.02	0.03
$50 mail-in rebate	**0**	**-1**	0.32	0.30	1.33	**0**	**-1**	0.41	0.39	1.59	**-1**	**-1**	0.34	0.39	1.30	**0**	**-1**	0.41	0.34	0.41
$100 mail-in rebate	**7**	**7**	0.40	0.40	-0.45	**8**	**6**	0.47	0.55	**2.01**	**7**	**6**	0.45	0.55	1.29	**8**	**7**	0.47	0.45	0.84
$150 mail-in rebate	**14**	**12**	0.73	0.68	1.10	**13**	**13**	0.84	0.96	-0.20	**13**	**13**	0.78	0.96	0.44	**13**	**13**	0.84	0.78	-0.69

COMMENT ON GOGLIA

ROBERT A. HART, JR.
GELB CONSULTING GROUP, INC.

Chris Goglia sets out to determine if the design of a choice-based conjoint study, specifically with respect to imposing corner restrictions on implausible combinations of alternatives and the use of logos rather than just names for brand, can produce better (or even different results). He also compares self-explicated ratings to the inferred ratings for features derived from CBC part-worth estimates.

With respect to study optimization, Chris finds that corner restrictions and the use of logos versus written names had no impact on attribute importance levels, relative valuation of feature levels, or on the predictive power of the models.

For corner restrictions, this finding makes absolute intuitive sense, and is actually a very powerful finding. Given that the power of conjoint analysis is in its ability to tell us how various features, and combinations of features, are valued through respondents' empirical choices rather than any self-explicated means OR by a priori imposing structure on the data, imposing corner restrictions is just another, albeit less pernicious, means of imposing our biases and values on the data, where none is needed.

If there truly are implausible combinations, then the data will tell us that empirically. And Chris' finding that imposing these restrictions did not improve model fit with reality is comforting, and also additional ammunition in favor of proceeding with a *tabula rasa* design strategy.

For the finding that logos do not affect valuation or choice, I find this a bit more curious, or at least did at first. What this tells us, in essence, is that in the fairly structured format of CBC, the logotype does not convey any new or compelling information to the respondent than does the brand name written out. From an ease of design perspective, CBC researchers should consider this good news.

But, it also should be noted that there may be other uses for graphical images in a CBC that WOULD convey information, such as testing competing labels of packaged goods, or showing a photo of an automobile rather than just listing the make, model and color. Testing this would provide an interesting extension of this research.

With respect to the self-explicated versus CBC results, this should be most exciting to CBC researchers. Chris finds that self-explicated ratings produce drastically different results than empirical conjoint part-worths. We've all been making the case for some time that self-explication is, at best, unreliable and less accurate than conjoint analysis. Chris' findings, especially with respect to brand, indicate that self-explicated ratings are much worse and downright misleading. Self-explicated ratings, in his study, reduce brand to a trivial product attribute. Conjoint part-worths indicate that brand is the single most important driver of product choice. This finding should be kept in the front pocket of every market research manager and market research professional who finds themselves faced with a client (or internal client) who does not believe that there are advantages to conducting conjoint analysis.

How Few is Too Few?: Sample Size in Discrete Choice Analysis

Robert A. Hart, Jr.
Gelb Consulting Group, Inc.
Michael Patterson
Probit Research

Introduction

One of the stiffest challenges facing market researchers is to balance the need for larger samples with the practical need to keep project budgets manageable. With discrete choice analysis (choice-based conjoint), one option available is to increase the number of choice tasks each respondent completes to effectively increase the sample without increasing the number of respondents (Johnson and Orme 1996; Orme 1998).
A concern is that there may be some minimum sampling threshold below which discrete choice estimates become unstable and the technique breaks down.

We address this concern in two ways. First, we conduct Monte-Carlo experiments with known population parameters to estimate the impact of reducing the sample of respondents (even while maintaining the size of the data used to estimate attribute coefficients by expanding the number of choice tasks) on the accuracy of and variance around part-worth estimates. We then conduct an online discrete-choice study to determine if our experimental findings match empirical realities.

Our goal is to refine existing heuristics regarding respondent sample size in discrete choice surveys. Thus this research will provide guidance to conference attendees concerning trade-offs between sample size and the number of choice tasks included in a study design.

Sample Size and Discrete Choice

Gaining adequate sample is one of the fundamental challenges to conducting sound market research. The cost of obtaining sample can double or triple the cost of conducting a piece of research, so we are constantly faced with getting "just enough" sample to be able to draw actionable business conclusions while not simultaneously reducing a year's budget to nothing. In addition to cost, though, are other constraints. Some questions are directed at a segment of the population whose numbers are small, and these "low-incidence" samples may only be able to produce a small number of potential respondents. In addition, timing can be a factor, whereas even with no cost or incidence constraints, it may be the case that there simply isn't the time necessary to develop sample with which the researcher would normally be comfortable.

Regardless of the reason, Johnson and Orme (1996) and Orme (1998) both present findings which suggest that respondents and choice tasks are, within reason, interchangeable sources of additional data. This finding is even more robust and important in light of the findings that the reliability of choice-based conjoint results

actually *increase* for "at least the first 20 tasks." (Johnson and Orme 1996:7) Thus, for a study that had initially only included ten choice tasks, the researcher could reliably double the sample of data by doubling the number of choice-tasks, rather than the more prohibitive method of doubling the number of respondents.

MAXIMUM-LIKELIHOOD IN SMALL SAMPLES

Maximum-Likelihood Estimation (MLE) is a robust estimation method whose asymptotic properties are much less constrained and assumption dependent than ordinary least squares. Yet little empirical research has been conducted to identify the small-sample properties of MLEs, which are not defined mathematically like the asymptotic properties.

One study conducted a series of Monte-Carlo simulations to observe the behavior of MLEs in a very simple, controlled environment (Hart and Clark 1997). This study generated normally distributed, random independent variable and error data and built dependent variable data based on these values. Bivariate probit models were then estimated (only a single independent variable was modeled for this paper) at various sample sizes.

The results of this work suggested that maximum-likelihood estimation can run into problems when the size of the data matrix used for estimation gets small. When $n \leq 30$, even with a single right hand side variable, the incidence of models not converging or converging large distances from the actual population parameters increased dramatically. At a sample size of ten, there was only approximately a 10% chance that the model would successfully converge, much less have the independent variable's coefficient stand up to a hypothesis test.

By contrast, least-squares was incredibly robust in this environment, estimating in all cases and even providing accurate coefficient estimates in the sample of ten. This is, of course, a function of the mathematical ease with which OLS arrives at estimates, and is why problems associated with least-squares are generally classified as problems of inference rather than problems of estimation.

Maximum-likelihood, for all of its desirable asymptotic properties, remains an information intensive estimation procedure, asking the data, in essence, to reveal itself. When the data is insufficient to perform this task, it does not merely converge with the blissful ignorance of its least-squares cousin, but often exhaustively searches for its missing maximum in vain.

Given that choice-based conjoint estimates part-worth utilities via maximum-likelihood multinomial logit estimation, it is our concern that CBC may experience some of the problems observed in the general probit scenario described above. Now, it may be the case that since increasing the number of choice tasks a la Johnson and Orme increases the actual size of the data matrix used for estimation that these problems will disappear. On the other hand, since the additional data is really only additional information about *a single respondent*, the lack of information problem may occur in this environment as well.

STUDY DESIGN

To address this issue, and to investigate the small respondent sample properties of CBC, we conduct two phases of research. In the first we run Monte Carlo simulations where fictitious respondents are created according to some preset standards who then perform a series of choice tasks, and the resulting data are used to estimate part-worth utilities. This is done at a variety of sample-size/choice-task combinations to determine if the MLE problems are present in smaller respondent samples.

The second phase of work utilizes empirical data collected from a study of IT professionals on their preferences over various server features. Subsets of respondents and choice tasks are drawn from the overall sample to determine how decreasing respondents relative to choice tasks behaves using real data, and if the patterns mimic what appears experimentally.

MONTE CARLO SIMULATION

Our first step is to create a sterile, controlled environment and determine if small samples of respondents cause any of the problems discussed above, or if there are other patterns of parameter estimates or the variance around those estimates that are unusual or different from what is expected from sampling theory.

To do so, a SAS program was written to generate groups of fictitious respondents, represented by utilities for hypothetical product features. The respondents were generated by randomly composing their utilities for each choice by additively combining the population utility parameters and adding a normally distributed error (~N(0,16)). In addition, when the actual choices for each fictitious respondent were modeled, their choices (which are a direct function of their generated utility) were also given some random error (~Gumbel(-.57,1.65)).

Table One presents the population attribute part-worths used to design the experiment.

Table One:
Population Part-Worth Parameters

	Level A	Level B	Level C
Attribute 1	-2	0	2
Attribute 2	-1	0	1
Attribute 3	1	0	-1
Attribute 4	-3	-1	4
Attribute 5	-2	1	1
Attribute 6	1	0	-1
Attribute 7	1	0	-1
Attribute 8	-3	0	3
Attribute 9	-4	0	4
Attribute 10	-5	0	5

To determine the effect of sample size on estimates, four separate sets of experiments were run, as follows:

1000 Respondents, 1 Choice Task

100 Respondents, 10 Choice Tasks

50 Respondents, 20 Choice Tasks

10 Respondents, 100 Choice Tasks

Notice that the design insures that the body of data available for analysis is constant (1000 data "points") such that our focus is really on the potential problems associated with having too much data come from a single respondent (and thus behave more like a single data point). For each set-up, 500 simulations are run, and the part-worth estimates are saved for analysis.

Table Two presents the average part-worth estimates for the A and B Levels for each attribute.

Table Two:
Mean Simulation Part-Worth Estimates

	PP*	1000/1	100/10	50/20	10/100
A1	-2	-2.36	-2.37	-2.38	-2.38
B1	0	0.01	0.01	0.00	0.00
A2	-1	-1.19	-1.18	-1.19	-1.18
B2	0	0.00	-0.01	0.00	-0.02
A3	1	1.18	1.19	1.18	1.19
B3	0	0.00	-0.01	0.00	0.00
A4	-3	-3.54	-3.56	-3.56	-3.58
B4	-1	-1.17	-1.18	-1.19	-1.19
A5	-1	-2.36	-2.36	-2.37	-2.38
B5	1	1.19	1.19	1.18	1.18
A6	1	1.18	1.19	1.18	1.19
B6	0	0.00	0.00	0.00	0.00
A7	1	1.20	1.20	1.20	1.20
B7	0	-0.01	-0.01	0.00	-0.01
A8	-3	-3.54	-3.55	-3.57	-3.57
B8	0	0.00	0.00	0.01	0.01
A9	-4	-4.71	-4.75	-4.76	-4.77
B9	0	-0.01	0.01	0.01	-0.01
A10	-5	-5.90	-5.92	-5.95	-5.96
B10	0	0.00	0.00	0.01	-0.01

The part-worth estimates are consistent and unbiased by sample size (which is of course predicted accurately by sampling theory), and the accuracy of the estimates, in the aggregate is comforting. Even a sample of ten respondents, given a large enough body of choice tasks (ignoring the effects reality would have on our robo-respondents), will, on average, produce part-worth estimates that are reflective of the population parameters.

Table Three presents the variance around those part-worth estimates at each sample-size/choice task combination.

This table, though, indicates that any given small sample runs the risk of being individually further away from the population, but this finding is not at all surprising and is once again consistent with sampling theory. In fact, what is most remarkable is that the variance around those estimates (when n=10) is not that much greater than for the much larger body of respondents.

Table Three:
Variance around Simulation Part-Worth Estimates

	PP*	1000/1	100/10	50/20	10/100
A1	-2	0.07	0.08	0.09	0.15
B1	0	0.06	0.06	0.06	0.07
A2	-1	0.07	0.07	0.07	0.09
B2	0	0.06	0.06	0.06	0.06
A3	1	0.06	0.06	0.07	0.09
B3	0	0.06	0.06	0.06	0.06
A4	-3	0.08	0.09	0.12	0.22
B4	-1	0.07	0.07	0.07	0.09
A5	-1	0.07	0.08	0.09	0.15
B5	1	0.06	0.06	0.08	0.09
A6	1	0.06	0.07	0.07	0.09
B6	0	0.06	0.06	0.06	0.06
A7	1	0.06	0.07	0.07	0.09
B7	0	0.06	0.06	0.06	0.07
A8	-3	0.07	0.09	0.12	0.22
B8	0	0.06	0.06	0.06	0.07
A9	-4	0.09	0.11	0.15	0.28
B9	0	0.06	0.06	0.06	0.07
A10	-5	0.10	0.13	0.18	0.34
B10	0	0.07	0.07	0.06	0.07

Most apparent, when looking at the individual part-worth estimate output, is that sample size did *not* wreak havoc on the MLE behind the scenes, and there is no evidence of estimation problems, even with a sample of ten respondents.

One issue not addressed in this analysis is the issue of heterogeneity. To see if heterogeneity is affecting the analysis, a series of simulations were run utilizing hierarchical-Bayes estimation, and these simulations were only run at the 100, 50 and 10 respondent levels.

Table Four shows the HB part-worth estimates. The magnitude of the coefficients are muted toward zero compared to their general logit counterparts. Most interesting is the fact that the estimates in the smallest sample size are decidedly different than the larger sample estimates.

Table Four:
Mean HB Part Worth Estimates

	PP*	100/10	50/20	10/100
A1	-2	-1.17	-1.25	-1.76
B1	0	0.02	-0.01	-0.01
A2	-1	-0.53	-0.63	-0.91
B2	0	-0.02	0.04	0.06
A3	1	0.59	0.67	0.90
B3	0	0.03	0.00	0.08
A4	-3	-1.72	-1.82	-2.72
B4	-1	-0.60	-0.64	-0.86
A5	-1	-1.12	-1.19	-1.79
B5	1	0.58	0.60	0.96
A6	1	0.61	0.63	0.86
B6	0	-0.02	-0.03	0.03
A7	1	0.60	0.65	0.92
B7	0	0.01	0.03	-0.04
A8	-3	-1.73	-1.84	-2.58
B8	0	0.00	-0.03	-0.12
A9	-4	-2.32	-2.50	-3.67
B9	0	-0.02	0.01	0.10
A10	-5	-2.89	-3.11	-4.52
B10	0	-0.03	-0.01	0.01

Even more alarming is the fact that the estimates for the smallest sample are in actuality the estimates that are closest to the population parameters. Unfortunately, we do not have a sound explanation for this finding, but will continue to explore the reason for its occurrence.

Table Five presents the variance around the hierarchical-Bayes part-worth estimates. The variances tend to be lower than for logit, but this could be due to scale factor. When the sample gets extremely small, though, the variance around the HB estimates gets much bigger. The observation of much greater variance is not terribly surprising, but the fact that this did not happen in the general logit case remains so.

Table Five:
Variance around HB Part-Worth Estimates

	PP*	100/10	50/20	10/100
A1	-2	0.05	0.07	0.57
B1	0	0.03	0.04	0.23
A2	-1	0.04	0.06	0.29
B2	0	0.03	0.05	0.21
A3	1	0.03	0.05	0.31
B3	0	0.03	0.05	0.21
A4	-3	0.06	0.09	0.92
B4	-1	0.03	0.06	0.37
A5	-1	0.06	0.07	0.39
B5	1	0.04	0.06	0.40
A6	1	0.04	0.05	0.33
B6	0	0.03	0.04	0.32
A7	1	0.04	0.05	0.29
B7	0	0.03	0.05	0.29
A8	-3	0.06	0.14	0.75
B8	0	0.04	0.06	0.27
A9	-4	0.11	0.16	1.28
B9	0	0.05	0.05	0.30
A10	-5	0.10	0.20	1.95
B10	0	0.04	0.04	0.24

EMPIRICAL DATA ANALYSIS

In an effort to validate some of the findings of the simulations, an online conjoint study was designed and fielded. IT professionals were recruited to our internally hosted CBC survey which was designed and programmed using Sawtooth Software's SSI Web software. Respondents were shown 20 choice tasks comprised of various industrial server attributes.

Table Six lists the attributes and their levels.

Table Six:
Empirical Study Attributes and Levels

Attribute	Level A	Level B	Level C
Number of processors capable	One	Two	Four
Battery-backed write cache	None	Optional	Standard
CD/Floppy availability	CD-Rom	Floppy	Both
Max number of hard drives	One	Two	Six
NIC type	1-10/100 NIC	2-10/100 NIC	2-10/100/1000 NIC
Processor type	Intel Xeon Processor	Intel Pentium III Processor	AMD Athlon Processor
Standard memory (MB)	256 MB RAM	512 MB RAM	1024 MB RAM
System Bus speed (MHz)	133 MHz	400 MHz	533 MHz
Hard drive type	ATA Hard Drive	Standard SCSI Hard Drive	High Performance SCSI RAID Hard Drive
Price	$3000	$5000	$7000

209 respondents were gathered over a several-day period. To determine the effects of sample size in this environment, many sub-samples of respondents and choice tasks were selected from the overall body of 209 respondents and 20 choice tasks to mirror, as best as possible, the structure of the experimental work.

Specifically, in Table Seven we report the part-worth coefficient estimates and variances for the first level of each variable (for simplicity only). The first two columns report the overall betas, and the second is an average, for 200 respondents, for multiple combinations of five choice tasks (such that there are 1000 data points). The final two columns report the estimates and variances for 100 respondents and 10 choice tasks and 50 respondents and 20 choice tasks.

Table Seven:
Empirical Part-Worth Estimates and Variances

209 Resp, 20 CTs	200 Resp. 5 CTs	100 Resp. 10 CTs		50 Resp. 20 CTs	
Beta	Beta	Beta	**Variance (100 runs)**	Beta	**Variance (100 runs)**
-0.37666	-0.37710	-0.38757	**0.00285**	-0.38674	**0.00799**
-0.18387	-0.17110	-0.23262	**0.00239**	-0.19010	**0.00477**
-0.28915	-0.31265	-0.35712	**0.00261**	-0.28848	**0.00596**
-0.15122	-0.19246	-0.17055	**0.00291**	-0.15492	**0.00349**
-0.04358	-0.04673	0.03677	**0.00224**	-0.04067	**0.00433**
0.37774	0.36652	0.35572	**0.00278**	0.38509	**0.00527**
-0.15353	-0.16358	-0.15466	**0.00244**	-0.16002	**0.00486**
-0.14144	-0.14570	-0.10562	**0.00246**	-0.14066	**0.00326**
-0.55779	-0.56611	-0.57577	**0.00435**	-0.56563	**0.00818**
0.17774	0.19742	0.17791	**0.00269**	0.18146	**0.00709**

The average part-worth estimates are consistent across sample sizes, which jibes with expectations and the experimental work. The variance around the estimates for 50 respondents is about twice that for 100 respondents, but still not beyond what is expected nor evidence of any estimation-related problems.

CONCLUSION

The end result of this first-cut research is a very good-news story indeed: the estimation issues MLE exhibits in other areas do not appear to affect the multinomial-logit component of a CBC study, provided there are ample data points created via more choice tasks. Although there are still inference issues in small samples, we have sampling theory to tell us how accurate our estimates are in those cases.

Researchers faced with hard to reach target populations can take some comfort that even a sample of 100 or possibly less can provide unbiased estimates of population preferences, although there will be ever widening margins of error around those estimates.

REFERENCES

Hart, Robert A., Jr., and David H. Clark. 1997. “Does Size Matter?: Maximum Likelihood in Small Samples.” Paper presented at the Annual Meeting of the Midwest Political Science Association. Chicago.

Johnson, Richard M. and Bryan K. Orme. 1996. “How Many Questions Should You Ask in Choice-Based Conjoint Studies?” *Sawtooth Software Research Paper Series.*

Orme, Bryan K. 1998. “Sample Size Issues for Conjoint Analysis Studies” *Sawtooth Software Research Paper Series.*

CBC VALIDATION

VALIDATION AND CALIBRATION OF CHOICE-BASED CONJOINT FOR PRICING RESEARCH

GREG ROGERS
PROCTER & GAMBLE
TIM RENKEN
COULTER/RENKEN

INTRODUCTION

Choice-based conjoint has been used for many years to measure price sensitivity, but until now we have only had a handful of in-market validation cases. Much of the efforts to date have focused on comparing 'base case' scenarios from conjoint simulators to market shares or holdout tasks. Consequently, they do not necessarily speak to the accuracy of the measured price sensitivity. We have compared the price sensitivity results of CBC versus the estimated sensitivity from econometric data (Marketing Mix Models).

We pursued this work for two reasons:

1. Assess the accuracy of CBC for pricing research.
2. Explore calibration methods to improve the accuracy.

Comparing the parallel studies highlighted the opportunity to make the correlation stronger by calibrating the CBC results. Two approaches were attempted to calibrate the results:

1. Adjustment of the exponential scalar in share of preference simulations. We explored the use of a common scalar value applied across all studies that would increase the correlation.
2. Multiple regression using brand and market data as variables to explain the difference between the CBC and Econometric data. Variables used were: unit share, distribution, % volume sold on deal, # of items on shelf, and % of category represented.

DESIGN AND DATA COLLECTION

As *Table 1* indicates, there were a total of 18 comparisons made between the econometric and CBC data. The time lag between the collection of the econometric data and CBC data is cause for concern, even though an item's price sensitivity tends not to change too drastically from one year to the next. Also worth noting is the coverage of the econometric data, estimated at about 80% of category sales for the outlets the models were based on.

All CBC studies used a representative sample of respondents with each respondent having purchased the test category at least once in the past 12 months. Choice tasks were

generated using complete enumeration, with conditional pricing used in each study. No prohibitions were included.

The econometric data, or marketing mix model, was constructed using multiple regression. Models for each outlet were used to estimate weekly target volume as a function of: each modelled and competitive SKU's price, merchandising (feature, display, feature & display, and temporary price reduction), advertising and other marketing activities. The model controls for cross-store variation, seasonality and trend.

Table 1

	Washing Powder (USA)	Salted Snacks (USA)	Facial Tissue (USA)	Potato Crisps (UK)
Data collection method	Central Site	Central Site	Central Site	Central Site
Base size	650	600	620	567
# of items in study	15	15	15	10
# of choice tasks in study	14	14	14	12
CBC data collection dates	May 2002	May 2002	June 2002	April 2002
Time period of Econometric data	52 wks ending Aug 2001	52 wks ending Sept 2001	52 wks ending Aug 2001	104 wks ending Aug 2001
Econometric data	Food + Mass	Food + Mass	Food + Mass	Top 5 Grocers
# of parallel cases (Econometric vs CBC)	3	3	5	7

COMPARING UNCALIBRATED CBC DATA TO ECONOMETRIC DATA

The choice model estimation was made using CBC/HB with constraints on the price utility. A share of preference model with respondent weighting by purchase frequency was used as the simulation method.

For comparison purposes the data were analysed at the +10% price increase and -10% price decrease levels. This means of the 18 cases of comparison we have 2 price levels to compare for each case, or a total of 36 comparison points across the entire data set.

Comparing the CBC results to the econometric data shows that CBC over estimates the price sensitivity, but far more so for price decreases than price increases. Table 2 provides a summary of the comparisons.

Table 2

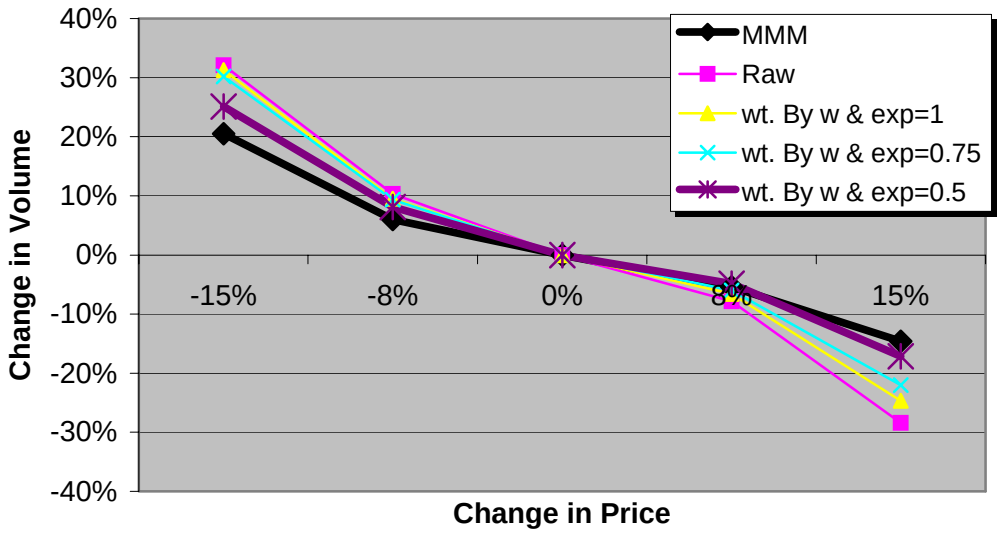
		No Calibration
ALL DATA		
	Mean	31%
	MAE	8.8%
PRICE INCREASES		
	Mean	12%
	MAE	5.1%
PRICE DECREASES		
	Mean	51%
	MAE	12.4%

CALIBRATING CBC DATA TO ECONOMETRIC DATA USING AN EXPONENTIAL SCALAR

The CBC data were modelled using the same methods described previously except we now introduce a scaling factor in the share of preference simulation model (1).

$$P(A) = \frac{e^{(U_{b_1} + U_p)\times 0.5}}{\sum_{b_1}^{b_n} e^{(U_{b_1} + U_p)\times 0.5}} \qquad (1)$$

This scaling factor will have the effect of dampening the differences between the utilities as shown in *Figure 1* (note "w" refers to purchase frequency).



The scaling factor helps to fit the CBC data to the econometric data much better, but the global scaling factor across all price levels results in the calibrated CBC data now being too insensitive to price increases whilst still being too sensitive to price decreases. This is shown in *Table 3*.

Table 3

		No Calibration	Scalar Calibration
ALL DATA			
	Mean	31%	-5%
	MAE	8.8%	5.6%
PRICE INCREASES			
	Mean	12%	-16%
	MAE	5.1%	4.1%
PRICE DECREASES			
	Mean	51%	7%
	MAE	12.4%	7.2%

It is also important to note that by making the calibration in this way the share of choices for the 'base case' (all parameters at their current market condition) move closer to market shares (*Table 4*).

Table 4

	Exp=1.0	Exp=0.5
MAE	5.1%	4.5%

CALIBRATING CBC DATA TO ECONOMETRIC DATA USING REGRESSION

The goal of this method is to calibrate by identifying systematic differences between econometric and conjoint price sensitivities. The dependent variable is the difference between the econometric and conjoint price sensitivities, and the independent variables are: past 52 week unit share, past 52 week distribution, percent volume sold on deal, number of items on shelf, and percent of the category represented on the shelf.

The regression model is described in Equation (2).

$$\frac{\ln\left(\frac{\tilde{v}_i}{\tilde{c}_i}\right)}{\ln(\tilde{p}_i)} = \beta_0 + \beta_1 \text{Share}_i + \beta_2 \text{Distrib}_i + \ldots + \varepsilon_i \qquad (2)$$

where $\tilde{p}_i \neq 1.0$ and

$\tilde{v}_i$ = volume (from scanner data) of product *i*
$\tilde{c}_i$ = choice share (from choice data) of product *i*
$\tilde{p}_i$ = price of product *i*.

Data are normalized such that at base prices $\tilde{v}_i = \tilde{c}_i = \tilde{p}_i = 1.0$. Thus, a 15% price increase over base price would enter the model as $\tilde{p}_i = 1.15$.
Please see Appendix A for derivations.

The output of the regression is shown in Table 5. It suggests that CBC under estimates the price sensitivity of larger share items, over estimates the price sensitivity of

items that sell a lot on deal, and over estimates the price sensitivity in experiments with few items on the shelf.

Table 5

Parameter	Regression Parameter Estimates	St Error	Pr > |t|
Intercept	2.024	0.582	0.0008
Past 52-week Unit Share	-14.581	5.943	0.0161
Past 52-week Distribution	0.190	0.601	0.7522
% of Volume Sold on Deal	1.542	0.532	0.0047
# of Items on Shelf	-0.097	0.011	0.0001
% of Category Represented	-1.844	1.361	0.1788

Sample Size = 96
RSquare = 0.479

CONCLUSIONS

This work has demonstrated that estimates of price sensitivity from CBC can be greatly improved by calibration. Interestingly, the relatively simple method of using an exponential scalar resulted in a similar improvement as the regression based calibration. The overall results are shown in Table 6.

Table 6

	No Calibration	Scalar Calibration	Regression Calibration
ALL DATA			
MAE	8.8	5.6	5.7
MAPE	69%	44%	48%
PRICE INCREASES			
MAE	5.1	4.1	4.1
MAPE	46%	36%	38%
PRICE DECREASES			
MAE	12.4	7.2	7.4
MAPE	91%	53%	57%

The scalar method has the advantage of easy implementation by practitioners, while the regression method helps to identify systematic divergence from econometric data. For this reason, there is more for the community at large to learn from the regression method, and further work would be best pursued in this area.

REFERENCES

Elrod, T. (2001). "Recommendations for Validations of Choice Models", In *Proceedings of the Sawtooth Software Conference* (No. 9, pp. 225-243). Sequim. WA: Sawtooth Software.

Feurstein, M., Natter, M. & Kehl, L. (1999). "Forecasting scanner data by choice-based conjoint models". In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 169-182). Sequim, WA: Sawtooth Software.

Orme, B.K., Alpert, M.I. & Christensen, E. (1997). "Assessing the validity of conjoint analysis — continued". *Research Paper Series,* Sawtooth Software, Sequim, WA.

Orme, B.K. & Heft, M.A. (1999). "Predicting actual sales with CBC: How capturing heterogeneity improves results". In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 183-200). Sequim, WA: Sawtooth Software.

Pinnell, J. & Olsen, P. (1993). "Using choice-based conjoint to assess brand strength and price sensitivity", *Sawtooth News* (No. 9 (3), pp. 4-5). Evanston, IL: Sawtooth Software.

APPENDIX A
DERIVATION OF REGRESSION BASED CALIBRATION

Assume that we can adequately represent the relationship between volume, as measured by scanner data, and price using a function of the form

$$v_i = \lambda_{0i} p_i^{\lambda_{1i}} \quad (1)$$

Assume we can adequately represent the relationship between choice share, as measured by hierarchical Bayes choice models, and price using a function of the form

$$c_i = \theta_{0i} p_i^{\theta_{1i}} \quad (2)$$

where v_i is the volume per week for product i, p_i is the price of product i, c_i is the choice share for product i, and λ_{0i}, λ_{1i}, θ_{0i}, and θ_{1i} are parameters. If we let $\bar{p}_i$ be the base price for product i and $\bar{v}_i$ and $\bar{c}_i$ be the associated volume and choice share for this base price, then we can normalize Equations (1) and (2) to read

$$\tilde{v}_i = \tilde{p}_i^{\lambda_{1i}} \quad (3)$$

and

$$\tilde{c}_i = \tilde{p}_i^{\theta_{1i}} \quad (4)$$

where $\tilde{v}_i = v_i / \bar{v}_i$ and $\tilde{c}_i = c_i / \bar{c}_i$. Dividing Equation (3) by Equation (4) gives us

$$\frac{\tilde{v}_i}{\tilde{c}_i} = \tilde{p}_i^{\lambda_{1i} - \theta_{1i}} \quad (5)$$

Let $\Omega_i = \lambda_{1i} - \theta_{1i}$. Ω_i then represents the difference between price elasticities derived from scanner data and those derived from choice data. We might speculate that differences between these elasticities may be systematic, that for example big brands (as measured by unit share) have lower price sensitivities in the choice exercise than in the marketplace. We thus let Ω_i be a function of explanatory variables such as share, distribution, etc:

$$\Omega_i = \beta_0 + \beta_1 \text{Share}_i + \beta_2 \text{Distrib}_i + .. + \varepsilon_i \quad (6)$$

where β_0, β_1, and β_2 are parameters, and ε_i is a normally distributed error term. Taking logs of Equation (5) and rearranging terms yields

$$\frac{\ln\left(\frac{\tilde{v}_i}{\tilde{c}_i}\right)}{\ln(\tilde{p}_i)} = \beta_0 + \beta_1 \text{Share}_i + \beta_2 \text{Distrib}_i + \ldots + \varepsilon_i \quad (7)$$

With data on the impact on volume $\tilde{v}_i$ and choice share $\tilde{c}_i$ of various price levels $\tilde{p}_i$, together with share, distribution and other information about each product i, we can use Equation (7) to estimate the parameters β_0, β_1, and β_2 with a standard regression model.

DETERMINANTS OF EXTERNAL VALIDITY IN CBC

BJORN ARENOE
SKIM ANALYTICAL

INTRODUCTION

Ever since the early days of conjoint analysis, academic researchers have stressed the need for empirical evidence regarding its external validity (Green and Srinivasan, 1978; Wittink and Cattin, 1989; Green and Srinivasan, 1990). Even today, with traditional conjoint methods almost completely replaced by more advanced techniques (like CBC and ACA), the external validity issue remains largely unresolved. Because conjoint analysis is heavily used by managers and investigated by researchers, external validity is of capital interest (Natter, Feurstein and Kehl, 1999).

According to Natter, Feurstein and Kehl (1999), most studies on the validity and performance of conjoint approaches rely on internal validity measures like holdout samples or Monte Carlo Analysis. Also, a number of studies deal with holdout stimuli as a validity measure. Because these methods focus only on the internal validity of the choice tasks, they are unable to determine the success in predicting actual behaviour or market shares. Several papers have recently enriched the field. First of all, two empirical studies (Orme and Heft, 1999; Natter, Feurstein and Kehl, 1999) investigated the effects of using different estimation methods (i.e. Aggregate Logit, Latent Class and ICE) on market share predictions. Secondly, Golanty (1996) proposed a methodology to correct choice model results for unmet methodological assumptions. Finally, Wittink (2000) provided an extensive paper covering a range of factors that potentially influence the external validity of CBC studies. Although these papers contribute to our understanding of external validity, two blind spots remain. Firstly, the number of empirically investigated CBC studies is limited (three in Orme and Heft, 1999; one in Natter, Feurstein and Kehl, 1999). This lack of information makes generalisations of the findings to 'a population of CBC studies' very difficult. Secondly, no assessment was made of the performance of Hierarchical Bayes or techniques other than estimation methods (i.e. choice models and methodological corrections).

OBJECTIVES

CBC is often concerned with the prediction of market shares. In this context, the external validity of CBC can be defined as the accuracy with which a CBC market simulator predicts these real market shares. The objective of this study is to determine the effects of different CBC techniques on the external validity of CBC. The investigated techniques include three methods to estimate the utility values (Aggregate Logit, Individual Choice Estimation and Hierarchical Bayes), three models to aggregate utilities into predicted respondent choices (First Choice model, Randomised First Choice with only product variability, Randomised First Choice with both product and attribute variability) and two measures to correct for unmet methodological assumptions (weighting respondents by their purchase frequency and weighting estimated product

shares by their distribution levels). A total of ten CBC studies were used to assess the effects of using the different techniques. All studies were conducted by Skim Analytical; a Dutch marketing research company specialised in CBC applications.

MEASURES OF VALIDITY

Experimental research methods can be validated either internally or externally. Internal validity refers to the ability to attribute an observed effect to a specific variable of interest and not to other factors. In the context of CBC, internal validation often refers to the ability of an estimated model to predict other observations (i.e. holdouts) gathered in the same artificial environment (i.e. the interview)[1]. We see that many authors on CBC techniques use internal validation as the criterion for success for new techniques (Johnson, 1997; Huber, Orme and Miller, 1999; Sentis and Li, 2001).

External validity refers to the accuracy with which a research model makes inferences on the real world phenomenon for which it was designed. External validation assesses whether the research findings can be generalized beyond the research sample and interview situation. In the context of CBC, external validation of the SMRT – CBC market simulator provides an answer to the question whether the predicted choice shares of a set of products are in line with the actual market shares. External validity obviously is an important criterion as it can legitimise the use of CBC for marketing decision-making. Very few authors provide external validation of CBC techniques although many do acknowledge its importance. A proposed reason for this lack of evidence is that organisations have no real incentive to publish such results (Orme and Heft, 1999).

External validity of CBC can be assessed by a comparison of predicted market shares with real market shares. One way to do this is to simulate a past market situation and compare the predicted shares with the real shares recorded during that time period. This approach is used in this study and in the two other important papers on external validity (Orme and Heft, 1999; Natter, Feurstein and Kehl, 1999). The degree of similarity in this study is recorded with two different measures: the Pearson correlation coefficient (R) between real and predicted shares and the Mean Absolute Error (MAE) between real and predicted shares.

TECHNIQUES

Three classes of CBC techniques are represented in this study. *Estimation methods* are the methodologies used for estimating utility values from respondent choices. Aggregate Logit estimates one set of utilities for the whole sample, hereby denying the existence of differences in preference structure between respondents. Individual Choice Estimation (ICE) tries to find a preference model for each individual respondent. The first step in ICE is to group respondents into segments (Latent Classes) that are more or less similar in their preference structure. During the second step, individual respondent utilities are calculated as a weighted sum of segment utilities. As ICE acknowledges heterogeneity in consumer preferences it is generally believed to outperform Aggregate Logit. In this study all ICE solutions are based on ten segments. Hierarchical Bayes (HB)

[1] Definitions by courtesy of Dick Wittink.

is another way to acknowledge heterogeneity in consumer preferences. This method tries to build individual preference models directly from respondent choices, replacing low quality individual information by group information if necessary. In general HB is believed to outperform ICE and Aggregate Logit, especially when the amount of choice information per respondent is limited.

Choice models are the methodologies used to transform utilities into predicted respondent choices. The First Choice model (FC) is the simplest way to predict respondent choices. According to this model, every consumer always chooses the product for which he has the highest predicted utility. In contrast, the Randomised First Choice model acknowledges that respondents sometimes switch to other preferred alternatives. It simulates this behaviour by adding random noise or 'variability' to the product or attribute utilities (Huber, Orme and Miller, 1999). RFC with product variability simulates consumers choosing different products on different occasions typically as a result of inconsistency in evaluating the alternatives. This RFC variant is mathematically equivalent to the Share of Preference (SOP) model. In other words: the Share of Preference model and the RFC model with product variability, although different in their model specifications, are interchangeable. RFC with product *and* attribute variability additionally simulates inconsistency in the relative weights that consumers apply to attributes. RFC with product and attribute variability is thought to generally outperform RFC with only product variability and FC. RFC with only product variability is thought to outperform FC. In order to find the optimal amounts of variability to add to the utilities, grid searches were used in this study (as suggested by Huber, Orme and Miller, 1999). This process took about five full working days to complete for all ten CBC studies.

Correctional measures are procedures that are applied to correct CBC results for unmet methodological assumptions. For instance, CBC assumes that all consumers buy with equal frequencies (every household buys an equal amount of product units during a given time period). Individual respondents' choices should therefore be duplicated proportionally to their purchase frequency. In this study, this is achieved by applying 'respondent weights' in Sawtooth's SMRT (Market Simulator) where every respondent's weight reflects the number of units that a respondent typically buys during a certain time period. These weights were calculated from a self-reported categorical variable added to the questionnaire. CBC assumes also that all the products in the base case have equal distribution levels. This assumption is obviously not met in the real world. In order to correct this problem predicted shares have to be weighted by their distribution levels and rescaled to unity. This can be achieved by applying 'external effects' in Sawtooth's SMRT. The distribution levels came from ACNielsen data and were defined as 'weighted distribution' levels: product's value sales generated by all resellers of that product as a percentage of the product category's value sales generated by all resellers of that product category. Finally, the assumption of CBC that respondents have equal awareness levels for all products in a simulated market is typically not met. Although a correction for unequal awareness levels was initially included in the research design it turned out that awareness data was unavailable for most studies.

HYPOTHESES

In the previous section some brief comments were provided on the expected performance of the techniques relative to each other. This expected behaviour resulted in the following research hypothesis:

With respect to estimation methods:

H1: ICE provides higher external validity than Aggregate Logit. (Denoted as: ICE > Aggregate Logit).

H2: HB provides higher external validity than Aggregate Logit. (Denoted as: HB > Aggregate Logit).

H3: HB provides higher external validity than ICE. (Denoted as: HB > ICE).

With respect to choice models:

H4: RFC with product variability provides higher external validity than FC. (Denoted as: RFC + P > FC).

H5: RFC with product and attribute variability provides higher external validity than FC. (Denoted as: RFC + P + A > FC).

H6: RFC with product and attribute variability provides higher external validity than RFC with product variability. (RFC + P + A > RFC + P).

With respect to correctional measures:

H7: Using the purchase frequency correction provides higher external validity than not using the purchase frequency correction. (Denoted as: PF > no PF).

H8: Using the distribution correction provides higher external validity than not using the distribution correction. (Denoted as: DB > no DB).

SAMPLE AND VALIDATION DATA

The sample consists of ten commercially conducted CBC studies involving packaged goods. All the studied products are non-food items. All the interviews were administered by high quality fieldwork agencies using computer assisted personal interviewing (CAPI). Names of brands are disguised for reasons of confidentiality towards clients. All studies were intended to be representative for the consumer population under study. The same is true for the sample of products that makes up the base case in every study. All studies are designed to the best ability of the responsible project managers of SKIM. All studies were conducted in 2001 except for study J that was conducted in 2002. A study only qualified if all the information was available to estimate the effects for all techniques. This includes external information like distribution and purchase frequency measures in order to test propositions P7 and P8. Refer to table 1 for an overview of the design characteristics of each of the studies.

Table 1.
Individual study characteristics

Study name	Product category	Country of study	Trade channel[a]	Attributes[b]	Sample size[c]	Base case size[d]	Market covered[e] (%)
A	Shampoo	Thailand	TT	Brand, price, SKU, anti-dandruff (y/n)	495	20	63
B	Shampoo	Thailand	MT	Brand, price, SKU, anti-dandruff (y/n)	909	30	53
C	Liquid surface cleaner	Mexico	Both	Brand, price, SKU, aroma, promotion	785	14	65
D	Fabric softener	Mexico	TT	Brand, price, SKU, promotion	243	12	78
E	Fabric softener	Mexico	MT	Brand, price, SKU, promotion	571	20	90
F	Shampoo	Germany	Both	Brand, price, SKU, anti-dandruff (y/n)	659	29	63
G	Dish washing detergent	Mexico	TT	Brand, price, SKU	302	14	92
H	Dish washing detergent	Mexico	MT	Brand, price, SKU	557	21	84
I	Female care	Brazil	both	Brand, price, SKU, wings (y/n)	962	15	59
J	Laundry detergent	United Kingdom	both	Brand, price, SKU, promotion, variant 1, variant 2, concentration	1566	30	51

[a] MT = Modern Trade; TT = Traditional Trade
[b] Attributes used in the CBC design
[c] Number of respondents
[d] Number of products in the base case
[e] Cumulative market share of the products in the base case

The interpretation of these characteristics is straightforward, except perhaps for the type of outlet channel studied. Each of the CBC studies is typically performed for either traditional trade, modern trade or for both trade types. *Traditional trade channels (TT)* is the term used for department stores, convenience stores, kiosks, etc. *Modern trade channels (MT)* consist of supermarkets and hypermarkets. Analysis of a separate trade channel is achieved by drawing an independent sample of consumers who *usually* buy the studied products through a certain trade channel.

The *real market share* of a product is defined as the unit sales of a product in a studied market as a fraction of the total unit sales of all the products in the studied market. The real market shares used for validation purposes were provided by the client and involve ACNielsen market share listings. These are typically measured through point of sale scanner data or through retail audits. Volume shares were converted to unit shares if necessary. Sales data are aggregated nationally over retailers, over two to three monthly

periods. The aggregation over such time periods is believed to neutralise any disturbing short-term promotional effects. Also the *real prices* during the studied time period were provided by the client.

METHODOLOGY

The ten CBC studies are analysed at the individual level. This means that a separate model is constructed for each CBC study, which describes the effects of using the techniques *within that particular CBC study*. An assessment of each hypothesis can now be made by *counting* the number of studies that support it. This limits the evaluation to a qualitative assessment, which is inevitable due to the small sample size (n=10).

The first step is to create a set of dummy variables to code the techniques. The first two columns in table 2 depict all the techniques described earlier. In order to transform all techniques into dummy variables, a base level for each class has to be determined. The base level of a dummy variable can be viewed as the 'default' technique of the class and the effects of the occurrence of the other techniques will be determined relative to the occurrence of the base level. For instance, in order to test hypothesis H1 (ICE > Aggregate Logit), Aggregate Logit has to be defined as the base level. The performance of ICE is now determined relative to that of Aggregate Logit. The last column assigns dummy variables to all techniques that are not base levels. Any dummy variable is assigned the value 0 if it attains the base level and the value 1 if it attains the corresponding technique. The coding used in table 2 is denoted as coding scheme 1.

The problem of coding scheme 1 is that hypothesis H3 (HB > ICE) and H6 (RFC+P+A > RFC+P) cannot be tested. This is because neither of the techniques considered in any one of these propositions is a base level in coding scheme 1. In order to test these two hypotheses we have to apply the alternative dummy variable coding depicted in table 3. This coding is denoted as coding scheme 2. The interpretation of table 3 is analogous to that of table 2.

Table 2.
Coding scheme 1 (used for testing H1, H2, H4, H5, H7 and H8)

Class of techniques	Technique	Base level	Dummy variable	Hypothesis to be tested
Estimation method	Aggregate Logit	*		
	ICE		dICE	H1
	HB		dHB	H2
Choice model	FC	*		
	RFC + P		dRFCP	H4
	RFC + P + A		dRFCPA	H5
Purchase frequency weighting	Not applied (no PF)	*		
	Applied (PF)		dPF	H7
Distribution weighting	Not applied (no DB)	*		
	Applied (DB)		dDB	H8

Table 3.
Coding scheme 2 (used for testing H3 and H6)

Class of techniques	Technique	Base level	Dummy variable	Hypothesis to be tested
Estimation method	Aggregate Logit		dLogit	
	ICE	*		
	HB		dHB2	H3
Choice model	FC		dFC	
	RFC + P	*		
	RFC + P + A		dRFCPA2	H6
Purchase frequency weighting	Not applied (no PF)	*		
	Applied (PF)		dPF2	
Distribution weighting	Not applied (no DB)	*		
	Applied (DB)		dDB2	

The approach to the analysis is to try and construct a full factorial experimental design with all the techniques. Three estimation methods, three choice models, and the application or absence of two different corrections thus result in 36 unique combinations of techniques (3*3*2*2). However, eight combinations are not possible because Aggregate Logit is not compatible with the First Choice model or with the purchase frequency correction. Therefore, the final design only consisted of 28 combinations of techniques. All 28 combinations were dummy variable coded according to coding scheme 1 and coding scheme 2. This double coding ensures the possibility of testing all hypotheses. Note that such a 'double' table was constructed for each of the ten CBC studies.

Each row in the resulting data matrix represents a unique design alternative. Each design alternative is fully described by either the first set of six dummies (coding scheme 1) or the second set of six dummies (coding scheme 2). The next step is to parameterise a market simulator according to the techniques within each row, thus 'feeding' the market simulator a specific design alternative. Although the real market shares of the products in a base case are fixed within each individual study, the way in which a market simulator *predicts* the corresponding choice shares is not. These choice shares are believed to vary with the use of the different techniques. Consequently, two unique external validity measures (MAE and R) can be calculated for each design alternative in the dataset.

The two measures of validity can each be regressed on the two sets of dummy variables. The resulting models describe the absolute effects on MAE and R when different techniques are applied. The estimation of all models was done by linear regression in SPSS. This assumes an additive relationship between the factors. Furthermore, no interaction effects between the techniques were assumed. Linear regression assumes a normally distributed dependent variable (Berenson and Levine, 1996). R and MAE have some properties that cause them to violate this assumption if they are used as a dependent variable. Because the distribution of the R-values is strongly left skewed, the R-values were transformed with Fisher's z' transformation before entering in the regression[2]. An attempt was made to transform MAE with a logistic

[2] The Fisher z' transformation is defined as: Z' = 0.5 ln (1+R / 1-R). The final coefficients were converted back into R-values.

transformation (ln [MAE]) but this did not yield satisfactory results. Therefore, no transformation was used for MAE.

As mentioned earlier, the First Choice model as well as the application of purchase frequency weighting is prohibited for the Aggregate Logit model. The interpretation of the estimated effects must therefore be limited to an overall determination of the magnitude of effects. The effects from the regression model are formally estimated as if all estimation methods could be freely combined with all choice models and purchase frequency correction schemes. Admittedly this is not completely methodologically correct. However, this approach was chosen for the strong desire to determine *independent* effects for estimation methods as well as for choice models. The omission of eight alternatives, all estimated with Aggregate Logit, resulted in a strong increase in collinearity between dummy variables dICE and dHB (correlation of R= -0.75; p=0.00; VIF for both dummies: 2.71). Similar, but weaker, effects occurred between the variables dRFCP and dRFCPA (correlation of R=-0.56; p=0.00; VIF for both dummies: 1.52). Between dummies from coding scheme 2, collinearity occurs to a lesser extent.

No correctional action was undertaken because the collinearity did not seem to affect the individual parameter estimates in either of the models (i.e. many models were able to estimate highly significant effects for both dummy variables within each pair of correlating dummy variables). Furthermore, in every model the bivariate correlations between the dummy variables of each correlating pair fell below the commonly used cut-off levels of 0.8 or 0.9 (Mason and Perreault, 1991). Finally, the VIF for neither variable in neither model fell above the absolute level of 10 which would signal harmful collinearity (Mason and Perreault, 1991).

In summary, ten datasets were generated according to coding scheme 1 and another ten datasets were generated according to coding scheme 2 (each dataset describes one original study). Each dataset consists of 28 combinations of techniques, 28 corresponding values for the dependent variable MAE and 28 values for the dependent variable Z'. The regression models used for hypotheses H1, H2, H4, H5, H7 and H8 are thus defined for every individual study as:

$$MAE_i = \alpha + \beta_1\, dICE_i + \beta_2\, dHB_i + \beta_3\, dRFCP_i + \beta_4\, dRFCPA_i + \beta_5\, dPF_i + \beta_6\, dDB_i + \varepsilon_i$$

$$Z_i = \alpha + \beta_7\, dI{+}CE_i + \beta_8\, dHB_i + \beta_9\, dRFCP_i + \beta_{10}\, dRFCPA_i + \beta_{11}\, dPF_i + \beta_{12}\, dDB_i + \varepsilon_i$$

where:

i	=	Design alternative where i = {1..28}
MAE_i	=	External validity measured by MAE for study alternative i.
Z_i	=	External validity measured by Z' for study alternative i.
β_1 - β_{12}	=	Unstandardized regression coefficients for the dummy variables that were coded according to data matrix 1
α	=	Intercept
ε_i	=	Error term for study alternative i.

The regression models that were used for hypotheses H3 and H6 are defined for every individual study as:

$$MAE_i = \alpha + \beta_1 dLogit_i + \beta_2 dHB2_i + \beta_3 dFC_i + \beta_4 dRFCPA2_i + \beta_5 dPF2_i + \beta_6 dDB2_i + \varepsilon_i$$

$$Z_i = \alpha + \beta_7 dLogit_i + \beta_8 dHB2_i + \beta_9 dFC_i + \beta_{10} dRFCPA2_i + \beta_{11} dPF2_i + \beta_{12} dDB2_i + \varepsilon_i$$

Where:

i	=	Design alternative where i = {1..28}
MAE_i	=	External validity measured by MAE for study alternative i.
Z_i	=	External validity measured by Z' for study alternative i.
$\beta_1 - \beta_{12}$	=	Unstandardized regression coefficients for the dummy variables that were coded according to data matrix 2
α	=	Intercept
ε_i	=	Error term for study alternative i.

Note that variables dLogit, dFC, dPF2 and dDB2 from coding scheme 2 are discarded after the model has been estimated because they are not relevant to the hypotheses.

The values of the regression coefficients are interpreted as the amount with which the external validity measure increases when the dummy variable switches from the presence of the base level to the presence of the technique (assuming that the other dummies in the six-dummy model remain constant). The medians (m_i) and means (μ_i) of the regression coefficients are indicative for the magnitude of the effects in general. The standard deviations (σ_i) of the regression coefficients give an indication of the stability of these estimates across the studies.

Effects for the dummy variables are estimated for each study independently. The eight hypotheses can thus be accepted or rejected *for each individual study*. A hypothesis is supported by a study if there exists a significant positive (R models) or significant negative (MAE models) effect for the respective dummy variable (at or below the 0.05 significance level). The final assessment of a hypothesis is accomplished by counting the number of studies that show a significant positive (R) or negative (MAE) effect. No hard criteria are formulated for the final rejection or acceptance.

Refer to tables 5 and 6 for an overview of individual model statistics. All models were significant at the 0.01 level. As can be seen, the quality of the models is generally high. However, R^2 values are somewhat artificially inflated because the observations are not independent. The bottom rows in each table show the minimum, maximum, median and mean validity measures observed in all studies as well as standard deviations.

Table 5.
Individual model statistics for MAE

	Individual study models									
Statistic	A	B	C	D	E	F	G	H	I	J
R^2	0.960	0.657	0.971	0.810	0.975	0.980	0.995	0.994	0.982	0.975
Std. Error	0.316	0.301	0.158	0.530	0.087	0.032	0.255	0.112	0.089	0.070
F	83.88	6.72	118.51	14.94	121.83	172.98	677.38	605.67	186.20	137.92
Observations: min	1.99	4.42	4.46	4.09	2.82	2.60	2.88	2.62	1.80	1.75
Observations: max	3.41	7.99	7.34	5.56	6.52	3.10	6.19	10.23	3.55	3.07
Observations: median	2.13	5.63	6.13	4.89	3.58	2.85	3.23	6.31	2.73	2.14
Observations: mean	2.33	5.85	5.87	4.73	4.01	2.84	3.93	6.05	2.58	2.18
Observations: std. dev	0.46	1.39	0.83	0.46	1.07	0.20	1.31	3.13	0.57	0.39

Table 6.
Individual model statistics for R (R^2, Std. Error and F are based on z' values)

	Individual study models									
Statistic	A	B	C	D	E	F	G	H	I	J
R^2	0.821	0.725	0.947	0.878	0.936	0.958	0.916	0.781	0.972	0.952
Std. Error	0.158	0.132	0.057	0.082	0.043	0.016	0.241	0.166	0.049	0.080
F	16.07	9.24	62.71	25.18	51.22	79.839	38.172	12.49	119.42	70.055
Observations: min	0.27	-0.14	-0.07	0.11	0.49	0.51	-0.30	0.04	0.72	-0.26
Observations: max	0.75	0.60	0.51	0.53	0.84	0.64	0.49	0.98	0.93	0.66
Observations: median	0.59	0.08	0.22	0.31	0.73	0.59	-0.09	0.61	0.87	0.26
Observations: mean	0.56	0.21	0.27	0.35	0.73	0.58	0.08	0.59	0.86	0.27
Observations: std. dev	0.15	0.29	0.20	0.13	0.11	0.05	0.29	0.34	0.07	0.28

RESULTS

Table 7 shows the unstandardized regression coefficients and their p-values needed for the evaluation of hypothesis H1 to H8 for validity measure MAE. All coefficients, as well as the median and mean values for the coefficients, indicate the absolute change of the Mean Absolute Error (in %-points) between real market shares and shares of choice, as a result of a switch from the base level technique to the technique described by the corresponding dummy variable. Note that positive coefficients denote a negative impact on validity, as MAE is a measure of error.

Table 8 shows the unstandardized regression coefficients and their p-values needed for the evaluation of hypothesis H1 to H8 for validity measure R. All coefficients, as well as the median and mean values for the coefficients, indicate the absolute change of the Pearson correlation coefficient between real market shares and shares of choice, as a result of a switch from the base level technique to the technique described by the corresponding dummy variable. Note that positive coefficients denote a positive impact on validity, as R is a measure of linear relationship.

Figure 9 shows the median and mean values of the regression coefficients for both the MAE models (top graph) and R models (bottom graph).

Table 7.
Unstandardized regression coefficients and p-values for MAE

Study	dICE[a]		dHB[a]		dHB2[b]		dRFCP[c]		dRFCPA[c]		dRFCPA2[d]		dPF[e]		dDB[f]	
	b	p	b	p	b	p	b	p	b	p	b	p	b	p	b	p
A	0.00	0.99	0.05	0.79	0.06	0.67	-2.81	0.00	-2.81	0.00	0.00	1.00	0.01	0.95	-0.83	0.00
B	-0.55	0.01	-0.14	0.47	0.41	0.00	-0.66	0.00	-0.73	0.00	-0.07	0.60	0.05	0.66	0.03	0.83
C	0.45	0.00	0.32	0.00	-0.12	0.07	-1.01	0.00	-1.01	0.00	0.00	1.00	-0.01	0.84	-1.23	0.00
D	-1.04	0.01	0.35	0.30	1.39	0.00	-1.49	0.00	-1.51	0.00	-0.02	0.95	0.11	0.62	-0.44	0.04
E	-0.08	0.18	-0.20	0.00	-0.12	0.00	-0.67	0.00	-0.68	0.00	-0.01	0.76	-0.04	0.27	-0.65	0.00
F	-0.01	0.59	-0.09	0.00	-0.08	0.00	0.00	0.76	0.00	0.76	-0.01	0.50	0.02	0.14	-0.38	0.00
G	-0.37	0.03	0.07	0.66	0.44	0.00	-0.54	0.00	-0.63	0.00	-0.09	0.45	0.03	0.77	-6.10	0.00
H	0.04	0.55	0.19	0.01	0.15	0.00	-2.79	0.00	-2.81	0.00	-0.02	0.70	0.04	0.40	-0.12	0.01
I	-0.05	0.34	0.47	0.00	0.53	0.00	-0.09	0.05	-0.11	0.02	-0.02	0.61	0.15	0.00	-0.97	0.00
J	0.04	0.38	-0.05	0.32	-0.09	0.01	-0.71	0.00	-0.76	0.00	-0.06	0.09	0.02	0.60	-0.36	0.00
Median:	-0.03		0.06		0.11		-0.69		-0.75		-0.02		0.03		-0.55	
Mean:	-0.16		0.10		0.26		-1.08		-1.11		-0.03		0.04		-1.11	
Std. Dev:	0.41		0.23		0.47		1.00		0.99		0.03		0.06		1.80	

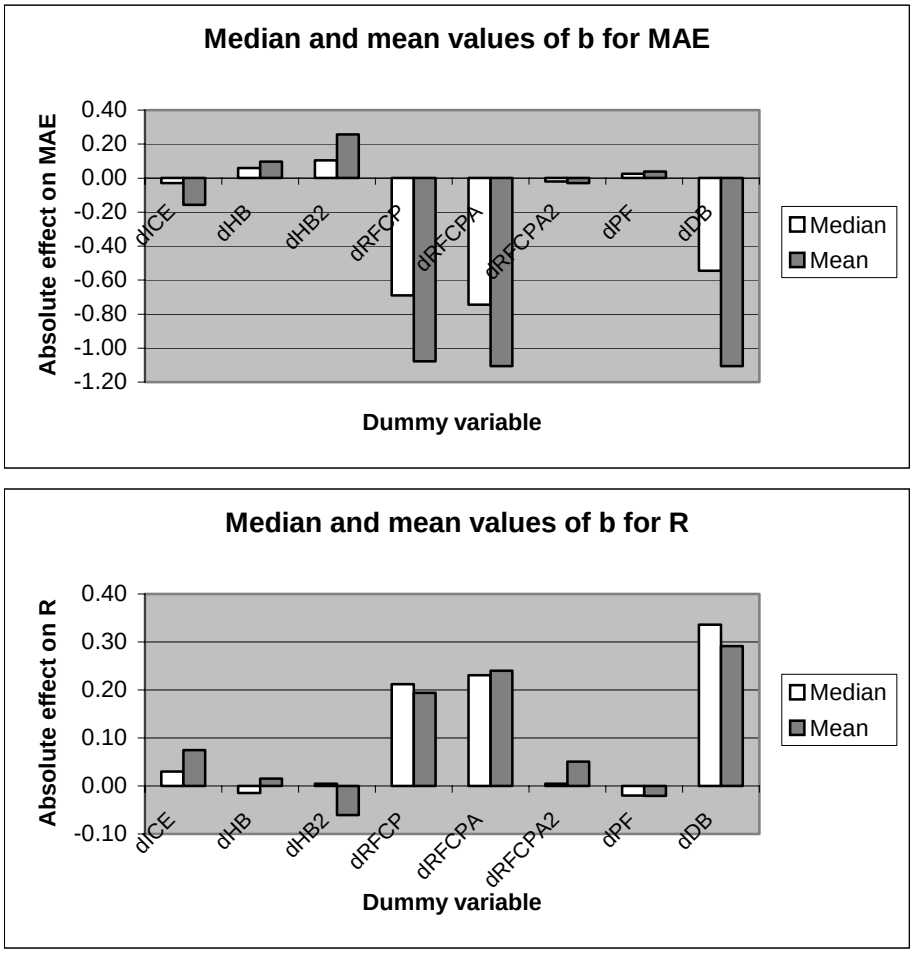
[a] base level: Aggregate Logit
[b] base level: Individual Choice Estimation
[c] base level: First Choice
[d] base level: RFC with product variability
[e] base level: No purchase frequency weighting
[f] base level: No distribution weighting

Table 8.
Unstandardized regression coefficients and p-values for R

Study	dICE[a]		dHB[a]		dHB2[b]		dRFCP[c]		dRFCPA[c]		dRFCPA2[d]		dPF[e]		dDB[f]	
	b	p	b	p	b	p	b	p	b	p	b	p	b	p	b	p
A	-0.03	0.73	-0.01	0.90	0.02	0.74	0.24	0.00	0.24	0.00	0.00	1.00	-0.03	0.61	0.49	0.00
B	0.45	0.00	0.27	0.00	-0.20	0.00	0.09	0.19	0.21	0.00	0.12	0.05	-0.05	0.39	-0.16	0.00
C	0.03	0.44	0.03	0.37	0.00	0.86	0.30	0.00	0.30	0.00	0.00	1.00	0.03	0.23	0.31	0.00
D	0.30	0.00	-0.04	0.45	-0.34	0.00	0.22	0.00	0.22	0.00	0.00	0.99	-0.03	0.46	-0.03	0.27
E	0.03	0.24	0.08	0.01	0.05	0.01	0.21	0.00	0.21	0.00	0.00	0.92	0.02	0.39	0.21	0.00
F	-0.11	0.00	-0.03	0.01	0.08	0.00	0.00	0.98	0.00	0.98	0.00	0.96	-0.01	0.12	0.09	0.00
G	0.16	0.31	0.17	0.27	0.01	0.90	0.14	0.26	0.31	0.01	0.18	0.11	-0.01	0.90	0.87	0.00
H	-0.13	0.22	-0.16	0.13	-0.03	0.67	0.32	0.00	0.37	0.00	0.06	0.43	-0.06	0.42	0.36	0.00
I	0.11	0.00	-0.14	0.00	-0.24	0.00	0.01	0.59	0.02	0.41	0.01	0.75	-0.06	0.01	0.41	0.00
J	-0.06	0.22	-0.02	0.63	0.04	0.25	0.41	0.00	0.52	0.00	0.14	0.00	-0.01	0.81	0.36	0.00
Median:	0.03		-0.01		0.00		0.21		0.23		0.00		-0.02		0.34	
Mean:	0.07		0.02		-0.06		0.19		0.24		0.05		-0.02		0.29	
Std. Dev:	0.18		0.13		0.14		0.13		0.15		0.07		0.03		0.29	

[a] base level: Aggregate Logit
[b] base level: Individual Choice Estimation
[c] base level: First Choice
[d] base level: RFC with product variability
[e] base level: No purchase frequency weighting
[f] base level: No distribution weighting

Figure 9.
Median and mean coefficient values for MAE (top) and R (bottom)



ESTIMATION METHODS

Table 7 indicates that the use of ICE over Aggregate Logit results in an average decrease in MAE (over all ten studies) of 0.16 %-points. Table 8 indicates that the same change in estimation methods results in an average increase in R of 0.07. The effects of using ICE over Aggregate Logit can thus be regarded as very modest. In the same manner, the average effects of using HB over Aggregate Logit and HB over ICE are very small.

However, although the *average* effects of the estimation methods across the ten studies are modest, the relatively high standard deviations at the bottom rows of tables 7 and 8 indicate large variance *between* the coefficients. In other words: it seems that extreme positive and negative coefficients cancel each other out. If we look for instance at the effect on MAE of using ICE instead of Aggregate Logit, we see a set of coefficients ranging from a low of –1.04%-points to a high of 0.45%-points. This not only indicates that effect sizes vary heavily between studies but also that the direction of the effects (whether increasing or decreasing validity) varies between studies.

The findings with regard to the estimation methods can be considered surprising. Although in theory, ICE and HB are often believed to outperform Aggregate Logit, the empirical evidence suggests that this does not always hold in reality. Also the superiority of HB over ICE in the prediction of real market shares cannot be assumed. In general, there seems to be no clearly superior method that 'wins on all occasions'. The performance of each method instead seems to be different for different studies and is dependent on external factors. Possible factors might be the degree of heterogeneity in consumer preferences or the degree of similarity in product characteristics. It is also believed that the design of the CBC study (number of questions per respondent, number of concepts per task) has an effect on the relative performance of HB over ICE.

Choice Models

The use of RFC with product variability over First Choice results in an average decrease in MAE of 1.08 %-points and an absolute increase in R of 0.19. These effects are much more pronounced than any of the average effects of the estimation methods. Furthermore, looking at the individual studies, we can see that the effects are much more stable. Randomised First Choice with product variability (RFC+P) as well as Randomised First Choice with both product and attribute variability (RFC+P+A) outperform First Choice (FC) on most occasions. However, RFC+P+A does not improve external validity much over RFC+P. Because RFC+P is equal to the SOP model, RFC+P+A seems to have limited added value over the much simpler and less time consuming SOP model. The process of determining the optimal amount of product and attribute variability in the RFC model is a tedious process, which does not really seem to pay off. Approximately 95% of the total data generation effort, being around fifty hours, went into the determination of these measures for all ten studies (although some optimising is required for SOP as well).

Note that it is no coincidence that all the effects for the choice models are zero or positive. RFC with only product variability is an extended form of FC where an optimal amount of random variability is determined. If adding variability results in a level of performance worse than FC, the amount of added variability can be set to zero and the RFC model would be equal to the FC model (this actually happens for study F). Hence, RFC can never perform worse than FC. The same holds for the performance of RFC with product and attribute variability over RFC with only product variability.

Purchase Frequency Correction

The use of purchase frequency weighting actually results in a (small) average decrease in validity (increase MAE of 0.04 %-points; decrease R of 0.02). A possible explanation for this finding is that people really buy different products with approximately equal frequency. However, this assumption seems implausible, as larger package sizes typically take longer to consume. It does also not explain the tendency towards decreasing validity. Therefore, a second explanation seems more plausible. Because purchase frequency was measured with a self-reported, categorical variable, it

can easily be the case that this variable was not able to capture enough quantitative detail necessary for these purposes. It could thus add more noise than it explains, resulting in decreasing validity.

Distribution Correction

The mean coefficients for the use of distribution weighting in the MAE model (-1.11%-points) as well as in the R model (0.29) indicate a strong average increase in external validity. At the level of individual studies, the distribution correction almost always results in an improvement in external validity. However, as with most techniques, the magnitude of the improvement can vary between studies and is dependent on external factors. The decision whether to apply distribution weighting or not can make or break a CBC study as it has the potential of turning an invalid study into an extremely valid one. A good example is study G where applying the distribution correction resulted in a reduction of MAE with more than 6%-points and an increase in R of almost 0.9 (although this is an extreme situation).

Assessment of the Hypotheses

A qualitative assessment of the hypotheses can be made by counting the number of studies with a significant negative (MAE) or positive (R) effect for each of the corresponding dummy variables (see table 10). Studies with a significant negative MAE effect or a significant positive R effect indicate an improvement in external validity and hence are considered supportive to the respective hypothesis. Also the number of studies with a significant but opposite effect is reported for each hypothesis.

Table 10.
Assessment of hypotheses (cells display number of studies from a total of 10)

Hypotheses	Description	Number of studies:			
		Supporting[a]		Supporting opposite[b]	
		MAE	R	MAE	R
H1	ICE > Logit	3	3	1	1
H2	HB > Logit	2	2	3	2
H3	HB > ICE	3	2	5	3
H4	RFC + p > FC	9	6	0	0
H5	RFC + p + a > FC	9	8	0	0
H6	RFC + p + a > RFC + p	0	2	0	0
H7	PF corr. > no PF corr.	0	0	1	1
H8	DB corr. > no DB corr.	9	8	0	1

[a] Number of studies that show a significant negative effect (MAE models) or positive effect (R models) for the dummy variable corresponding to the hypothesis at or below the 0.05 significance level.

[b] Number of studies that show a significant positive effect (MAE models) or negative effect (R models) for the dummy variable corresponding to the hypothesis at or below the 0.05 significance level.

I will not provide any hard criteria for the assessment of the hypotheses. I believe every reader has to decide for himself what to take away from the summary above. However, I believe it is fair to state that H1, H2, H3, H6 and H7 cannot be confirmed

with respect to CBC studies for packaged goods *in general*. Accordingly, H4, H5 and H8 *can* be confirmed for these situations.

RECOMMENDATIONS

It seems that utilities from a CBC study should be estimated with all three methods (Aggregate Logit, ICE and HB) if possible. The market simulations resulting from all three methods should be compared on external validity and the best performing method should be chosen. It is advised to try and relate the performance of the methods to some specific external variables that are known in advance. Finding such a relationship (which makes it possible to exclude certain methods in advance) could save time as some of the methods typically take considerable time to estimate (i.e. HB). Candidates for such variables are measures for the heterogeneity between the respondents or the similarity of the attributes and levels of the products in the base case.

If there are no objections against the RFC model, than it can be used instead of the First Choice model. If there are objections against the RFC model, the Share of Preference model can be used as an alternative to the RFC model. Objections to the RFC model could exist because the model is difficult to understand and because the time needed to find the optimal amount of product and attribute variability is quite considerable.

Weighting respondents' choices by their purchase frequency as measured with categorical variables could actually make the results less valid. It is advisable however to experiment with other kinds of purchase frequency measures (e.g. quantitative measures extracted from panel data). Weighting products' shares of choice by their weighted distribution should always be tried as it almost always improves external validity.

DIRECTIONS FOR FUTURE RESEARCH

Future research in the area of external validation of CBC should focus on the following questions. Firstly, what are the determinants of the performance of Aggregate Logit, ICE and HB? Potential determinants include the amount of heterogeneity in consumer preferences, the degree of similarity in product characteristics and study design characteristics like number of choice tasks per respondent.

Secondly, what other factors (besides the techniques investigated in this study) determine external validity? Potential candidates are study design characteristics, sample design characteristics and characteristics of consumers and products in a particular market.

Thirdly, what is the effect of purchase frequency weighting if quantitative instead of qualitative variables are used for the determination of the weights? Consumer panel diaries or POS-level scanning data could perhaps be used to attain more precise purchase frequency measures.

And finally, what are the effects of the investigated techniques for products other than fast moving consumer goods? Because the structure of consumer preference typically differs between product categories, the performance of the techniques is probably

different as well. For instance, a decision about the purchase of a car differs considerably from a decision about the purchase of a bottle of shampoo. Because consumers are expected to engage in less variety seeking when it comes to cars, the performance of RFC over FC will probably be less pronounced.

REFERENCES

Berenson, Mark L. and Levine, David M. (1996), *Basic Business Statistics, Concepts and Applications,* Sixth edition, New Jersey: Prentice Hall, p. 736

Golanty, John (1996), 'Using Discrete Choice Modelling to Estimate Market Share', *Marketing Research*, Vol. 7, p. 25

Green, Paul E. and Srinivasan V. (1978), 'Conjoint Analysis in consumer research: issues and outlook', *Journal of Consumer Research*, 5, p. 338-357 and 371-376

Green, Paul E. and Srinivasan V. (1990), 'Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice', *Journal of Marketing*, 4, p. 3-19

Huber, Joel, Orme, Bryan and Miller, Richard (1999), 'Dealing with Product Similarity in Conjoint Simulations', *Sawtooth Software Conference Proceedings,* p. 253-266

Johnson, Richard M. (1997), 'Individual Utilities from Choice Data: A New Method', *Sawtooth Software Conference Proceedings,* p. 191-208

Mason, Charlotte H. and Perreault, William D. Jr. (1991), 'Collinearity, Power, and Interpretation of Multiple Regression Analysis', *Journal of Marketing Research*, Volume XXVIII, August, p. 268-80

Natter, Martin, Feurstein, Markus and Leonhard Kehl (1999), 'Forecasting Scanner Data by Choice-Based Conjoint Models', *Sawtooth Software Conference Proceedings,* p. 169-181

Orme, Bryan K. and Mike Heft (1999),'Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results', *Sawtooth Software Conference Proceedings,* p. 183-199

Sentis, Keith and Li, Lihua (2001), 'One Size Fits All or Custom Tailored: Which HB Fits Better?', *Sawtooth Software Conference Proceedings, p. 167-175*

Wittink, Dick R. (2000), 'Predictive Validation of Conjoint Analysis', *Sawtooth Software Conference Proceedings,* p. 221-237

Wittink, Dick R., Cattin, Philippe (1989) 'Commercial Use of Conjoint Analysis: An Update', *Journal of Marketing*, Vol.53 (July), p. 91-96

COMMENT ON ARENOE AND ROGERS/RENKEN

DICK R. WITTINK
SCHOOL OF MANAGEMENT, YALE UNIVERSITY

Conjoint analysis in one or another of its various implementations has been around for more than thirty years. The approach is now widely accepted, and its users obviously believe the results have a high degree of external validity. Yet there is scant evidence of the extent to which conjoint results allow users to offer strong and convincing support of the proposition that the effects of *changes* in actual product or service characteristics and price are accurately predicted. Arenoe (2003) and Rogers and Renken (2003) deserve support for providing new insights into the external validity of CBC. Arenoe addresses the determinants of CBC's ability to predict market shares in store audit or scanner data. Rogers and Renken compare the price sensitivity inferred from CBC with results derived from econometric models applied to scanner data.

Importance. Research on external validation is important for several reasons. One is that the client wants to understand how simulation results relate to the marketplace. It is ultimately impossible for the client to interpret the effects of "what if" simulations if those effects cannot be translated into real-world effects. It is now well known that conjoint-based simulations rarely provide results that correspond perfectly to the marketplace. Due to systematic differences between survey and market data, conjoint researchers often refer to the simulation output as "preference" or "choice" shares (as opposed to market shares). This is to make users sensitive to the idea that adjustments are required before observable market shares can be successfully predicted. For example, marketplace behavior is influenced by decision makers' awareness of and access to alternative products/services. In addition, customers differ in the likelihood or in the frequency and volume of category purchases.

The marketing literature contains relevant papers that address such systematic differences. For example, Silk and Urban (1978) proposed a concept-testing or pretest market model called ASSESSOR. Urban and Hauser (1980) provide external validity test results for this approach. Predicted shares for concepts are compared with actual shares observed in a test market for products introduced. Importantly, the market simulations allow the user to specify alternative awareness and distribution levels for a new product that is a test market candidate. Urban and Hauser (p. 403) show for 25 products that the predictive validities improve considerably if the *actual* awareness and distribution levels are used instead of the levels management planned to achieve. The Mean Absolute Deviation (MAD) between actual and predicted test market shares is 0.6 percentage points if actual levels are used but 1.5 points with the planned levels.

In conjoint simulations one can actually use respondent-specific awareness and distribution levels. The advantage over using aggregate adjustments, as in ASSESSOR-based share predictions, is that interaction effects between awareness/availability of alternatives and changes in product characteristics can be accounted for. For example, suppose there are two segments of potential customers that differ in the sensitivity to changes in, say, product performance. If the product of interest is currently unknown to

the segment that is sensitive to product performance (but known to the segment with little performance sensitivity), then a proper market simulation will show how the share can be increased with improved product performance if this segment is also made aware of the product. Thus, the benefit of changes in price and product characteristics may depend on the awareness and availability of the product to potential customers. By allowing this dependency to be shown, users can assess the benefits and costs of such related phenomena.

It should be clear that any attempt to determine external validity must adjust for awareness and distribution. Arenoe found that distribution was the factor with the highest average impact on his external validity measures. He lacked brand awareness information for several data sets and therefore omitted this variable from his analysis. Surprisingly, however, the variable capturing frequency of purchase multiplied by purchase volume per purchase occasion did not help. This is surprising because CBC only captures choice of an item *conditional* upon category purchase. It is noteworthy that researchers studying supermarket purchase behavior use three models to predict all the elements that pertain to purchase behavior: (1) a model of category purchase incidence; (2) a model of brand choice, conditional upon category purchase; and (3) a model of quantity or volume, conditional upon the brand chosen (e.g. Gupta 1988; Bell et al. 1999). Thus, CBC simulation results cannot be expected to predict market shares accurately unless category purchase incidence and volume are taken into account. For categories with fixed consumption levels, such as detergents, proper measurements of household-level purchase or consumption should improve the predictive validity of CBC results. For categories with expandable consumption levels, such as soft drinks or ice cream, it may be useful to also have separate approaches that allow users to predict purchase incidence and volume as a function of price and product characteristics.

Arenoe used market share data that pertained to a few months *prior* to the timing of data gathering for CBC. The use of past purchase data is understandable given that the client wants to relate the CBC-based simulations to the latest market data. Nevertheless, external validation must also focus on future purchases. Ultimately, management wants to make changes in product characteristics or in prices and have sufficient confidence that the predicted effects materialize. Thus, we also need to know how the predicted changes in shares based on changes in products correspond to realized changes. Wittink and Bergestuen (2001) mention that there is truly a dearth of knowledge about external validation for changes in attributes based on conjoint results.

Determinants. Although it is a fair assumption that conjoint in general will provide useful results to clients, we can only learn how to make improvements in aspects such as data collection, measurement, sampling, and question framing if we relate variations in approaches to external validation results. There is a vast amount of research in the behavioral decision theory area that demonstrates the susceptibility of respondents' preferences and choices to framing effects. For example, the subjective evaluations of lean/fat content of beef depend on whether we frame the problem as, say, 75% lean or 25% fat. Behavioral researchers create variations in choice sets to demonstrate violations of basic principles implicitly assumed in conjoint. An example is the compromise effect that shows conditions under which an alternative can achieve a higher share if it is one of three options than if it is one of two options (see Bettman et al., 1998 for a review).

Given that marketplace choices are influenced by characteristics of the choice environment, including how a salesperson frames the options, how products are advertised or how the alternatives are presented on supermarket shelves, it is important that we take such characteristics into account in conjoint study designs. The more one incorporates actual marketplace characteristics into a study design, the stronger the external validity should be. The SKIM designs in Arenoe's data include, for example, a *chain-specific* private label alternative.

It seems reasonable to argue that the effects of variations in estimation methods can be appropriately studied based on *internal* validity. However, the effects of variations in, say, the number of attributes, the number of alternatives in choice sets, and the framing of attributes should be determined based on *external* validity. Whenever external validity is measured, it is useful to include internal validity so that one can learn more about the conditions under which these two criteria converge.

CBC may be especially attractive to clients who want to understand customers' price sensitivities. Nevertheless, for consumer goods, managers should also consider using scanner panel data to estimate price sensitivities. Household scanner panel data can be used to show how purchase incidence, brand choice and quantity vary (in a separate model for each component) as a function of temporary price cuts and other promotions (see Van Heerde et al. 2002 for a new interpretation of the decomposition of sales effects into primary and secondary components). However, these data rarely allow for meaningful estimation of *regular* price effects (because regular prices vary little over time). If CBC is used to learn about regular price effects, respondents must then be carefully instructed to approach the task accordingly. In fact, for an analysis of regular price effects, it should be sufficient to focus exclusively on conditional brand choice in a CBC study. In that case, the ability to use scanner data to determine the external validity of price sensitivity measures is compromised. That is, temporary price cut effects generally do not correspond to regular price effects.

Measures. Just as researchers increasingly compare the performance of alternative approaches on internal validity measures relative to the reliability of holdout choices, it is important that external validity measures are also computed relative to corresponding measures of uncertainty. Apart from such aspects as sampling error and seasonality, in the United States aggregate measures based on store sales exclude sales data from Walmart. Importantly, Walmart accounts for an increasing part of total retail sales of many product categories. Disaggregate measures based on household scanner panels can overcome this limitation. Household scanner panel data also allow for the accommodation of household heterogeneity in a manner comparable to the approaches used for CBC.

Finally, it is worth noting that for both internal and external validity there are aggregate and disaggregate measures. Elrod (2001) argues that the hit rate is subject to limitations and he favors log likelihood measures. But clients cannot interpret log likelihoods. It seems more appropriate to let the measure depend on the nature of the application. In traditional market research applications, clients want to predict market shares. In that case, it is meaningful to use MAD (Mean Absolute Deviation) between actual and predicted shares as the criterion. However, for a mass customization application, the natural focus is the prediction of individual choices. The hit rate is the

best available measure for the prediction of discrete choices. It is worth noting that the accuracy of the hit rate combines bias and sampling error. Thus, for mass customization applications, researchers should balance the quality of data collection and analysis with simplicity. By contrast, the prediction of market shares tends to be maximized with the approach that best captures each respondent's true preferences, revealed in the marketplace. In other words, for the best MAD results, bias should be minimized since the uncertainty of individual-level parameter estimates is largely irrelevant. That is, the uncertainty of individual predictions that influences the hit rates plays a decreasing role in aggregate measures such as MAD, as the number of respondents increases.

REFERENCES

Arenoe, Bjorn (2003), "Determinants of External Validity of CBC", presented at the tenth Sawtooth Software Conference.

Bell, David B., Jeongweng Chiang and V. Padmanabhan (1999), "The Decomposition of Promotional Response: An Empirical Generalization", *Marketing Science*, 18 (4) pp. 504-26.

Bettman, James R., Mary Frances Luce and John W. Payne, "Constructive Consumer Choice Processes", *Journal of Consumer Research*, 25 (December), pp. 187-217.

Elrod, Terry (2001), "Recommendations for Validation of Choice Models", *Sawtooth Software Conference Proceedings*, pp. 225-43.

Gupta, Sunil (1983), "Impact of Sales Promotion on When, What and How Much to Buy", *Journal of Marketing Research*, 25 (November), pp. 342-55.

Rogers, Greg and Tim Renken (2003), "Validation and Calibration of CBC for Pricing Research", presented at the tenth Sawtooth Software Conference.

Silk, Alvin J. and Glen L. Urban (1978), "Pre-Test-Market Evaluation of New Packaged Goods: A Model and Measurement Methodology", *Journal of Marketing Research*, 15 (May), pp. 171-91.

Urban, Glen L. and John R. Hauser (1980), *Design and Marketing of New Products*. Prentice Hall.

Van Heerde, Harald J., Sachin Gupta and Dick R. Wittink (2002), "Is 3/4 of the Sales Promotion Bump due to Brand Switching? No, it is 1/3", *Journal of Marketing Research*, forthcoming.

Wittink, Dick R. and Trond Bergestuen (2001), "Forecasting with Conjoint Analysis", in: J. Scott Armstrong (ed.) *Principle of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer, pp. 147-67.

CONJOINT ANALYSIS APPLICATIONS

LIFE-STYLE METRICS: TIME, MONEY, AND CHOICE

THOMAS W. MILLER
RESEARCH PUBLISHERS, LLC

> The irretrievability of the past, the inexorable passage of the present, the inevitable approach of the future, must at some time have given pause to every thinking person — a progression that, whatever its content, is ceaseless and unremitting, yet the movement of which is virtually unintelligible, is not literally motion at all, and, for the most part, seems irrelevant to the nature of events by whose sequence it is constituted and measured. If we were not habitually puzzled by all this, it is only through indifference bred of perpetual familiarity (Error E. Harris 1988, p. xi).

We have two things to spend in life: time and money. Who we are, what we do, how we live — these are defined, in large measure, by how we spend time and money. We make choices about lifestyles, just as we make choices about products. Lifestyle, goods, and service choices are inextricably intertwined.

Time is the fundamental currency of life. Days, hours, minutes, seconds — we measure time with precision across the globe. Time is the great equalizer, a common denominator resource spent at the same, constant rate by everyone. It can't be passed from one person to another. It can't be stored or restored. Time is the ultimate constraint upon our lives.

Money, a medium of exchange, is easily passed from one person to another. Money not used today can be saved for future days. Money invested earns money. Money borrowed costs money. Accumulated money, associated with social status, is passed from one generation to the next. Much of economics and consumer research has concerned itself with price or the money side of the time-money domain. Transactions or trades are characterized as involving an exchange of goods and services for money. Money for labor, money for goods — this is a way of valuing time spent and property purchased.

Time is also important in markets for goods and services. Economists and consumer researchers would be well advised to consider tradeoffs between products and the time spent to acquire, learn about, use, and consume them. This paper provides an introduction to the expansive literature relating to time, citing sources from economics, the social sciences, and marketing. It also introduces a series of choice studies in which time and money (or prices) were included as attributes in lifestyle and product profiles. These studies demonstrate what we mean by lifestyle metrics.

TIME, MONEY, AND CHOICE LITERATURE

Literature regarding time, money, and choice is extensive. This section reviews sources from economics, the social sciences, and marketing. We focus upon time, providing discussion about time perception and time allocation. For the money side of the time-money domain, a comprehensive review may be found in Doyle (1999).

ECONOMICS AND THE LABOR-LEISURE MODEL

Figure 1 shows the classic labor-leisure model of labor economics. The model shows leisure and consumption as benefits or goods. We imagine that people desire these and that more leisure time hours *H* and more units of consumption *C* are better. Individuals differ in their valuation of leisure time versus consumption, as reflected in the utility function *U(H,C)*.

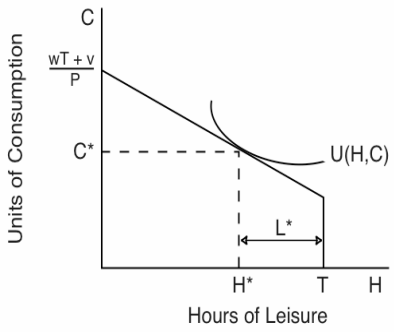


Figure 1. Labor-Leisure Model

Time and money resources are limited. The maximum hours in the day *T* is twenty-four hours. And, ignoring purchases on credit, units of consumption are limited by earnings from hours of labor *L*, income from investments *v*, and the average price per unit of consumption *P*. If an individual worked twenty-four hours a day, units of consumption would be at its maximum, shown by the intersection of the leisure-consumption budget line with the vertical axis. Working twenty-four hours a day, however, would leave no time for leisure.

The intersection of the leisure-consumption budget line with an individual's utility function provides the optimal levels of leisure and consumption, shown as *H** and *C**, respectively. Utility functions differ across individuals with some people choosing to work more and others to work less. Further discussion of the labor-leisure model may be found in introductions to labor economics (e.g. Kaufman 1991).

For conjoint and choice researchers, the labor-leisure model provides a useful review of economic principles, utility concepts, and tradeoffs that underlie much of standard conjoint and choice modeling. The typical marketing study focuses upon tradeoffs between product features or between product features and prices. Much of conjoint and choice research concerns the consumption component of the labor-leisure model, while ignoring the time component.

SOCIAL SCIENCES LITERATURE

Time has been an important topic of research and discussion in the social sciences. Since the days of William James (1890), psychologists have observed that our perception of time varies with the activities in which we engage. Time seems to pass quickly when we are active, slowly when inactive or waiting. Time may seem to pass quickly when we are doing things we like. Social researchers like Csikszentmihalyi (1990) and Flaherty (1999) have provided psychological and phenomenological analyses of how we experience time (time-consciousness).

Time allocation studies have been a staple of the sociologist's repertoire for many years. Juster and Stafford (1991) reviewed technologies associated with such studies, citing the advantages of time diaries over recall reports. Time allocation has also been a subject of considerable debate among social scientists. Schor (1992) argued that people in the United States have less leisure time today than in previous years, that they sacrifice leisure for consumption. Hochschild (1989, 1997) built upon similar themes, focusing upon the special concerns of working women. Citing the results of time diary studies, Robinson and Godbey (1997) disputed the claims of Schor, arguing that Americans have more leisure time today than ever before. People's perception of free time may be distorted by the fact that they spend so much of it watching television.

Social psychologists, sociologists, and anthropologists have observed cultural differences in time perception and allocation. Hall (1983) noted how keeping time, rhythms, music, and dance vary from culture to culture. People in some cultures are monochromic, focused upon doing one thing at a time. They keep schedules and think of activities as occurring in a sequence. Many insist that things be done on time or "by the bell." People in other cultures are polychromic, willing to do more than one thing at a time and not concerned about being on time.

Levine (1997) conducted field research across various countries and cultures. He and his students observed the pace of life in thirty-one countries, noting walking speeds, clock accuracy, and postal delivery times. Similar studies were conducted across thirty-six U.S. cities. Among countries studied, those with the fastest pace of life were Switzerland, Ireland, Germany, and Japan; those with the slowest pace of life were El Salvador, Brazil, Indonesia, and Mexico. The United States and Canada fell toward the middle of the list. Across the United States, cities in the Northeast had the fastest pace of life, whereas cities in California and in the South had the slowest pace.

Usunier and Valette-Florence (1994) proposed a measure of individual differences in time perceptions or time orientations, identifying five factors:

- Economic time (degree to which people follow schedules),
- Orientation toward the past,
- Orientation toward the future,
- Time submissiveness (acceptance of tardiness), and
- Usefulness of time (productivity versus boredom).

TIME IN MARKETING

The role of time in consumer research has been the subject of numerous review papers. Jacoby, Szybillo, and Berning (1976) cited economic theorists Stigler (1961) and Becker (1965), as well as work in psychology and sociology. They discussed time pressures in consumer searching and shopping and the role of time in brand loyalty. Ratchford (2001) reviewed economic concepts relevant to marketing and consumer research, noting the importance of time investments in consumption.

Graham (1981) identified three ways of perceiving time: linear-separable, circular-traditional, and procedural-traditional, noting that much of consumer research presupposes a Western linear-separable conception. Bergadaà (1990) argued that consumer research needed to acknowledge the importance of time in consumer decisions.

Time is important to product choice. When we commute to work, we choose to drive a car, ride a bike, or use public transportation, largely on the basis of the time needed for various modes of transportation. The decision to buy a cell phone is sometimes related to the expectation that time will be saved by being able to participate in concurrent activities.

Transaction costs are related time costs. We spend time specifying and ordering products, searching for the right product at the right price. The appeal of online shopping is associated with time and price savings. Switching costs are associated with time costs; the purchases of products within a category are affected by previous experiences with products within that category. Brand loyalty may be thought of as the embodiment of many product experiences. When we choose a software application, such as a word processor, we consider the time investment we have made in similar applications, as well as the time it will take to learn the new application.

Time has been an important feature of empirical studies in transportation. Much original work in discrete choice modeling concerned transportation mode options, with transit time and cost being primary attributes affecting choice (Hensher and Johnson 1981; Ben-Akiva and Lerman 1985). Recent studies, such as those reviewed by Louviere, Hensher, and Swait (2000), illustrate the continued importance of time components in transportation choice.

Waiting time has been an important consideration in service research, with many studies showing a relationship between waiting time and service satisfaction (Taylor

1994; Hui and Tse 1996). Perceptions of service waiting time vary with the waiting experience (Katz, Larson, and Larson 1991), the stage of service (Hui, Thakor, and Gill 1998), and the type of individual doing the waiting (Durrande-Moreau and Usunier 1999).

CONJOINT AND CHOICE STUDY EXAMPLES

This section provides examples of conjoint and choice studies with time included explicitly within product profiles or scenarios. Scenarios with time, money, and lifestyle attributes provide an evaluation of lifestyles and a potential mechanism for segmentation. Scenarios with product attributes, as well as time and money attributes, provide an evaluation of products in terms of time and money tradeoffs. Study designs and respondent tasks are reviewed here. Analysis methods, time-money tradeoff results, and consumer heterogeneity will be the subjects of future papers.

STUDY 1. COMPUTER CHOICE

This study, conducted in cooperation with Chamberlain Research Consultants of Madison, Wisconsin, involved a nationwide sample of home computer buyers. The objective of the study was to determine factors especially important to home computer buyers. Conducted in the fall of 1998, just prior to the introduction of Microsoft Windows 98, the study examined benefits and costs associated with switching between or upgrading computer systems. It considered learning time as a factor in computer choice and tradeoffs between price and learning time in consumer choice.

An initial survey was conducted by phone. Consumers were screened for their intentions to buy home computers within two years. Respondent volunteers were sent a sixteen-set choice study with each set containing four computer system profiles. Respondents were contacted a second time by phone to obtain data from the choice task. Exhibit 1 shows the attributes included in the study.

Exhibit 1. Computer Choice Study Attributes

- Brand name (Apple, Compaq, Dell, Gateway, Hewlett-Packard, IBM, Sony, Sun Microsystems)
- Windows compatibility (65, 70, 75, 80, 85, 90, 95, 100 percent)
- Performance (just, twice, three, or four times as fast as Microsoft Windows 95)
- Reliability (just as likely to fail versus less likely to fail than Microsoft Windows 95)
- Learning time (4, 8, 12, 16, 20, 24, 28, 32 hours)
- Price ($1000, $1250, $1500, $1750, $2000, $2250, $2500, $2750)

The study showed that learning time could be included in product profiles, in conjunction with product and price attributes. Respondents encountered no special difficulty in making choices among computer profiles in this study. Aggregate study results showed that learning time was an important determinant of computer choice, though not as important as Windows compatibility, price, or system performance.

STUDY 2. JOB AND LIFESTYLE CHOICES

This study concerned student job and lifestyle choices. Respondents were students enrolled in an undergraduate marketing management class at the University of Wisconsin–Madison. Time and money factors were included in hypothetical job descriptions presented in paper-and-pencil and online surveys. The choice task involved twenty-four paired comparisons between profiles containing six or ten job and lifestyle attributes. Measurement reliability was assessed in a test-retest format. Exhibit 2 shows attributes included in the choice profiles.

Exhibit 2. Job and Lifestyle Study Attributes

- Annual salary ($35K, $40K, $45K, $50K)
- Typical work week (30, 40, 50, 60 hours)
- Fixed work hours versus flexible work hours
- Annual vacation (2 weeks, 4 weeks)
- Location (large city, small city)
- Climate (mild with small seasonal changes versus large seasonal changes)
- Work in small versus large organization
- Low-risk, stable industry versus high-risk, growth industry
- Not on call versus on call while away from work
- $5,000 signing bonus versus no signing bonus

Results from this study, reported in Miller et al. (2001), showed that students could make reliable choices among job and lifestyle profiles. Paper-and-pencil and online modalities yielded comparable results.

STUDY 3. TRANSPORTATION CHOICE

This study involved transportation options in Madison, Wisconsin. Exhibit 3 shows attributes from the choice task. As with many transportation studies, time and money attributes were included in the transportation profiles or scenarios. Respondents included students registered in an undergraduate course in marketing management at the University of Wisconsin-Madison. Students had no difficulty in responding to thirty-two choice sets with four profiles in each set. Paper-and-pencil forms were used for this self-administered survey. Results will be reviewed in future papers.

Exhibit 3. Transportation Study Attributes

- Transportation mode (car, bus, light rail, trolley)
- Good weather (sunny or partly cloudy) versus bad weather (cold with rain or snow)
- Cost ($2, $3, $4, $5)
- Total travel time (20, 30, 40, 50 minutes)
- Wait time (none, 5, 10, 15 minutes)
- Walk (no walk, 2, 4, 8 blocks)

STUDY 4. CURRICULUM AND LIFESTYLE CHOICES

Conducted in cooperation with researchers at Sawtooth Software and MIT, this study involved adaptive conjoint and choice tasks in a test-retest format administered in online sessions. Exhibit 4 shows the extensive list of study attributes. Students were able to successfully complete the conjoint and choice tasks, providing reliable data for the analysis of time-money tradeoffs concerning curricula and lifestyle options. Results were summarized in Orme and King (2002).

Exhibit 4. Curriculum Study

Business Program and Grade Attributes

- Number of required courses/credits for the business degree (12/36, 16/48, 20/60, 24/72)
- Major options from current list of ten possible majors (choice of one, two, or three majors; choice of one or two majors; choice of one major; general business degree with no choice of majors)
- No mandatory meetings with academic advisor versus mandatory meetings
- Opportunity to work on applied business projects and internships for credit versus no opportunity to earn credit
- Students not required to provide their own computers versus students required to own their own computers
- Grades/grade-point-average received (A/4.0, AB/3.5, B/3.0, BC/2.5)

Time and Money Attributes

- Hours per week in classes (10, 15, 20, 25)
- Hours per week spent working in teams (0, 5, 10, 15)
- Hours of study time per week (10, 15, 20, 25)
- Hours per week spent working at a job for pay (0, 5, 10, 15, 20)
- Spending money per month ($150, $300, $450, $600, $750)

CONCLUSIONS

Time can be included in product and lifestyle profiles for use in many research contexts. Adult consumer and student participants experience no special difficulties in responding to conjoint and choice tasks involving time-related attributes. When included within conjoint and choice tasks, time-related attributes provide a mechanism for lifestyle metrics.

By including time in our studies, we can assess the importance of learning time in product decisions. We can see how waiting and travel time can affect choice of transportation mode. We can see how people make time-money tradeoffs among job and lifestyle options. Time is important to lifestyle and product choices, and we have every reason to include it in conjoint and choice research.

This paper provided a review of relevant literature and described study contexts in which time could be used effectively. We reviewed study tasks and cited general results based upon aggregate analyses. Future research and analysis should concern individual differences in the valuation of time-related attributes. Analytical methods that permit the modeling of consumer heterogeneity should prove especially useful in this regard.

REFERENCES

Becker, G. S. 1965, September. A theory of the allocation of time. *The Economic Journal* 75:493–517.

Ben-Akiva, M. E. and S. Lerman 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, Mass: MIT Press.

Bergadaà, M. M. 1990, December. The role of time in the action of the consumer. *Journal of Consumer Research* 17:289–302.

Csikszentmihalyi, M. 1990. Flow: *The Psychology of Optimal Experience*. New York: Harper and Row.

Doyle, K. O. 1999. *The Social Meanings of Money and Property*. Thousand Oaks, Calif.: Sage.

Durrande-Moreau, A. and J-C. Usunier 1999, November. Time styles and the waiting experience: An exploratory study. *Journal of Service Research* 2(2):173–186.

Flaherty, M. G. 1999. *A Watched Pot: How We Experience Time*. New York: University Press.

Graham, R. J. 1981. The role of perception of time in consumer research. *Journal of Consumer Research* 7:335–342.

Hall, E. T. 1983. *The Dance of Life: The Other Dimension of Time*. New York: Anchor Books.

Harris, E. E. 1988. *The Reality of Time*. Albany, N.Y.: State University of New York Press.

Hensher, D. A. and L. W. Johnson 1981. *Applied Discrete-Choice Modeling*. New York: Wiley.

Hochschild, A. R. 1989. *The Second Shift*. New York: Avon.

Hochschild, A. R. 1997. *The Time Bind: When Work Becomes Home & Home Becomes Work*. New York: Metropolitan Books.

Hui, M. K., M. V. Thakor, and R. Gill 1998, March. The effect of delay type and service stage on consumers' reactions to waiting. *Journal of Consumer Research* 24:469–479.

Hui, M. K. and D. K. Tse 1996, April. What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing* 60:81–90.

Jacoby, J., G. J. Szybillo, and C. K. Berning 1976, March. Time and consumer behavior: An interdisciplinary overview. *Journal of Consumer Research* 2:320–339.

James, W. J. 1890. *The Principles of Psychology*. New York: Dover. Originally published by Henry Holt & Company.

Juster, F. T. and F. P. Stafford 1991, June. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature* 29(2):471–523.

Katz, K. L., B. M. Larson, and R. C. Larson 1991, Winter. Prescription for the waiting-in-line blues: Entertain, enlighten, and engage. *Sloan Management Review* 44:44–53.

Kaufman, B. E. 1991. *The Economics of Labor Markets* (third ed.). Orlando, Fla.

Levine, R. 1997. *The Geography of Time: The Temporal Misadventures of a Social Psychologist, or How Every Culture Keeps Just a Little Bit of Time Differently*. New York: Basic Books.

Louviere, J. J., D. A. Hensher, and J. D. Swait 2000. *Stated Choice Models: Analysis and Application*. Cambridge: Cambridge University Press.

Miller, T. W., D. Rake, T. Sumimoto, and P. S. Hollman 2001. Reliability and comparability of choice-based measures: Online and paper-and-pencil methods of administration. *2001 Sawtooth Software Conference Proceedings*. Sequim, Wash.: Sawtooth Software.

Orme, B. and W. C. King 2002, June. Improving ACA algorithms: Challenging a 20-year-old approach. Paper presented at the 2002 Advanced Research Techniques Forum, American Marketing Association.

Ratchford, B. T. 2001, March. The economics of consumer knowledge. *Journal of Consumer Research* 27:397–411.

Robinson, J. P. and G. Godbey 1997. *Time for Life: The Surprising Ways Americans Use Their Time*. University Park, Penn.: The Pennsylvania State University Press.

Stigler, G. S. 1961, June. The economics of information. *Journal of Political Economy* 59:213–225.

Taylor, S. 1994, April. Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing* 58:56–69.

Usunier, J-C. and P. Valette-Florence 1994. Individual time orientations: A psychometric scale. *Time and Society* 3(2):219–241.

MODELING PATIENT-CENTERED HEALTH SERVICES USING DISCRETE CHOICE CONJOINT AND HIERARCHICAL BAYES ANALYSES

CHARLES E. CUNNINGHAM[1]
MCMASTER UNIVERSITY
DON BUCHANAN
MCMASTER CHILDREN'S HOSPITAL
KEN DEAL
MCMASTER UNIVERSITY

ACKNOWLEDGEMENTS

This research was supported by grants from the Ontario Early Years Challenge Fund and the Hamilton Health Sciences Research Development Fund. Dr. Cunningham's participation was supported by the Jack Laidlaw Chair in Patient Centred Care in the Faculty of Health Sciences at McMaster University. The authors express their appreciation for the research support provided by Heather Miller.

INTRODUCTION

North American epidemiological studies suggest that a considerable majority of children with psychiatric disorders do not receive professional assistance (Offord, et al., 1987). While these data reflect the limited availability of children's mental health services, utilization studies also suggest that, when demonstrably effective children's mental health programs are available, a significant majority of families who might benefit do not use these services. Families whose children are at higher risk are least likely to enroll. As part of a program of school-based interventions, for example, (Boyle, Cunningham, et al., 1999; Hundert, Boyle, Cunningham, Duku, Heale, McDonald, Offord, & Racine; 1999), Cunningham, et al., (2000) screened a community sample of 1498 5 to 8 year children. Parents were offered school-based parenting courses. Only 28% of the parents of high risk children (externalizing t-score $\geq$ 70) enrolled in these programs (Cunningham, et al., 2000). This level of utilization is consistent with other studies in this area (Barkley, et al., 2000; Hawkins, von Cleve, & Catalano, 1991). Low utilization and poor adherence means that the potential benefits of demonstrably effective mental health services are not realized (Kazdin, Mazurick, & Siegel, 1994) and that significant economic investments in their development are wasted (Vimarlund, Eriksson, & Timpka, 2001).

[1] Charles E. Cunningham, Ph.D., Professor, Department of Psychiatry and Behavioral Neurosciences, Jack Laidlaw Chair in Patient-Centred Health Care, Faculty of Health Sciences, McMaster University. Don Buchanan, Clinical Manager, Child and Family Centre, McMaster Children's Hospital. Ken Deal, Ph.D., Professor and Chair, Department of Marketing, Michael DeGroote School of Business, McMaster University

A growing body of evidence suggests that low utilization, poor adherence, and premature termination reflect failures in the design and marketing of children's mental health services. For example, it is unclear whether advertising strategies reach parents who might be interested in these programs. In a school-based program in which all parents were sent flyers regarding upcoming parenting courses, follow-up interviews suggested that a significant percentage were not aware that these services were available (Cunningham, et al., 2000).

Second, parents may not understand the longer-term consequences of early childhood behavior problems, the risks associated with poor parenting, or the benefits of parenting programs.

Third, low utilization suggests that advertisements regarding parenting services may not be consistent with the needs of different user groups. Readiness for change models, for example, suggest that users need different information at different stages of the health service delivery process (Cunningham, 1997). Parents at a precontemplative stage, who have not considered the changes they might make to improve their child's mental health, or those at the contemplative stage, who are considering change, require information regarding the potential benefits of a treatment related change, the consequences of failing to change, and assurance that the costs, risks, or logistical demands of change can be managed (Cunningham, 1997). Patients at a preparatory stage, need information regarding the advantages and disadvantages of treatment options and the details needed to plan the change process (e.g. times and locations of parenting groups). Patients at the action and maintenance stage require information regarding the strategies needed to execute and sustain change.

Finally, when we reach prospective users with effective advertising messages, logistical barriers often limit the utilization of potentially useful children's mental health services (Cunningham, et al., 1995; 2000; Kazdin, Holland & Crowley, 1997; Kazdin & Wassell, 1999). Cunningham, et al., (2000) for example found that most parents attributed their failure to participate in school-based parenting programs to inconveniently timed workshops and busy family schedules.

THE CURRENT STUDIES

To develop children's prevention or intervention programs which are consistent with the unique needs of different segments of the very diverse communities we serve, potential participants must be involved in the design of the services they will receive. This study, therefore, employed choice-based conjoint analysis to consult parents regarding the design of services for preschool children. While conjoint analysis has been used extensively to study consumer preferences for a variety of goods and services, and has more recently been applied to the study of consumer views regarding the design of the health care services (Morgan, Shackley, Pickin, & Brazier, 2000), symptom impact (Osman, et al., 2001), treatment preferences (Maas & Stalpers, 1992; Singh, Cuttler, Shin, Silvers, & Neuhauser, 1998), and health outcome choices (Ryan, 1999; Stanek, Oates, McGhan, Denofrio, & Loh, 2000), the use of conjoint analysis to understand the preferences of mental health service users has been very limited (Spoth & Redmond, 1993).

METHODS

Sampling Strategies

In this study, we involved parents in the design of programs available to all families of young children (Offord, Kraemer, Jensen, 1998). While participants in parenting programs can provide us with information regarding factors influencing the decision to enroll, utilization research suggests that service users represent a select subset of the larger population of potential participants. The perspective of parents who were not reached by our marketing strategies, were uninterested in our advertising messages, or unable to attend the programs scheduled, are not represented in service user samples. Representative community samples of prospective users should be most useful in identifying marketing and service design attributes which would maximize utilization and enrollment. Our preference modeling studies, therefore, begin with community samples of parents whose children attended local child care (n = 434) and kindergarten programs (n = 299). As Orme (1998) has recommended, our sample of 600 allows for at least 200 cases per segmentation analysis group. To increase participation, we offered 100.00 for each center returning more than 50% of their surveys and an additional 400.00 for the center returning the greatest percentage of surveys. Return rates ranged from 37 to 69% across centers.

While prospective users can provide information regarding factors that would encourage the use of parenting programs, service users can provide a more informed perspective regarding the attributes of our programs which would improve adherence and reduce dropouts. These might include the learning process in our parenting services, the knowledge and skills of workshop leaders, and the utility of the strategies acquired. Our preference modeling studies, therefore, include a sample of 300 parents enrolled in existing programs. The results presented below summarize the response of 434 prospective service users.

SURVEY ATTRIBUTE DEVELOPMENT

We employed a three-stage approach to the development of attributes and attribute levels. We began from the perspective of readiness for change research, an empirical model of factors influencing the willingness to make personal changes. This research suggests that change proceeds in incremental stages. Most individuals confronted with the need to change begin at a *precontemplative stage* where the possibility of change has not been considered. While parents of a challenging child may appreciate the need to improve their child's behavior, they may not have anticipated the need to change their parenting strategies. At the *contemplative stage*, individuals consider the advantages and disadvantages of change. Parents might weigh the likelihood that a program would improve their child's behavior against travel time, duration of the program, and the risk that their child may defy their efforts to change management strategies. At the *preparatory stage* individuals begin planning the change process. They may, for example, seek information regarding parenting programs or enroll in a parenting workshop. Individuals make changes in the *action stage* and attempt to sustain changes during the *maintenance stage*. Research in this area suggests that movement through these stages in governed by decisional balance: the ratio of the anticipated benefits of

change over the logistical costs and potential risks of change. The application of readiness for change models to the utilization of children's mental health services suggests that parents will enroll in a program when they believe the benefits outweigh the logistical costs of participation. According to this model, we could improve the motivation to change by either reducing the logistical costs of participating or increasing the anticipated benefits of the program. Our model, therefore, included attributes addressing both the logistical demands of participation (course times, duration, locations, distance from home, availability of child care) and different messages regarding the potential benefits of change (e.g. improving skills or reducing problems). We derived potential cost and benefit attributes from both parental comments and previous research on factors influencing the utilization and outcome of children's mental health services (Cunningham, et al., 1995; 2000; Kazdin, Holland & Crowley, 1997; Kazdin & Wassell, 1999). Next, a group of experienced parenting program leaders composed a list of attribute levels that encompassed existing practice and pushed our service boundaries in significant but actionable steps. To inform our segmentation analyses, we included a series of demographic characteristics which epidemiological studies have linked to service utilization and outcome. These included parental education, income level, family status (single vs two parent), and child problem severity. Finally, we field tested and modified the conjoint survey. The program attributes and attribute levels included in this study are summarized in Table 1.

Table 1
Survey Attributes and Attribute Levels

Attribute	***Attribute Levels***
1. Time and Day Courses are Scheduled	The course meets on weekday mornings The course meets on weekday afternoons The course meets on weekday evenings The course meets on Saturday mornings
2. Course Location	The course is at a hospital or clinic The course is at a school The course is at a recreation center The course is at a parent resource center
3. Course Duration	The course meets once a week for 1 week The course meets once a week for 4 weeks The course meets once a week for 8 weeks The course meets once a week for 12 weeks
4. Distance to Meetings	It takes 10 minutes to get to the course It takes 20 minutes to get to the course It takes 30 minutes to get to the course It takes 40 minutes to get to the course
5. Learning Process	I would learn by watching a video I would learn by listening to a lecture about new skills I would learn by watching a leader use the skill I would learn by discussing new skills with other parents
6. Child Care	There is no child care There is child care for children 0-3 years of age There is child care for children 3- 6 years of age There is child care for children 0-12 years of age
7. Positively Worded Program Benefits	The course will improve my relationship with my child The course will improve my child's school success The course will improve my parenting skills The course will improve my child's behavior
8. Negative Worded Program Benefits	The course will reduce my child's difficult behavior The course will reduce conflict with my child The course will reduce the chances my child will fail at school The course will reduce mistakes I make as a parent
9. Leader's Experience	The course is taught by parents who have completed a similar course The course is taught by preschool teachers The course is taught by child therapists The course is taught by public health nurses
10. Evidence Supporting the Program	The course is based on the facilitator's experience as a parent The course is new and innovative but unproven The course is based on the facilitator's clinical experience The course is proven effective in scientific studies

SURVEY METHODS

Using Sawtooth Software's Choice-Based Conjoint module (version 2.6.7) we composed partial profile paper and pencil surveys from a list of 10 4-level attributes. As depicted in Table 2, for each choice task, participants read written descriptions of 3 parenting service options described by 3 attributes each. While a larger number of attributes per choice task increases statistical efficiency by reducing error in the estimation of model parameters, respondent efficiency decreases linearly as a function of the number of attributes per choice task (Patterson & Chrzan, 2003).

Increasing the number of choice tasks has been shown to reduce error in conjoint analyses (Johnson & Orme 1998).

Table 2
Sample Choice Task

Program 1	Program 2	Program 3
The course meets weekday morning	The course meets weekday evenings	The course meets Saturday morning
The course is 10 minutes from your home	The course is 30 minutes from your home	The course is 40 minutes from your home
The program will improve your parenting skills	The course will improve your relationship with your child	The course will improve your child's school success

Indeed, doubling the number of choice tasks included is comparable to doubling sample size (Johnson & Orme, 1998). While we piloted a survey with 20 choice tasks, we minimized informant burden by reducing our final survey to 17 choice tasks. With 7 different versions of this survey, efficiency approached 1.0, relative to a hypothetical partial profile orthogonal array.

To reduce the probability that parents would avoid effortful choices and to simplify our analysis, we did not include a no-response option. This is consistent with recommendations regarding the design of partial profile choice-based conjoint studies (Pilon, 2003). It has been suggested, however, that if informants who are not likely to enroll in parent training programs have different preferences than those who are more likely to participate, the absence of a no response option may generate utilities that do not reflect the perspectives of potential users (Frazier, 2003).

RESULTS AND DISCUSSION

Using Sawtooth Software's Hierarchical Bayes module, we calculated individual utilities for each member of the sample. Next, we computed a principal components latent class segmentation analysis using SIMCA-P software. We replicated this segmentation analysis using Latent Gold's latent class analysis program. The SIMCA-P plot depicted in the Figure 1 figure revealed two of the most strategically important segments emerging from this analysis: (1) a demographically low risk group of parents who were better educated and employed and (2) a demographically high risk segment with lower education and employment levels.

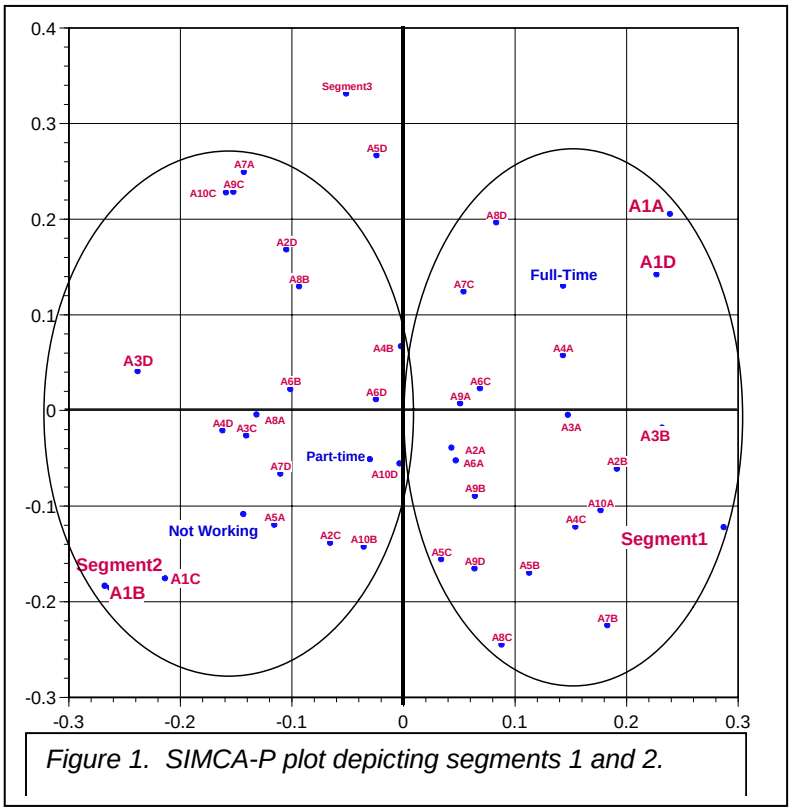


Figure 1. SIMCA-P plot depicting segments 1 and 2.

While the *proportion* of children from segment 2's high risk families experiencing mental health problems will be greater than the proportion of children in segment 1's low risk families, a majority of all childhood problems will emerge from the larger, lower risk segments of the population. This epidemiological principle, termed the *prevention paradox* (Rose, 1985), suggests that maximizing the population impact of prevention and intervention services requires the development of programs consistent with the preferences of both high and low risk segments of the community.

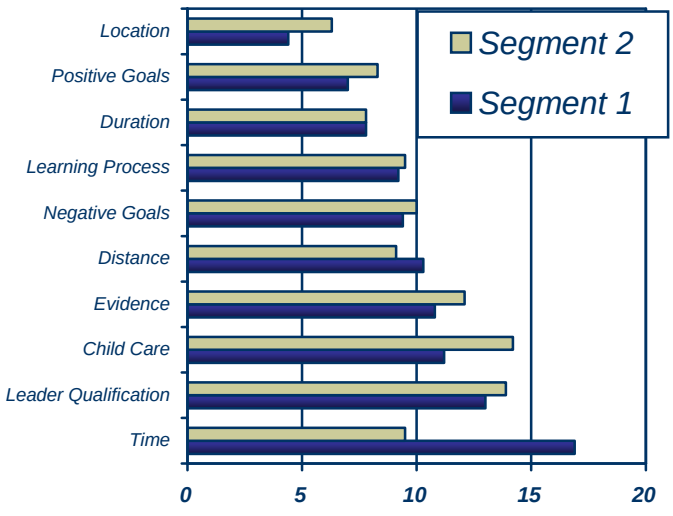


Figure 2. Importance scores for segments 1 and 2.

Figure 2 depicts importance scores for segments 1 and 2. As predicted by readiness for change research, parental preferences were influenced by a combination of the logistical demands and the anticipated benefits of participation. For both segments, logistical factors such as

workshop times, travel time to workshops, and the availability of child care exerted a significant influence on parental enrollment choices. The anticipated benefits of the program such as the qualifications of the leader and the level of evidence supporting the program were also important determinants of parental choice. The importance of logistical factors as barriers which may limit participation in parenting services is consistent with the reports of parents who did not use parenting programs in previous studies (Cunningham, et al., 1995; 2000) and those who fail to complete children's mental health services (Kazdin, Holland, & Crowley, 1997).

The utility values (zero-centered differences) for each attribute presented in Table 3 showed that different course times and advertising messages would be needed to maximize enrollment by parents from Segments 1 and 2. Figure 3, for example, suggests that, while segment 2's unemployed parents could flexibly attend either day or evening workshops, segment 1's employed parents expressed a strong preference for weekday evening or Saturday morning workshops. Segment 1 parents were more interested in building parenting skills, reducing parenting mistakes, and improving their child's success at school. Segment 2 parents, in contrast, were more interested in reducing behavior problems and improving their relationship with their child.

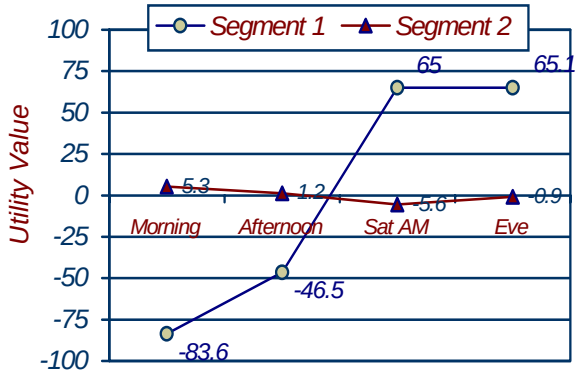


Figure 3. Workshop time utility values for Segments 1 and 2.

Table 3
Average Zero-Centered Utility Values for Segments 1 and 2

Attribute Levels	Utility Values	
	Segment 1	Segment 2
Logistical Demands of Participation		
Workshop Time		
Weekday Mornings	-83.6	5.3
Weekday Afternoons	-46.5	1.2
Weekday Evenings	65.1	-.9
Saturday Mornings	65.0	-5.6
Availability of Child Care		
No Child Care	-40.1	-55.4
Child Care for 0-3 Year Olds	-21.7	-8.9
Child Care for 3-6 Year Olds	19.2	8.5
Child Care for 0-12 Year Old	43.3	55.8
Travel Time to Workshop		
10 Minutes	39.2	28.6
20 Minutes	6.5	6.7
30 Minutes	11.5	2.0
40 Minutes	-57.1	-37.2
Location of the Program		
Hospital or Clinic	-11.1	-5.5
School	-4.1	-13.5
Recreation Center	4.l	-13.5
Parent Resource Center	9.3	25.4
Meeting Frequency		
Once a Week for a Week	9.0	-17.1
Once a Week for 4 Weeks	35.5	18.9
Once a Week for 8 Weeks	-13.8	-2.1
Once a Week for 12 Weeks	-29.8	.2
Benefits of Participation		
Leader's Skill and Experience		
Parents who Have Completed Course	-66.5	-57.8
Preschool Teachers	15.2	12.1
Public Health Nurse	4.2	2.5
Child Therapist	47.2	43.1
Evidence Supporting the Program		
New and Innovative But Unproven	-46.6	-56.3
Facilitators Parenting Experience	-10.2	9.0
Facilitators Clinical Experience	11.3	20.8
Scientific Studies	45.4	26.5
Learning Process		
Watch Video	-47.5	-39.8
Listen to a Lecture About New Skills	13.6	11.9
Watch a Leader Use Skills	19.7	12.8
Discuss New Skills with other Parents	14.2	15.4
Positive Focused Benefits		
Improve Relationship with Child	-7.2	12.1
Improve Child's School Success	14.8	-1.6
Improve My Parenting Skills	10.6	4.6
Improve My Child's Behavior	-18.2	1.6
Negatively Focused Benefits		
Reduce My Child's Difficult Behavior	-26.2	5.3
Reduce Conflict with My Child	-4.3	-1.6
Reduce Chances of School Failure	-3.7	-0.1
Reduce Mistakes I Make as Parent	15.7	-3.6

For both segments 1 and 2, utility values suggest that workshops of 4 weeks duration, located in a parent resource center, equipped with child care, 10 to 20 minutes from the family's home, would maximize utilization. Both segments 1 and 2 chose programs with professional versus paraprofessional leaders, an active learning process (discussion and modeling), and a program based on scientific evidence versus clinical experience. The importance of evidence in parental enrollment decisions is consistent with a shift by service providers to more evidence-based approaches to mental health programs.

As noted above, the day and time parenting workshops were scheduled exerted an important influence on workshop choices. While most children's mental health services are available during the day, parenting program schedules are service attributes which could be changed. We used Sawtooth Software's randomized first choice simulation module to predict preference shares for workshops scheduled in the afternoon, weekday evening, or Saturday morning. Figure 4 shows that, in comparison to weekday morning workshops, more than 80% of the preference shares were assigned to weekday evening or Saturday morning workshops.

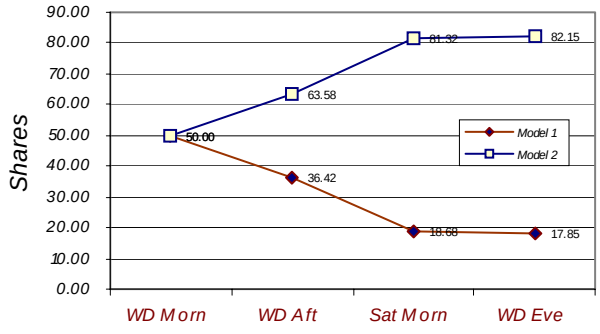


Figure 4. Sensitivity Analysis plotting preference shares for Weekday AM (WD Morn), Weekday Afternoon (WD Aft), and Saturday AM (Sat Morn) versus Weekday Morning workshops.

Given the importance which prospective users placed on travel time to parenting services, we simulated preference for parenting services located within 10 minutes versus 40 minutes of a parents home. As predicted, a considerable majority of preference shares were allocated to services were allocated to sessions within 10 minutes of the family's home.

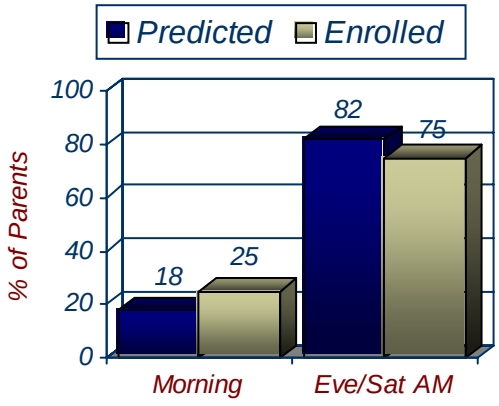


Figure 5. Comparing preference shares to utilization of clinic programs (further from home) versus redesigned community locations (closer to home).

Validating the Conjoint Analysis

We approached the validation of our conjoint analyses in three ways. First we examined predictive validity by comparing estimated patterns of utilization to field trial data in our own clinic. Next, we compared the predictions of our conjoint models to utilization

data from randomized trials of our parenting programs. Finally, we examined construct validity by testing several predictions regarding the process and outcome of parenting services whose design is consistent with the results of our preference modeling studies.

Aggregate utility values showed a strong preference for evening and Saturday morning versus weekday morning courses. Figure 5 presents Fall 2002 field trial data from our clinic. The percentage of parents enrolling in weekday morning versus weekday evening and Saturday morning courses is compared to the preference shares predicted from randomized first choice simulations. As our conjoint analysis predicted, a considerable majority of parents chose evening or Saturday versus weekday morning workshops.

Segmentation analysis revealed that, while the time programs were scheduled did not exert a significant influence on the choices of high risk Segment 2 families, Segment 1 parents showed a strong preference for evening and Saturday morning courses. In response to these findings, we added Saturday morning parenting courses to our Fall 2002's weekday morning and evening workshops. Saturday morning groups accounted for 17% of our program's capacity. Indeed, 17% of the parents participating in Fall 2002 workshops chose Saturday morning times. As our segmentation analysis predicted, all of the participants in this program were from segment 1's two parent families. Interestingly, for the first time in many years of conducting parenting programs, all fathers attended Saturday morning programs. Since utilization of parenting programs is consistently lower for fathers than mothers, our next series of conjoint studies will compare the program preferences of mothers and fathers.

Utility values for both Segment 1 and 2 parents revealed a strong preference for programs in close proximity to the homes of participants. To determine the validity of these findings, we reexamined utilization data from a previously conducted randomized trial (Cunningham, Bremner & Boyle, 1995). In this study, parents of 3564 children attending junior kindergarten programs completed a brief screening questionnaire regarding behavior problems at home. We randomly assigned parents of children who were more difficult than 93% of their peers to either a community-based parenting program located in neighborhood schools and recreation centers, a clinic-based parenting program located in a central clinic, or a waiting list control group. Community based programs were, on average, 17 minutes from the homes of participants. Clinic-based programs, in contrast, were located 36 minutes from their homes. As predicted, Segment 2 families assigned to the community condition were significantly more likely to enroll in programs than those assigned to clinic-based programs. This general preference for community-based groups located in closer proximity to the homes of participants was particularly pronounced in three groups of parents who are often members of our higher risk segment 2: parents of children with more severe problems, parents speaking English as a second language, and parents who are immigrants. Figure 6, for example, compares the percentage of the preference shares of segment 2 which were associated with programs located 10 and 40 minutes from the homes of families with utilization levels for immigrant families assigned to clinics (36 minutes from home) versus community (17 minutes from home) conditions.

While logistical factors and advertised benefits should influence enrollment in parenting programs, utility values suggested that the learning processes employed in each session of a program should influence ongoing participation. Although parenting skills are often taught didactically, utility values revealed a general preference for programs which involve group discussion rather than a lecture by the leader. This preference for discussions versus lectures and videotaped demonstrations was more pronounced among participants in parenting programs than in prospective user samples. As a measure of construct validity, we predicted that parents would respond more favorably to programs that are consistent with their preferences—more specifically, that participants would respond differently to parenting programs in which skills were taught via discussion and problem solving versus more didactic lectures or videotaped demonstrations. This prediction is consistent with the results observed in previously conducted studies. Cunningham, et al., (1993), for example, randomly assigned participants in a parenting program for staff/parents in residential treatment settings to one of two options: (1) a parenting program in which leaders taught skills more didactically, or (2) a program in which leaders used a problem solving discussion to teach new skills. The results of this study showed that, as predicted, participants in programs teaching new strategies via discussion attended a greater percentage of sessions, arrived late for significantly fewer sessions, completed more homework assignments, and engaged in less resistant behavior during homework reviews. Participants in discussion groups reported a higher sense of self-efficacy and were more satisfied with the program than those assigned to groups that were taught more didactically. These findings support the construct validity of our conjoint findings.

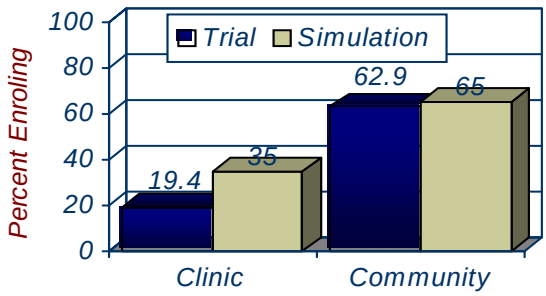


Figure 6. Comparing simulation (preference shares) for Segment 2 with utilization of redesigned community parenting workshops (closer to home) versus clinic services (further from home).

A large body of previous research suggests that parental depression, family dysfunction, and economic disadvantage, factors that place children at higher risk, reduce participation in traditional mental health services. As a test of the construct validity of our conjoint analyses, we hypothesized that participation in programs which were consistent with the preferences of parents would be less vulnerable to the impact of risk factors which reduce participation in more traditionally designed clinic services. We examined data from a trial studying the utilization of parent training programs by 1498 families of 5 to 8 year old children (Cunningham, et al., 2000). The parenting programs in this study were consistent with the logistical design preferences which emerged from our conjoint analysis. Courses were conducted in the evening, offered child care, were located at each child's neighborhood school, and were lead by a child therapist using a discussion/problem solving format. As predicted, logistic regression equations showed

that income level, family stress, family dysfunction, and parental depression were unrelated to enrollment (Cunningham, et al., 2000).

The validity checks reviewed above suggest that parenting programs which are consistent with user preferences improve utilization by high risk segment 2 families, improve attendance and homework completion, reduce resistance, and minimize the impact of family risk factors. As a final measure of construct validity, we would, therefore, predict that programs consistent with parental preferences would yield better outcomes. Cunningham et al., (1995) examined the outcome of a randomized trial comparing a community-based parent training program with more traditional clinic-based services. As we would predict, community-based programs consistent with the preferences of parents yielded larger effect sizes than a clinic based service. As a large group model, this community-based alternative was offered at $1/6^{th}$ the cost of individual clinic alternatives.

Applying Conjoint Analysis in Health Care Settings

We have applied the results of our conjoint analyses in several ways. First, knowledge of those logistical factors which are most important to parents has shaped the development of a new generation of family-centred parenting services. We have, for example, increased the availability of weekday evening and Saturday morning workshops which were critical to the participation of strategically important Segment 1 families. Interestingly, fathers, who are less likely to participate in parenting services, are much more likely to enroll in Saturday morning courses. In an effort to increase participation by fathers, this finding has prompted a series of follow-up studies examining differences in the service and advertising preferences of mothers and fathers.

The task of selecting brief, simple, relevant advertising messages describing complex parenting services is a challenge. In the past, we have composed these messages intuitively. We now use the results of our conjoint analyses to developed advertising messages highlighting those features of our programs that are consistent with the preferences of strategically important segments of our community. Our flyers, which are sent three times per year to families of all children enrolled in Hamilton area schools, emphasize that our services are scheduled at convenient times and locations, feature child care, and are offered in comfortable community settings. In addition, we include anticipated outcomes consistent with the motivational goals of different segments: parenting courses build parenting skills and reduce child behavior problems. Finally, given the importance that parents placed on the evidence supporting parenting service choices, we emphasize that these programs are supported by scientific research.

Benefits of Choice Based Conjoint in Health Service Planning

Choice-based conjoint provided a realistic simulation of the conditions under which parents make choices regarding parenting services. For example, the description of parenting service options in our conjoint analyses are similar to the format in which services are described in the flyers advertising our Community Education Service's many parenting courses and workshops.

The paper and pencil survey process employed in this study was also consistent with the data gathering strategies used in other Children's Hospital quality initiatives. Our patient satisfaction surveys, for example, are administered prior to or immediately after service contacts. Partial profile choice-based conjoint analysis could be completed in a 10 to 15 minute period before or after a service was provided. This ensured a high return rate and representative findings. Our validity analyses suggest that, while our brief partial profile surveys posed a minimum burden on respondents, the utilization of Hierarchical Bayes to calculate individual parameter estimates provided remarkably accurate and very useful estimates of shares of preference.

The inclusion of attribute levels reflecting existing service parameters provided an alternative source of user preference data regarding specific components of our programs. More importantly, choice-based conjoint allowed relative preferences for existing service options to be compared with actionable alternatives. For example, although most children's mental health services are available during the day and we have never offered weekend services, we included evening and Saturday morning workshops as alternative attribute levels. Segmentation analyses revealed that a strategically important subgroup of our participants preferred evening and Saturday morning parenting services. Moreover, fathers, a difficult to engage group of users, have consistently enrolled in our Saturday morning workshops.

Conjoint analyses allowed us to unpack the contribution of attributes which are often confounded in clinical trials. For example, we have suggested that the improved utilization observed when parenting services are offered in community locations, such as neighborhood schools, reflected the fact that these are more comfortable settings than outpatient clinics (Cunningham, et al., 1995; 2000). An alternative explanation is that community settings improve utilization by reducing travel time. The results of our conjoint analysis suggested that travel time provided a better explanation for the utilization advantages of community settings. Moreover, utility values suggested that the family resource centers included as an actionable alternative attribute level, are preferable to both schools and clinics.

Health service providers operate in a context of significant financial constraint. Before embarking on time consuming and expensive service delivery innovations, managers need convincing cost/benefit models. Randomized first choice simulations provide an empirical alternative to more intuitive approaches to service redesign. The consistency between our predictions, clinic field trials, and previously conducted randomized trials has provided convincing support regarding the predictive validity of these simulations and the utility of these methods.

Randomized controlled trials represent the gold standard in health service evaluation. Trials, however are typically limited to a small number of preconceived program alternatives, take 3 to 5 years to complete, and are conducted at considerable cost. If the services included in a randomized trial are poorly designed, advertised, or implemented, a potentially useful program might be rejected. It is difficult to repeat a trial with an alternative set of service parameters. Our conjoint analyses, for example, suggested that scheduling programs at times which do not reflect the preferences of strategic segments of the population, locating courses at inconvenient settings, or failing to offer child care

would limit enrollment and compromise the trial of parenting programs. Conjoint analysis simulations to optimize service delivery parameters and develop effective advertising messages will, therefore, be a preliminary step in the design of our next series of randomized trials.

Consumers are demanding a more important role in the design of the health services they receive (Maloney & Paul, 1993). Conjoint analysis represents an affordable, empirically sound method of involving users in the design of the health services they receive. Our findings suggest that the more patient-centred services that emerge when users are consulted via conjoint analysis may well improve health service utilization, adherence, and health outcomes.

REFERENCES

Barbour, R. (1999). The case for combining qualitative and quantitative approaches in health services research. *Journal of Health Services Research and Policy, 4*, 39-43.

Barkley, R. A., Shelton, T. L., Crosswait, C., Moorehouse, M., Fletcher, K., Barrett, S., Jenkins, L., & Metevia, L. (2000). Multi-method psychoeducational intervention for preschool children with disruptive behavior: Preliminary results at post-treatment. *Journal of Child Psychology and Psychiatry, 41,* 319-332.

Boyle, M. H., Cunningham, C. E., Heale, J., Hundert, J., McDonald, J., Offord, D. R., & Racine, Y. (1999). Helping children adjust-A Tri-Ministry Study: 1. Evaluation methodology. *Journal of Child Psychology and Psychiatry, 40, 1051-1060.*

Cunningham, C. E., Davis, J. R., Bremner, R. B., Rzass, T., & Dunn, K. (1993). Coping modelling problem solving versus mastery modelling: Effects on adherence, in session process, and skill acquisition in a residential parent training program. *Journal of Consulting and Clinical Psychology, 61,* 871-877.

Cunningham, C. E, Bremner, R., & Boyle, M. (1995). Large group community-based parenting programs for families of preschoolers at risk for disruptive behavior disorders. Utilization, cost effectiveness, and outcome. *Journal of Child Psychology and Psychiatry, 36,* 1141-1159.

Cunningham, C. E. (1997). Readiness for change: Applications to the design and evaluation of interventions for children with ADHD. The ADHD Report, 5, 6-9.

Cunningham, C. E., Boyle, M., Offord, D., Racine, Y., Hundert, J., Secord, M., & McDonald, J. (2000). Tri-Ministry Study: Correlates of school-based parenting course utilization. *Journal of Consulting and Clinical Psychology, 68,* 928-933.

Frazier, C. (2003). Discussant on modeling patient-centred children's health services using choice-based conjoint and hierarchical bayes. Sawtooth Software Conference. San Antonio, Texas.

Hawkins, J. D., von Cleve, E., & Catalano, R. F., Jr. (1991). *Journal of the American Academy of Child and Adolescent Psychiatry, 30,* 208-217.

Hundert, J. Boyle, M.H., Cunningham, C. E.. Duku, E., Heale, J., McDonald, J., Offord, D. R., & Racine, Y. (1999). Helping children adjust-A Tri-Ministry Study: II. Program effects. *Journal of Child Psychiatry and Psychology, 40,* 1061-1073.

Johnson, R. M & Orme, B. K. (1998). How many questions should you ask in choice based conjoint studies? *Sawtooth Software Research Paper Series.* Downloaded from: http://sawtoothsoftware.com/download/techpap/howmanyq.pdf

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,* 453-463.

Kazdin, A. E. & Mazurick, J. L. (1994). Dropping out of child psychotherapy: distinguishing early and late dropouts over the course of treatment. *Journal of Consulting and Clinical Psychology, 62,* 1069-74.

Kazdin, A. E. & Mazurick, J. L. & Siegel, T. C. (1994). Treatment outcome among children with externalizing disorder who terminate prematurely versus those who complete psychotherapy. *Journal of the American Academy of Child and Adolescent Psychiatry. 33,* 549-57.

Kazdin, A. E. & Wassell, G. (2000). Predictors of barriers to treatment and therapeutic change in outpatient therapy for antisocial children and their families. *Journal of Mental Health Services Research, 2,* 27-40.

Kazdin, A. E. & Wassell, G. (2000). Barriers to treatment participation and therapeutic change among children referred for conduct disorder. *Journal of Clinical Child Psychology, 28,* 160-172.

Maas, A. & Staplers, L. (1992). Assessing utilities by means of conjoint measurement: an application in medical decision analysis. *Medical Decision Making, 12,* 288-297.

Maloney, T. W. & Paul, B. (1993). Rebuilding public trust and confidence.

Gerteis, S. Edgman-Levital, J. Daley, & T. L. Delabanco (Eds.) Through the Patient's Eyes: Understanding and promoting patient-centred care. pp. 280-298. San Francisco: Jossey-Bass.

Morgan, A., Shackley, P., Pickin, M., & Brazier, J. (2000). Quantifying patient preferences for out-of-hours primary care. *Journal of Health Services Research Policy, 5,* 214-218.

Offord, D. R., Boyle, M. H., Szatmari, P., Rae-Grant, N., Links, P. S. Cadman, D. T., Byles, J. A., Crawford, J. W., Munroe-Blum, H., Byrne, C., Thomas, H., & Woodward, C. (1987). Ontario Child Health Study: II Six month prevalence of disorder and rates of service utilization. *Archives of General Psychiatry, 37,* 686-694.

Offord, D. R., Kraemer, H. C., Kazdin, A.R., Jensen, P.S., & Harrington, M. D. (1998). Lowering the burden of suffering from child psychiatric disorder: Trade-offs among clinical, targeted, and universal interventions. *Journal of the American Academy of Child and Adolescent Psychiatry,* 37, 686-694.

Orme, B. K. (1998), Sample Size Issues for Conjoint Analysis Studies. Available at www.sawtoothsoftware.com.

Osman, L. M., Mckenzie, L., Cairns, J., Friend, J. A., Godden, D. J., Legge, J. S., & Douglas, J. G. (2001). *Patient weighting of importance of asthma symptoms, 56,* 138-142.

Patterson, M. & Chrzan, K. (2003). Partial profile discrete choice: What's the optimal number of attributes. Paper presented at the Sawtooth Software Conference, San Antonio, Texas.

Pilon, T. (2003). Choice Based Conjoint Analysis. Workshop presented at the Sawtooth Software Conference, San Antonio, Texas.

Rose, G. Sick individuals and sick populations. International Journal of Epidemiology, 14, 32-38.

Ryan, M. (1999). Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilization. *Social Science and Medicine, 48,* 535-546.

Ryan, M. Scott, D. A. Reeves, C. Batge, A., van Teijlingen, E. R., Russell, E. M., Napper, M., & Robb, C. M. Eliciting public preferences for health care: A systematic review of techniques. Health Technology Assessment, 5, 1-186.

Sawtooth Software Technical Paper Series (2001). Choice-based conjoint (CBC) technical paper. Sequim, WA: Sawtooth Software, Inc.

Singh, J., Cuttler, L., Shin, M., Silvers, J. B. & Neuhauser, D. (1998). Medical decision-making and the patient: understanding preference patterns for growth hormone therapy using conjoint analysis. *Medical Care, 36,* 31-45.

Spoth, R. & Redmond, C. (1993). Identifying program preferences through conjoint analysis: Illustrative results from a parent sample. American Journal of Health Promotion, 8, 124-133.

Stanek, E. J., Oates, M. B., McGhan, W. F., Denofrio, D. & Loh, E. (2000). Preferences for treatment outcomes in patients with heart failure: symptoms versus survival. *Journal of Cardiac Failure, 6, 225-232.*

Vimarlund, V., Eriksson, H., & Timpka, T. (2001). Economic motives to use a participatory design approach in the development of public-health information systems. *Medinfo, 10.* 768-772.

CONJOINT ANALYSIS EXTENSIONS

COMPLEMENTARY CAPABILITIES FOR CHOICE, AND PERCEPTUAL MAPPING WEB DATA COLLECTION

JOSEPH CURRY
SAWTOOTH TECHNOLOGIES, INC.

Hundreds of successful studies have been carried out with "off–the-shelf" software for conjoint, choice-based conjoint, and perceptual mapping. As happens with many software applications, there are users with needs that go beyond what these packages offer. This can occur when technology outpaces software development and when projects require more complex capabilities. Web interviewing, in particular, creates research challenges and opportunities for studies that continually demand more advanced software features. These advanced needs occur in the data collection, analysis, and modeling phases of studies.

This paper addresses solutions for advanced data collection needs. It describes the work of three researchers who satisfied their studies' requirements by using the newest generation of Web questionnaire-authoring software to extend the capabilities of their choice-based and perceptual mapping software.

This paper characterizes the inherent limits of off-the-shelf conjoint, choice, perceptual mapping and other similar software and uses three case examples to illustrate ways to work beyond those limits with separate, complementary data collection software.

CHARACTERIZING THE LIMITS

In the years before Sawtooth Software's ACA (Adaptive Conjoint Analysis), CBC (Choice-Based Conjoint), and CPM (Composite Product Mapping) were created, Rich Johnson, Chris King and I worked together at a marketing consulting firm in Chicago that collected conjoint and perceptual mapping data using computer interviewing. We custom-designed the questionnaire, analysis, and modeling components for most of our studies, for many reasons: to get around the basic assumptions of the techniques; to deal with complexities of our clients' markets; to make use of respondent information during the interview; to combine techniques—all to provide the best information for our clients. We were able to do this customization because we had the flexibility of writing our questionnaires, analysis routines, and models using programming languages such as Basic and FORTRAN.

Based on the experience we gained from that work, we went on to develop commercial software packages for conjoint analysis and perceptual mapping. That software achieved widespread acceptance among researchers for several reasons: First, it provided the complete set of data collection, analysis, and modeling tools needed to employ the techniques. Second, it was easy to use. Third, it was relatively foolproof, since we were careful to include only what we knew worked and was the least prone to misuse. Finally, it significantly decreased the cost of conducting conjoint and mapping

studies and broadened the range of product categories and situations to which the techniques could be applied.

Products that are easy to use and that ensure quality results necessarily have limits in the scope of capabilities that can be included. For most studies, these limits are not significant, but nearly all experienced users have bumped up against them. Technology and users' increasingly complex research needs often race ahead of the software designers.

To set the stage for describing how some users get past these limits, this paper characterizes the techniques using the two highly correlated dimensions shown in *Figure 1*. The horizontal dimension represents how closely the techniques mimic reality. The vertical dimension represents the techniques' ability to deal with market complexity. The dimensions are correlated, since techniques that more closely mimic reality and are better able to deal with market complexities generally yield results that have greater predictive validity.

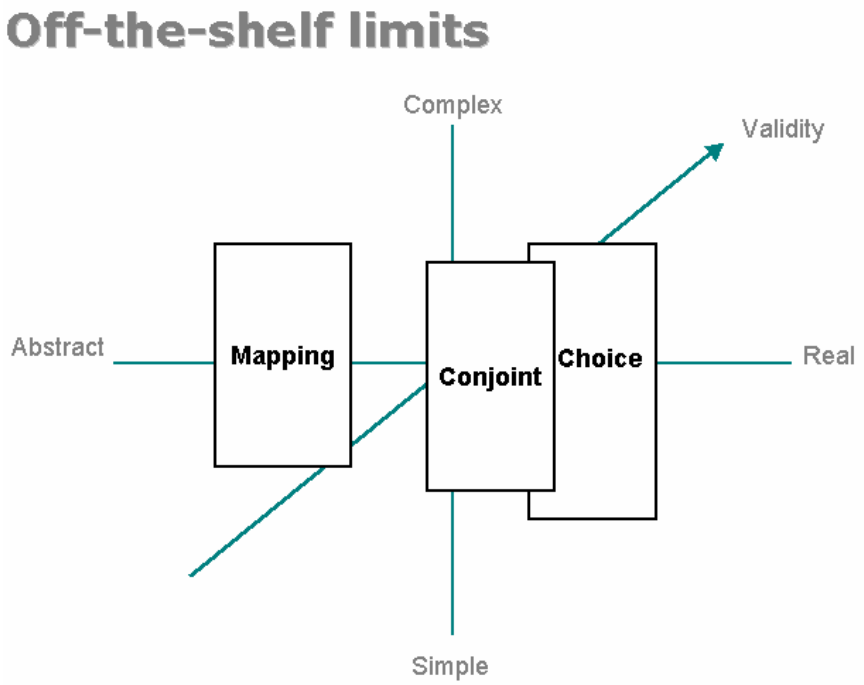


Figure 1

The techniques are represented schematically in the figure as boxes. The edges of the boxes represent the limits of the techniques. The figure illustrates that choice mimics reality better than conjoint, and that both choice and conjoint mimic reality better than perceptual mapping. All three techniques are shown as having roughly the same limits with respect to dealing with complexity.

As *Figure 1* implies, the off-the-shelf packages for these advanced techniques allow us to go only so far. When our studies require that we go farther, we have two choices: wait for future releases or create our own data collection, analysis and modeling

extensions. This paper focuses specifically on overcoming the off-the-shelf software's current data collection limits for Web interviewing by creating extensions of the techniques with advanced data collection software.

The two factors that make creating extensions for collecting complex conjoint, choice and perceptual mapping data on the Web possible are shown in *Figure 2*. The first is the availability of questionnaire-authoring tools for customizing Web-based questionnaires. These tools replace and sometimes enhance the built-in data collection tools available in the conjoint, choice, and perceptual mapping software packages.

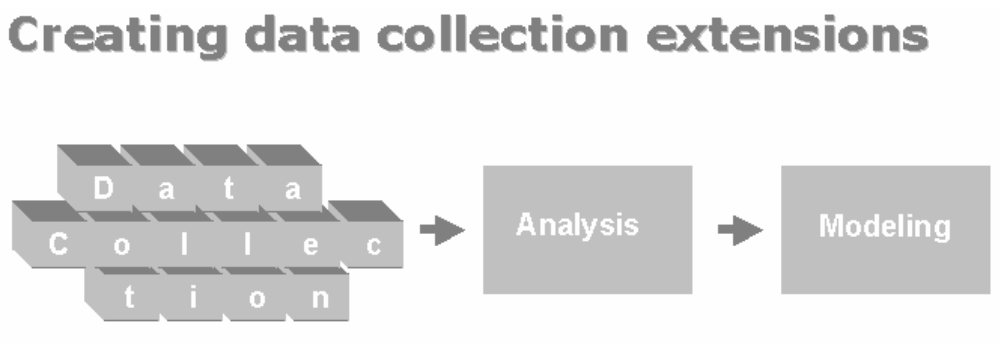


Figure 2

Second, the advanced-techniques software generally allows data collected by other methods to be imported for analysis and modeling.

The following examples illustrate how three Sawtooth Technologies' Sensus Web interviewing software users successfully created data collection extensions for their choice and perceptual mapping projects—studies with requirements that were too complex for their existing choice and perceptual mapping software tools. The product categories in these examples have been disguised to preserve clients' confidentiality.

CONDITIONAL PRICING

Richard Miller of Consumer Pulse, Inc. (Birmingham, Michigan) needed to do a choice-based conjoint study of a market with a complicated price structure. To test some new concepts for his client, Miller needed to create choice tasks that matched the pricing complexities of that market. The price ranges for the task concepts had to be conditioned on the levels of other attributes in the concepts. An example of this type of situation—which is becoming increasingly more common—is shown in *Figure 3*.

Conditional pricing

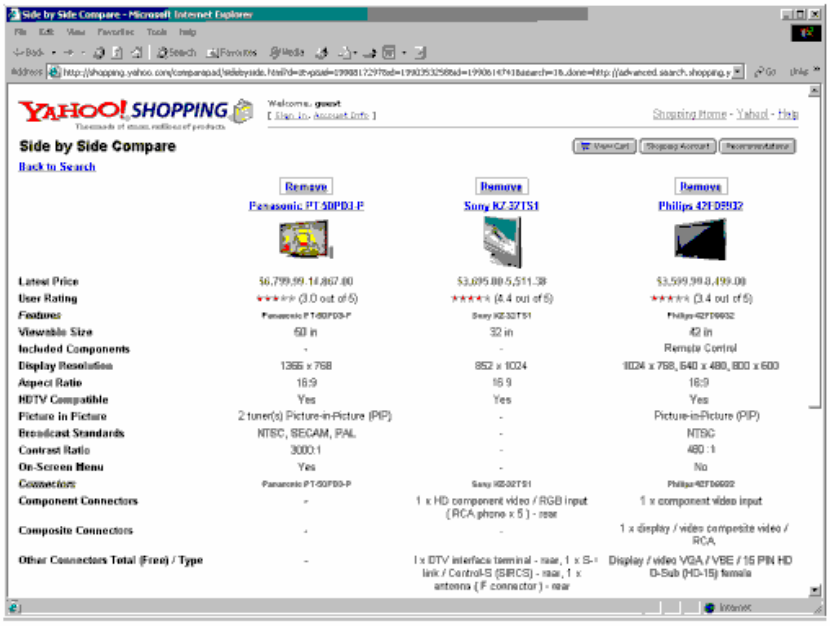


Figure 3

Here, a Web site shopping page displays a side-by-side comparison of the features of three high definition televisions (HDTV's). In this example, the price ranges of the HDTV's depend on the brand, the form of the technology (plasma, direct view, or projection) and screen size.

To collect this type of information, Miller needed to use choice-based conjoint with conditional pricing. This required the construction of a conditional pricing table, such as the one shown in *Figure 4* below. For each combination of brand, technology and screen size there is a range of prices with a high, medium and low value.

Conditional pricing table

	Panasonic			Philips			Sony		
	Low Price	Likely Price	High Price	Low Price	Likely Price	High Price	Low Price	Likely Price	High Price
Plasma									
32"	$3,495	$4,495	$5,495	$3,995	$4,995	$5,995	$4,995	$5,995	$6,995
37"	$4,495	$5,495	$6,495	$4,995	$5,995	$6,995	$5,995	$6,995	$7,995
42"	$5,495	$6,495	$7,495	$5,995	$6,995	$7,995	$6,995	$7,995	$8,995
47"	$6,495	$7,495	$8,495	$6,995	$7,995	$8,995	$7,995	$8,995	$9,995
Direct View									
32"	$1,095	$1,295	$1,495	$1,195	$1,395	$1,595	$1,295	$1,495	$1,695
37"	$1,295	$1,495	$1,695	$1,395	$1,595	$1,795	$1,495	$1,695	$1,895
42"	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
47"	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Rear Projection									
32"	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
37"	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
42"	$1,995.00	$2,195.00	$2,395.00	$2,095.00	$2,295.00	$2,495.00	$2,195.00	$2,395.00	$2,595.00
47"	$2,295	$2,495	$2,695	$2,395.00	$2,595.00	$2,795.00	$2,495.00	$2,695.00	$2,895.00

Figure 4

Miller uses the conditional price table as follows: He starts by entering the table into his choice software to generate task sets of fixed designs. Each of his tasks has three product concepts and a "none" option. The price ranges for the concepts within the tasks are initially labeled as high, medium, or low. For each concept in a task, the software looks up the range of prices from the conditional pricing table based on the levels of brand, technology and screen size that make up the concept description. It then substitutes the price within that range for the label, based on whether the price level for the concept is the high, medium, or low price point in that range. An example of a task constructed in this way is shown in *Figure 5*.

Figure 5

Miller could implement this design for the Web only by creating the questionnaire with advanced data collection software. The questionnaire included randomizing the tasks within sets of choice tasks, and randomizing choice task sets across respondents. The entire process from questionnaire design to analysis is shown in *Figure 6*.

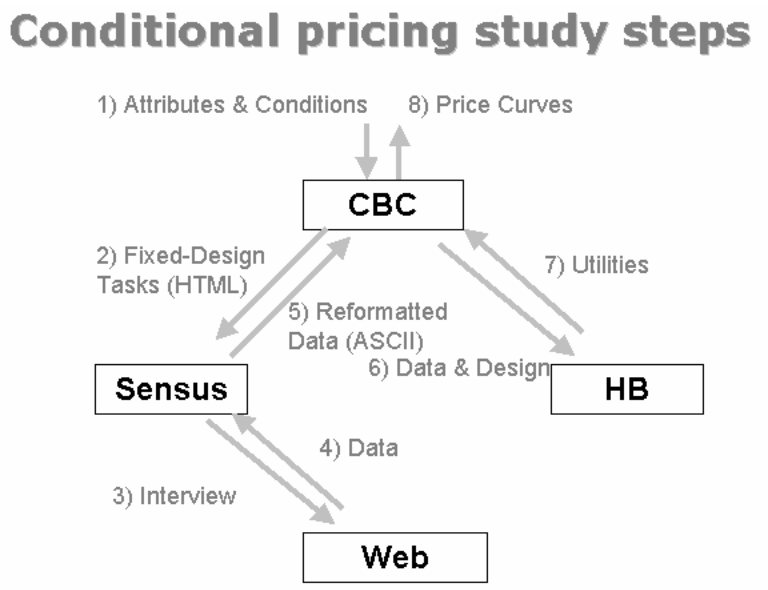


Figure 6

In *Figure 6*, "CBC" is Sawtooth Software's Choice-Based Conjoint software and SMRT simulator, and "HB" is Sawtooth Software's Hierarchical Bayes utility estimator.

Referring back to *Figure 1*, Miller extended the limit of choice-based conjoint in the vertical direction, increasing the complexity that the technique could handle, beyond the limits of the off-the-shelf system. What did he accomplish? Miller states that his price curves are more realistic, the precision with which he can model changes in price is increased significantly, and the overall results of market simulations are more credible.

VISUALIZATION OF CHOICE TASKS

Dirk Huisman of SKIM Analytical (Rotterdam, The Netherlands) has an important client in the consumer package goods industry. He and his client maintain that in reality most communications involved with routine and impulsive purchases of fast-moving consumer goods are non-verbal. To capture this important aspect of consumer behavior when doing market studies, they think it is essential that tasks in choice-based interviews mimic reality as closely as possible. The simulated store-shelf version of the choice task in *Figure 7* clearly mimics reality better than a task that presents the concepts using only words.

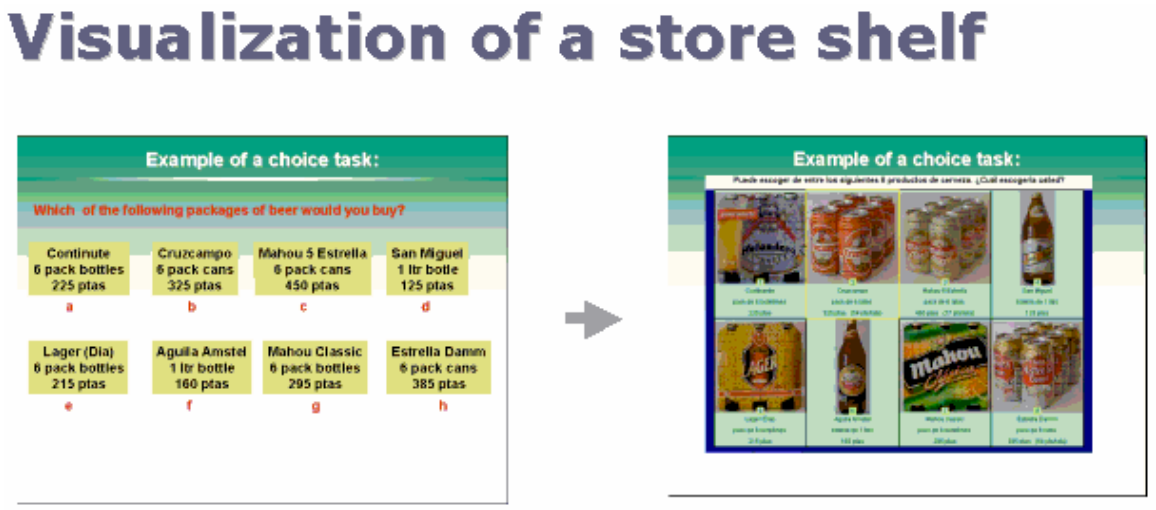


Figure 7

Huisman generates choice tasks for testing the impact of attributes that convey non-verbal information by combining images for those attributes in real time to form product concepts. He uses a generalized form of the conditional pricing scheme (described in the previous example) to ensure that the concepts he constructs make sense visually. Figure 8 shows an example of a choice task generated in this way for electric toothbrushes.

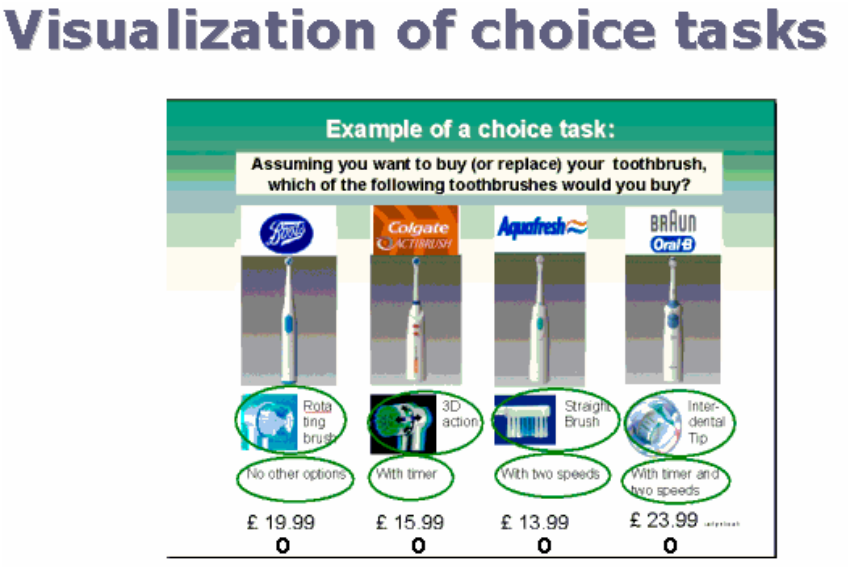


Figure 8

The steps Huisman follows in executing studies that use visualized choice tasks are shown in *Figure 9*. With the exception of how the Web interview is created, the steps are the same as those Miller follows. Miller uses CBC to generate fixed choice-task designs and then enters them into Sensus Web software. Huisman uses Sensus Web to create an interview template that generates the choice tasks during the interview, with the resulting advantage of being able to test for interactions. For any given study, Huisman simply

specifies the lists of attributes and conditions for constructing the choice-task concepts using Sensus Web, and imports the sets of images for the questionnaire.

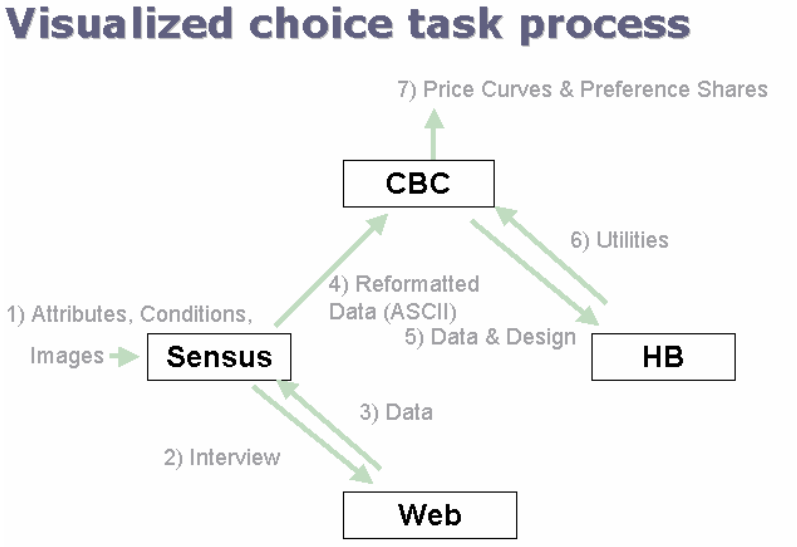


Figure 9

Huisman plans to conduct a study to test a number of hypotheses associated with his visualization-of-choice-task approach. For example, he'll explore whether including non-verbal information in choice tasks leads to better share predictions, less sensitivity to price, and a higher impact of promotions.

RANDOMIZED COMPARATIVE SCALES

The final example uses perceptual mapping based on multiple discriminant analysis. Tom Pilon of TomPilon.com (Carrollton, Texas) created a Web version of a longitudinal mapping study that he had been conducting for his client using disk-by-mail. It was important that the Web version be as close to the disk-by-mail version as possible, so that the change of the interview modality did not cause data discontinuities.

The length of the interview was another critical issue. The disk-by-mail interview included 120 product ratings and respondents were asked for each rating as separate questions. Pilon wanted to make the Web version more efficient by asking respondents for multiple ratings in a single question, something that was not possible with the software used for the disk-by-mail version.

Pilon used Sensus Web to create a three-part questionnaire. In the first part, respondents rated their familiarity with a number of products (PR firms in our example) using a five-point semantic rating scale (*Figure 10*). In the second part, respondents rated the importance of a number of attributes for selecting a product (also *Figure 10*).

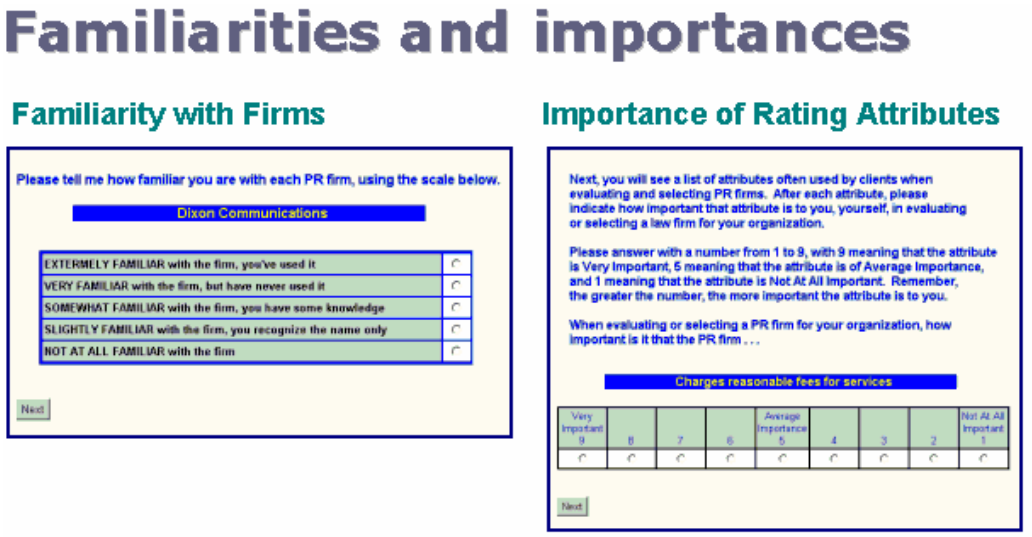
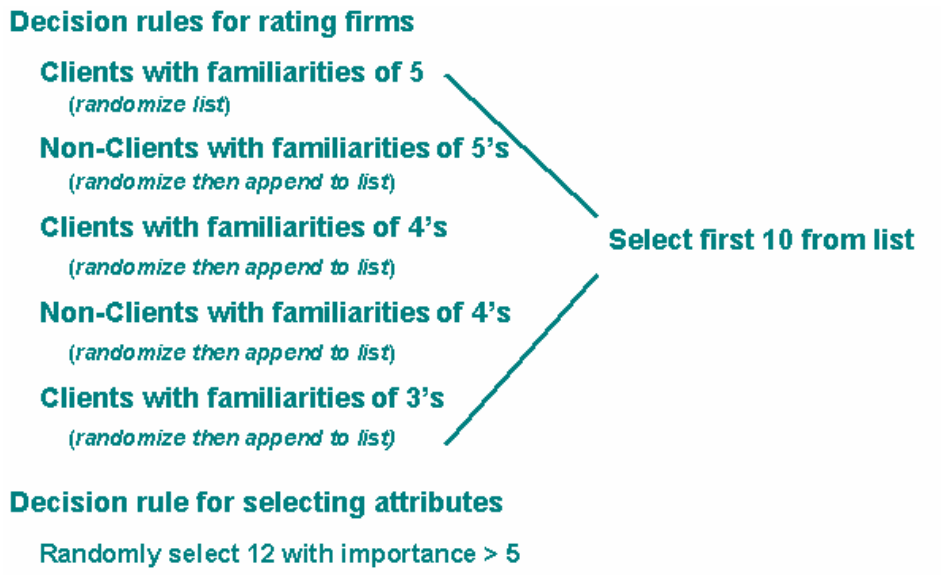


Figure 10

Based on the respondent's familiarity with the firms and the importance ratings of the attributes, Pilon had to invoke the following decision rules to determine which firms and attributes the respondent would use in the third part, the ratings section.



In the ratings section, Pilon wanted respondents to rate all of the firms within each attribute and he wanted to randomize the firms and attributes. To make completing the ratings efficient and reliable, he wanted respondents to be able to enter all of their ratings for a given attribute at the same time (*Figure 12*), rather than in a sequence of individual questions.

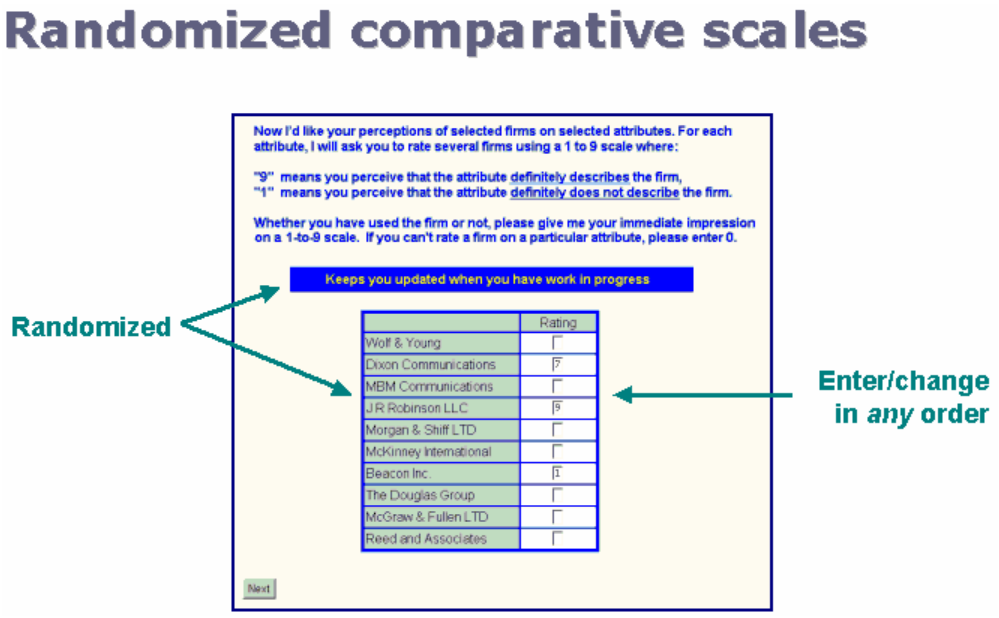


Figure 12

The process for carrying out Pilon's design is shown in *Figure 12*. The list of firms and rating attributes were entered into Sensus Web for creating the questionnaire and into Sawtooth Software's CPM software for use in the analysis phase. A questionnaire template was set up using Sensus Web for administering the three parts of the interview. The attributes and firms were entered into the template, and the questionnaire was deployed to the Web for administration.

Once the data were collected, they were downloaded to Sensus Web for export to CPM. CPM performed the discriminant analysis and created the maps.

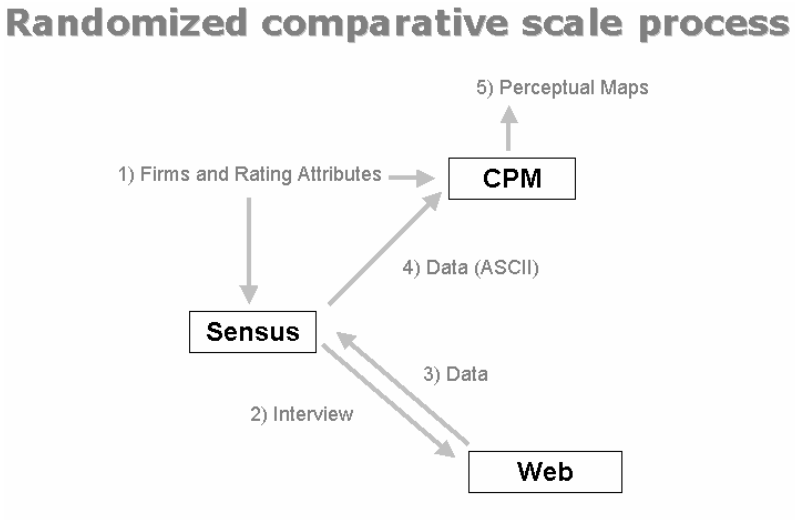


Figure 13

Pilon maintains that being able to implement the questionnaire where respondents can rate all firms at once increases measurement reliability, shortens the administration of the questionnaire, and results in better overall perception measurement.

CONCLUDING REMARKS

In the late 1980's I wrote a paper, *Interviewing By PC, What You Couldn't Do Before*. In that paper I described how researchers could use the PC and PC-based research tools to provide their clients with results that were more strategic and insightful. These PC tools have migrated to the Web, and today the opportunities for advanced strategy and insight are even greater.

The tools being ported from the PC to the Web have undergone nearly two decades of testing, evolution and refinement. They let us deal with far greater complexity and let us mimic reality more closely than even the tools of just five years ago.

Sometimes, software systems cannot keep up with technology advancements and research demands. Combining tools, we can extend capabilities and overcome limitations. Researchers don't have to wait; they can remain relevant *and* innovate. The tools for quality, advanced research are available now, as the works of Miller, Huisman and Pilon illustrate.

BRAND POSITIONING CONJOINT: THE HARD IMPACT OF THE SOFT TOUCH

MARCO VRIENS AND CURTIS FRAZIER
MILLWARD BROWN INTELLIQUEST

SUMMARY

Including brand-positioning attributes in a conjoint study has been complicated and has not been typically pursued by users of conjoint. Brand positioning attributes do play an important role in consumer choice behavior. For example, when considering to buy a car, concrete attributes like price, power of the engine, extras (airbags), design, etc. will have an impact on consumers' choices, but perceptions of the brand in terms of "Reliability," "Safety," "Sporty," "Luxurious," etc. will also play a role. Brand positioning attributes cannot be included in a conjoint study directly because it is difficult to define such attributes (or perceptual dimensions) in terms of concrete attribute levels (which is needed in order to design the conjoint experiments), and consumers (probably) already have perceptions of how the various brands perform on such positioning dimensions, making it difficult for them to engage in a task where they need to ignore their own perceptions as they would have to do in a typical conjoint exercise. In this paper we describe a practical approach to deal with the issue that we have found to work very well in practice.

INTRODUCTION

Conjoint analysis is probably the most popular tool in marketing research today for assessing and quantifying consumer preferences and choices. The conjoint approach can be used for a variety of marketing problems including product optimization, product line optimization, market segmentation, and pricing. Usually market simulations are performed to facilitate decision making based on the conjoint results. Recent developments, such as the discrete choice modeling, latent class analysis, hierarchical Bayes techniques, and efficient experimental designs, have made new application areas possible, such as studying tradeoffs when different product categories are involved, etc.

An important attribute in consumer tradeoffs is brand: a typical conjoint study will include a series of alternatives that are defined on a number of concrete attributes, and price. From such a design we can assess the value (utility) of the included brand names. Hence, conjoint can be used for brand equity measurement. Concrete attributes are often the basis for product modification or optimization, while more abstract attributes are often the basis for brand positioning. However, to include more abstract brand-positioning attributes in a conjoint study so that these attributes can become part of predicting preference shares of hypothetical market situations, i.e. including brand-positioning attributes in market simulations, has been more complicated and has not been typically pursued by users of conjoint.

In many product categories it is difficult to position a brand and maintain a strategic advantage based on concrete attributes alone. Any advantage, as perceived by customers, that is the result of specific concrete attributes can often times fairly easily be copied or imitated by the competition unless there exists a patent to prevent this. Brand positioning attributes are much better suited for creating a sustainable advantage, and it has been shown that they do play an important role in consumer choice behavior. We encounter brand-positioning attributes in consumer markets: for example, when considering to buy a car, concrete attributes like price, power of the engine, extras (airbags), trunk volume, warranty, design, etc. will have an impact on consumers' choices, but perceptions of the brand in terms of "Reliability," "Safety," "Sporty," "Luxurious," etc. will also play a role. We also encounter brand-positioning attributes in many business-to-business technology markets for products such as servers, enterprise software, storage solutions, etc. For example, for buyers of business/enterprise software, concrete attributes like price, total cost of ownership, licensing terms, etc. will play a role, but so will attributes that are brand-related such as "This is a brand that knows me," "This is a pro-active brand," and "This is a brand that is innovative." Concrete attributes can be evaluated in the choice situation, be it a hypothetical choice situation in a survey, or be it in a real-life choice situation in a store when comparing alternatives. Brand positioning attributes (abstract) attributes are more likely to be retrieved from memory. Prior to the choice situation a consumer may have been exposed to brand attribute information because they used the brand, heard about it from others, or saw it in advertising. Such exposures will lead to brand information stored in memory as abstract attributes (see Wedel et al. 1998).

Hence, there are three reasons why brand-positioning attributes cannot be included in a conjoint study directly, and why the integration of brand positioning attributes in conjoint analysis is problematic:

- First, it is difficult to define such attributes (or perceptual dimensions) in terms of concrete attribute levels (which is needed to design the conjoint experiments),
- Second, consumers (probably) already have perceptions of how the various brands perform on such positioning dimensions as a result of previous exposures. This makes it difficult for them to engage in a conjoint task where they need to ignore their own perceptions as they would have to do in a typical conjoint exercise, and
- Third, often by including both concrete and more abstract attributes, the sheer number of attributes becomes a problem in itself: the conjoint task would become prohibitively difficult or fatiguing.

The above reasons have prevented the conjoint approach to be fully leveraged for the purposes of brand equity and brand positioning research. As a result the research literature has developed a separate class of techniques to deal with brand positioning attributes such as multi-dimensional scaling, tree structure analysis, etc. However, such methods do not allow the research to understand the joint impact of changes in both concrete attributes and brand positioning dimensions upon consumer brand choices. To understand the impact of brand positioning attributes it needs to be a part of a trade-off methodology. An early pioneering paper by Swait et al. (1993) demonstrated how discrete choice conjoint is a powerful method to measure brand equity in terms of what

consumers are willing to pay extra for a brand relative to competing brands. Park and Srinivasan (1994) discussed how a self-explicated approach could be used to measure brand equity and to understand the sources of brand equity (in their approach attribute-based sources and non-attribute based sources). Neither paper, however, discusses how to assess the impact of changes in performance on soft attributes on hard measure such as preference shares as derived with conjoint. In this paper we describe a practical approach to deal with the issue that we have found to work very well in practice.

CONJOINT BRAND POSITIONING

Our approach is conceptually shown in Exhibit 1, and involves the following steps:

1. Identify the key decision attributes that can concretely be defined (e.g. brand, price, etc.). Using this set of concrete attributes, a conjoint experiment is designed to derive individual-level brand utilities. In its simplest form, we could design a brand-price trade-off exercise. More complicated designs, i.e. involving more attributes, can be used as long as they include brand name. Key here is that the data must be analyzed in such a way as to achieve individual level brand utilities. When a traditional ratings-based conjoint is used one can usually estimate directly at the individual-level, when a choice-based conjoint is used we need to apply hierarchical Bayesian techniques to obtain the required individual-level utilities,

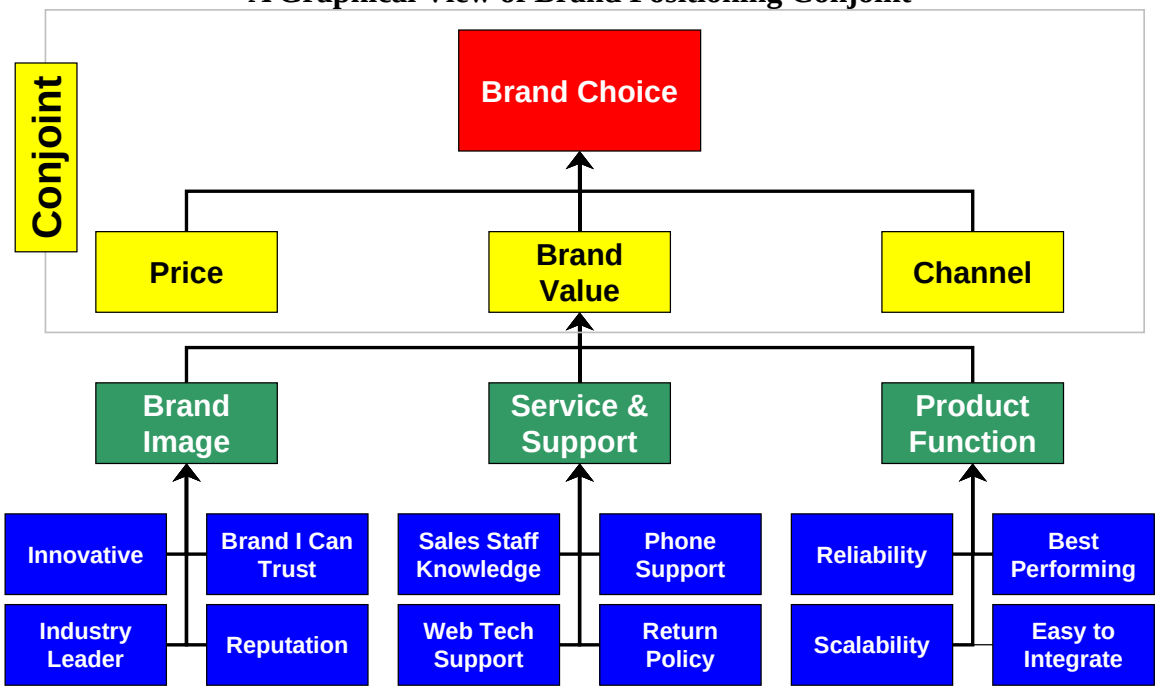
2. Identify the brand positioning attributes that are potentially important for the positioning of the brands and that are expected to play a role in consumer decision-making. The respondents evaluate all potentially relevant brands on these more abstract dimensions,

3. Use the individual-level brand utilities as the dependent variable in a linear or non-linear regression model with the performance perceptions of the abstract brand positioning attributes as independent variables. Essentially, the brand utilities become a dependent variable and are modeled as a function of brand positioning attributes. By asking respondents to evaluate each of the brands tested in the conjoint on a series of brand performance questions, we can construct a common key drivers model. The difference from a standard key drivers model is that rather than modeling overall brand value from a stated brand preference/value question, we are modeling derived brand value from the conjoint stage. The conjoint analysis and regression analysis can be executed simultaneously by specifying a hierarchical Bayesian model where the brand parameters are specified to be a function of the brand positioning perceptions, and where for the non-brand conjoint parameters a normal distribution is assumed.

4. Use the relative regression weights to calculate pseudo-utilities for the different levels of the brand positioning attributes, and

5. Utilize the comprehensive consumer choice model to build a simulator that allows the manager to evaluate different scenarios, including those that involve anticipated or planned changes in the brand positioning perceptions.

Exhibit 1.
A Graphical View of Brand Positioning Conjoint



AN ILLUSTRATION

We have tested this approach in a variety of situations including consumer and B-to-B markets, and on hardware and software technology products. Our illustration is derived from a recent study where respondents did a web-based interview that included 14 discrete choice tasks (presented in random order). In each of these tasks, respondents were shown profiles defined on only brand and price. Respondents were asked which, if any, of the options shown they would actually purchase. The "none of these" option is important because it allows estimation of the minimum requirements for a product to become considered. The conjoint exercise was followed by a series of brand positioning and relationship attributes. Respondents were asked familiarity with each of the brands tested in the conjoint. For those brands with sufficient familiarity, they were asked to indicate how they perceived the brands on these soft-touch attributes. These attributes included questions about brand reliability and performance, as well as less tangible attributes, such as "a brand I trust." The full list of brand positioning attributes is presented in Exhibit 2.

Exhibit 2:
Importance of Brand Positioning Attributes

Example Results Based on Studies of 3 Products in Business-to-Business and Consumer Spaces (6 studies total)		
Brand Positioning Attributes	*Minimum Importance Found*	*Maximum Importance Found*
Brand	29%	58%
Reliability	4%	12%
Performance	1%	11%
Service and support	6%	14%
Value for the price	5%	15%
Products with latest technology	0%	14%
Is a market leader	10%	19%
Product meets my needs	8%	38%
Is a brand that I trust	9%	16%
Stable, Long-term player	9%	19%
Easy-to-use	3%	11%
Appealing design/style	3%	14%

The analyses comprised three stages as discussed in the previous section. In the first stage the conjoint choice data are analyzed using hierarchical Bayesian methods. This methodology means that we are able to obtain unique conjoint utilities for each respondent in our sample. These unique utilities are what allow us to estimate the second piece of our model. In the second stage we first need to merge the brand utilities back into the survey data. At this point the analysis can take two different directions. Stage two can be done either at the market level or can be done brand-specific. Analysis at the market level means we estimate the relationships between brand positioning perceptions and brand utilities across all brands: in other words we assume that the importance of brand positioning attributes is the same for all brands. The model for this data format would specify that brand utility is a function of brand positioning attributes such as 'Trust," "Performance," "reliability," etc.

The alternative strategy is analysis at the brand level. There is no need for stacking the data, because the individual is the appropriate level of analysis. Rather than a single equation that applies equally well to each brand, we create unique equations for each brand. Hence the brand utility for brand 1 is modeled as a function of brand positioning attributes, the brand utility of brand 2 is modeled this way, etc.

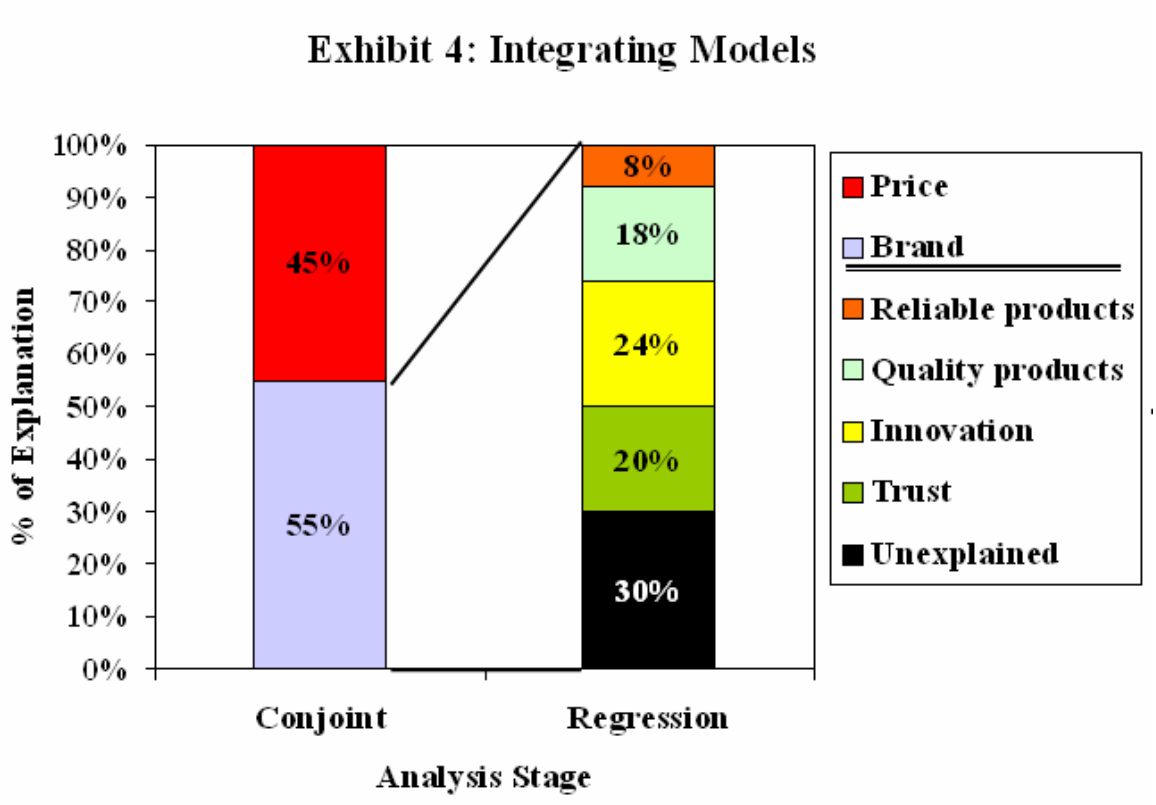
Analysis at the market level has several advantages. The most important of these involve sample size and reporting. In terms of sample size, the stacking process essentially replicates the data in such a way that our final regression analysis has k x N cases, where k equals our number of brands and N equals our number of respondents. Analysis at the market level also has the advantage of being easier to report/interpret. Rather than having attributes with differential importances, depending on which brand is being discussed, the analysis at the market level illustrates the importance across brands.

Exhibit 3

Data Format for Analysis at Market Level						
ID	**Brand #**	**Brand Utility**	**Trust**	**Performance**	**Reliability**	**Innovative**
1	**1**	1.2	6	6	4	5
1	**2**	0.3	5	6	4	5
1	**3**	-1.5	5	3	3	4
2	**1**	-0.7	4	4	2	4
2	**2**	0.3	4	5	5	3
2	**3**	0.4	5	5	5	4

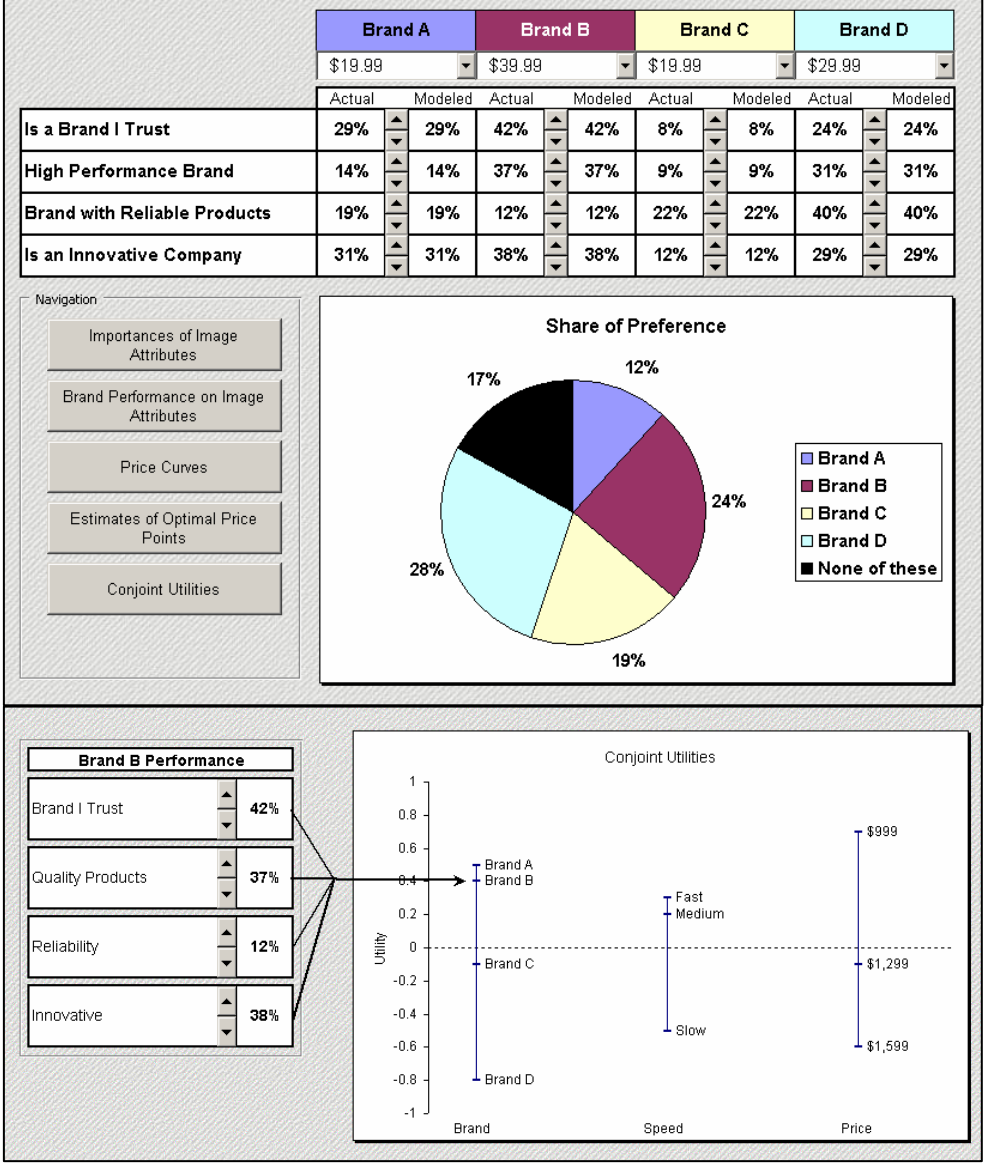
Although analyzed using a smaller effective base size, analysis at the brand level has some important advantages. Foremost among these is that it does not impose the assumption that the equations are consistent across brands. This assumption, while valid in some markets, is tenuous, at best, in others. For example, the utility of Apple/Macintosh may be driven more by *fulfills a need* or *compatible with my other systems*, whereas the utility of Gateway may be more driven by *reliability* or *performance*. Alternatively, the brand equity of smaller brands might be driven largely by awareness and familiarity, while larger brands may be driven by brand image.

In the third stage of the analysis we integrate the results from the first two stages. The basic process for model integration has already been discussed – using conjoint results as inputs into the hierarchical regression models. In this stage we re-scale the regression coefficients to the same scale as the conjoint utilities. The process of re-scaling is relatively simple. Attribute importances for the conjoint stage are calculated in the standard way. The importances in the regression stage are calculated in the standard way, except that they are scaled to sum up to equal the adjusted R^2. Once the model integration is completed we have a set of (pseudo) utilities that can be used as input for an integrated decision support tool. We note that the brand positioning perceptions don't predict brand utility completely, i.e. the regression equation has an explained variance of less than a 100%. We have found that the predictive power can range from high (e.g. over 80% explained variance) to low (e.g. 20% explained variance). See exhibit 4 for this.



Decision support tools are fairly common in conjoint studies because they enhance and facilitate how the product/market managers can study and work with the results. An example of how a simulator tool looks when brand-positioning attributes are included is shown in exhibit 5. However, by allowing the user of the tool to manipulate not only the tangible product features, but also brand positioning and relationship attributes, a more complete marketing picture is created. As the user manipulates the brand positioning attributes, these changes are adding, or subtracting, value from the utility for the brand(s). This re-scored brand utility value is then used in the share of preference calculations.

Exhibit 5: Example Decision Support Tool



We have applied our approach in both consumer and business-to-business markets. We can't present the results of individual studies because of their proprietary nature. However, in exhibit 2 we show the ranges we have found for the relative importance estimates of series of commonly used brand-positioning attributes.

The technique described in this paper extends conjoint analysis by allowing for a second set of research questions to be asked. Through conjoint, we know the answers to questions like "*what* do respondents want?" The technique described here allows to answers to questions like "*why* do they prefer it?" and "what can we do to make it more attractive?"

Our approach is useful to deal with situations where one wants to assess the impact of softer attributes. Our approach can also be used to deal with situations where one has a

large number of attributes that can't all be included in the tradeoff exercise. The approach summarized in this paper can be extended in several useful ways. First, other variables than brand could be used to make the integration between conjoint and non-conjoint variables. In our illustration we used brand as the variable that connects the two stages, but we also used channel (retail versus web), technology type (CD versus DVD versus tape), and other attributes as a link between conjoint and non-conjoint attributes. Second, we only have one non-conjoint level in our illustration (using simple OLS). This model simplicity does not have to be the case. The second stage can be a set of hierarchical regressions in which brand attributes are regressed on attribute subcomponents (this is actually the situation shown in exhibit 1). For example, brand equity may be a function of *service* and *image*, while *service* is modeled as a function of *web tech support* and *phone tech support*. By creating this hierarchical model, the results of the second stage move towards being more actionable. The second stage could also employ other more sophisticated designs. The second stage might use factor analysis or structural equations to model brand equity. They could use latent class regression or HB regression techniques. With any of these designs, the basic framework remains the same – utilities derived from a conjoint are used as dependent variables in a second stage regression-based analysis. However, by applying latent-class or Hierarchical Bayes techniques we could use our approach for segmentation purposes. It is very likely that different groups of consumers are looking for different things, not only at the level of concrete attributes but also at the level of brand positioning attributes.

Finally, we could apply our approach to study consideration set issues. In complex markets consumers often screen-out alternatives they do not wish to evaluate in detail. We believe that for consumers the most efficient way of screening-out alternatives is using perceptions they already have in their mind, instead of looking at concrete attributes, since it requires no mental searching costs at all.

Our method is not new, and several commercial market research firms probably apply our method in one form or another. However, in a lot of branding research the focus is on 'just' measuring a brand's position on identified branding variables, such as image attributes, brand personality, and brand relationship attributes without explicit empirical link to how people make choices and trade-offs. By linking brand perceptions to brand choices a researcher is able to develop a framework that enables a Return on Investment analysis. Hence, we believe that any brand approach can benefit from the basic notions outlined in this paper.

ADDITIONAL READING

Park, C. S. and V. Srinivasan (1994), "A Survey-Based Method for Measuring and Understanding Brand Equity and its Extendibility", *Journal of Marketing Research*, 31, 271-28.

Swait, J. T. Erdem, J. Louviere, and C. Dubelaar (1993), "The Equalization Price: A Measure of Consumer-Perceived Brand Equity", *International Journal of Research in Marketing*, 10, 23-45.

Vriens, M. and F. ter Hofstede (2000), "Linking Attributes, Benefits and Values: A Powerful Approach to Market Segmentation, Brand Positioning and Advertising Strategy", *Marketing Research*, 3-8.

Wedel, M., M. Vriens, T. Bijmolt, W. Krijnen and P. S. H. Leeflang (1998), Assessing the Effects of Abstract Attributes and Brand Familiarity in Conjoint Choice Experiments, *International Journal of Research in Marketing*, 15, 71-78.

COMMENT ON VRIENS AND FRAZIER

DAVID BAKKEN
HARRIS INTERACTIVE

This paper represents an extension to methods for understanding the brand part-worths yielded by conjoint methods. Typically, brand perceptions obtained outside the conjoint task are entered as predictors in a regression analysis, with the brand part worth as the dependent variable. Vriens and Frazier describe a method for incorporating the regression coefficients directly into a conjoint simulator, so that the impact of changes in brand perceptions can be estimated.

In reviewing this paper, three questions come to mind. First, what are we modeling in the brand term? Vriens and Frazier base their method on the assumption that the brand part-worth reflects brand "image," a vector of perceptions on image-related attributes. The object of the regression analysis is to determine the relative weight of each relevant attribute. This view is subject to all of the usual considerations with respect to model specification and identification. For example we must assume that statements used to measure the perceptions encompass the entire domain of brand positions for the category.

However, there are at least two other interpretations of the brand part-worth. First, the brand part-worth might simply reflect the expected probability that an alternative will deliver on its promise. In other words, "brand" is simply an indicator of the likelihood that the promised benefits will be delivered. Second, the brand part-worth might be an indivisible component of utility reflecting individual history with the brand.

The second question that came to mind is "What came first, the image or the brand?" If the "image" comes first—that is, marketing activities create a unique set of perceptions for the brand that drives preference, then it may be reasonable to assume that changes in perceptions will lead to changes in brand utility. However, even if perceptions are the *initial* driver of brand utility, it may be difficult to change utility once it has been established. On the other hand, it is possible that the perceptions are a consequence of experience with the brand and, rather than causing brand utility, simply covary with it. In that case, changes in brand perceptions — as measured by attribute ratings — may have less impact on actual behavior than might be expected from the regression model used by Vriens and Frazer.

The final question concerns the appropriate representation of the image component. Vriens and Frazier estimate an aggregate model for the brand image attributes. Some method that accounts for heterogeneity across respondents would be preferable. One possibility is the incorporation of the brand perceptions into the estimation of the conjoint model. This might be accomplished by creating "indicator" variables to reflect the brand perceptions. The brand part-worth would be replaced by part-worths associated with each of the perceptions, plus an intercept or residual term for the brand.

DATA FUSION WITH CONJOINT ANALYSIS

COMBINING SELF-EXPLICATED AND EXPERIMENTAL CHOICE DATA

AMANDA KRAUS
CENTER FOR NAVAL ANALYSES
DIANA LIEN
CENTER FOR NAVAL ANALYSES
BRYAN ORME
SAWTOOTH SOFTWARE

INTRODUCTION

This paper explores the potential for using hybrid survey designs with multiple preference elicitation methods to increase the overall information gained from a single questionnaire. Specifically, we use data generated from a survey instrument that included self-explicated, as well as partial- and full-profile CBC questions to address the following issues:

- Can self-explicated data be used with standard CBC data to improve utility estimates?
- Can the combined strengths of partial and full-profile designs be leveraged to improve the predictive power of the model?

PROJECT BACKGROUND

Study goals

The hybrid survey was designed for the development of a choice-based conjoint (CBC) model of Sailors' preferences for various reenlistment incentives and other aspects of Naval service. The study sponsor was the US Navy, and the main goal of the study was to quantify the tradeoffs Sailors make among compensation-based incentives and other, non-compensation job characteristics when making their reenlistment decisions.

Analysis of behavioral data from personnel files already provides good estimates of the effect of compensation on reenlistment rates. However, much less is known about how compensation-based reenlistment incentives compare with other, non-compensation factors that can be used to influence a Sailor's reenlistment decision. In particular, behavioral data cannot shed much light on the retention effects of most non-pay factors because we typically cannot observe which factors were considered in an individual's decision. However, CBC survey data overcome this drawback by effectively setting up controlled experiments in which Sailors make decisions about specific non-compensation aspects of Navy life.

Traditional reenlistment models

To manage the All-Volunteer Force, it was considered necessary to develop an understanding of the relationship between reenlistment behavior and military pay. Thus, there have been numerous efforts to quantify this relationship in terms of the pay elasticity of reenlistment, which measures the percentage change in the reenlistment rate due to a one percent change in military pay. Goldberg (2001) summarizes the results of 13 such studies conducted during the 1980s and 1990s. The studies all indicate that reenlistment behavior is responsive to changes in pay, but the estimates of the degree of responsiveness vary substantially. Specifically, the range of elasticity estimates is from as low as 0.4 to as high as 3.0, depending on the model used and the definition of pay.

In these studies, reenlistment is traditionally modeled as a discrete choice — to reenlist or not — that is a function of some measure of Navy compensation, [1] the individual characteristics of Sailors, and other variables that control for a Sailor's likely civilian opportunities. There have also been several studies that have included explanatory variables that capture the effects of working and living conditions on reenlistment rates. Examples in the former category include duration and frequency of time at sea, promotion rates, and measures of overall job satisfaction; examples in the latter category include type of housing and the availability of child care and recreational facilities. Finally, these models are typically estimated using logit or probit.

Why use CBC?

The Navy is interested in exploring the application of CBC surveys and models to personnel planning because survey data are frequently needed to fill gaps in personnel and other administrative data. Generally, these gaps arise for one of the following three reasons. First, is the typical "new products" case in which planners are considering implementing new programs or policies for which historical behavioral data simply don't exist. Second, the Navy does not collect administrative data to track the use of all its programs. Therefore, in some cases, data don't exist for programs that have been in place for substantial periods of time. Finally, even when data on the policies of interest do exist, they are often inappropriate for use in statistical analyses because the variables have too little variability over time or across individuals, are highly collinear with other variables in the model, or their levels are determined endogenously with reenlistment rates. For example, basic pay in the military is fully determined by specific characteristics such as rank and years of service. This means that there is very little variation across individuals.

Attention was focused on the CBC approach because, compared with other survey methods, models based on CBC data are more consistent, both behaviorally and statistically, with the reenlistment models that are currently in use. As noted above, traditional reenlistment models are discrete choice models, and underlying the statistical models are behavioral models of labor supply that are based on random utility theory

[1] Although people are responsive to changes in compensation, across-the-board increases in basic pay are considered an expensive way to increase reenlistment. Thus, the Navy created the Selective Reenlistment Bonus (SRB), which is a more targeted pay-based reenlistment incentive and which has become the Service's primary tool for managing reenlistment and retention. Because of its importance as a force management tool, many of the studies discussed above estimate the impact of changes in the SRB along with the impacts of pay.

(RUT). At a more intuitive level, CBC questions were also appealing because they better mimic real choice processes than do other types of questions.

THE HYBRID SURVEY DESIGN

13 Attributes

The attributes in the survey were chosen with several considerations in mind. First, to answer the study question, the survey had to include both measures of pay and non-pay job characteristics. And, within the non-pay category of characteristics, it was important to include attributes that captured both career and quality-of-life aspects of Naval service. In addition, attributes were chosen to reflect the concerns of Sailors on one hand, and policy makers on the other.

With so many criteria to fulfill, the final attribute list included 13 job characteristics: five compensation-based attributes related to basic pay, extra pay for special duties, reenlistment bonuses, and retirement benefits; second-term obligation length; two attributes related to the assignment process; changes in promotion schedules; time spent on work related to Navy training; time for voluntary education; and two attributes related to on- and off-ship housing.

The hybrid design mitigates problems associated with many attributes

Currently, there is not complete agreement among researchers regarding the maximum number of attributes that can be included in a CBC survey. According to Sawtooth Software's CBC documentation (Sawtooth Software, 1999), the number of attributes is limited by the human ability to process information. Specifically, Sawtooth Software suggests that options with more than six attributes are likely to confuse respondents. More generally, Louviere (Louviere, et al, 2000) indicates that the survey results may be less reliable statistically if the survey becomes too complex. However, Louviere also points out that some very complicated survey designs have been quite successful in practice.

The relationship between the quality of data collected with a CBC survey and the complexity of the tasks within it makes it necessary to make trade-offs between accommodating respondents' cognitive abilities to complete the tasks versus creating accurate representations of reality and collecting enough information to generate statistically meaningful results. In particular, one of the main problems associated with including a large number of attributes is that it may become necessary for respondents to adopt simplification heuristics to complete the choice tasks, which may lead to noisier data. Thus, in the reenlistment application, the primary issue was including enough attributes to fully capture the important determinants of quality of service in the Navy, without overwhelming respondents with too many job factors.

We used two strategies to address this potential problem and to minimize its effects. First, we chose as our target respondent population Sailors who were nearing their first actual reenlistment decisions, and were thus likely to have fairly well developed preferences regarding different aspects of Navy life. Second, we developed the three-part hybrid survey design with one section in which respondents were asked to provide explicit preference ratings for the survey attributes and two sections in which they were

asked to make discrete choices among options with different combinations of survey attributes and attribute levels. Each section and its purpose is described below and each description includes a sample survey task.

Survey Section 1 – Self-explicated questions

In the first section of the survey, respondents were instructed to rate each job characteristic, and then indicate how important getting their most preferred levels would be in making their reenlistment decisions. A sample task from Section 1 is shown in Figure 1.

Figure 1.
Sample task – self-explicated question
Change in Expected Promotion Date After Reenlistment

How much do you like or dislike each of the following promotion schedules? (Check 1 box for each item)	☹ - **1**	- **2**	- **3**	- **4**	😐 **5**	- **6**	- **7**	- **8**	☺ **9**
Get promoted 6 months **later** than expected	☐	☐	☐	☐	☐	☐	☐	☐	☐
Get promoted on expected date	☐	☐	☐	☐	☐	☐	☐	☐	☐
Get promoted 6 months **sooner** than expected	☐	☐	☐	☐	☐	☐	☐	☐	☐
Get promoted 12 months **sooner** than expected	☐	☐	☐	☐	☐	☐	☐	☐	☐
Considering the promotion schedules you just rated, how important is it to get the best one instead of the worst one?	Not Very Important								Extremely Important
	☐	☐	☐	☐	☐	☐	☐	☐	☐

One of the overriding objectives of the study was to obtain relatively stable individual-level estimates for all 13 attributes and all 52 attribute levels. As a safety net to increase the likelihood of this, we included self-explicated questions. We tested two ways to combine the self-explicated data with the choice data. The specific tests and their results are described later in the paper. One side benefit of this section was that it can ease respondents into the more complex choice tasks in sections 2 and 3 by introducing them to all 13 attributes, and can help them begin to frame reliable trade-off strategies.

Survey Section 2 – Partial-profile questions with no "none" option

Section 2 of the survey is the first of the two choice sections. Each of the 15 tasks in this section included four concepts, and each concept was defined by a different combination of only four of the 13 attributes. These partial-profile tasks did not include a "none" option. A sample task is shown in Figure 2.

Figure 2.
Sample task – partial-profile question, without none

Which of the following pay, work, and benefits packages is best for you?
Assume the packages are identical in all ways not shown. (Check only one box.)

Package 1 ☐	Package 2 ☐	Package 3 ☐	Package 4 ☐
Spend 95% of your time using skills and training	Spend 30% of your time using skills and training	Spend 50% of your time using skills and training	Spend 75% of your time using skills and training
Get promoted 12 months **sooner** than expected	Get promoted 6 months **sooner** than expected	Get promoted 6 months **later** than expected	Get promoted on expected promotion date
No change in shipboard living space	Increased shipboard recreational (study, fitness) space	Increased shipboard storage and locker space	Increased shipboard berthing space
Live in 3- to 4-person barracks	Live in 1- to 2-person barracks	Live on ship while in port	Get BAH and live in civilian housing

The partial-profile choice tasks were included in the survey to address the possibility that responses to full-profile tasks might not yield stable utility estimates for all 52 attribute levels. Specifically, partial-profile tasks impose a lighter information-processing burden on respondents. This increases the likelihood that respondent choices will systematically be due to the net differences among concepts, and reduces the likelihood that simplification heuristics will be adopted. The partial-profile approach to estimating utility values was tested by Chrzan and Elrod (Chrzan and Elrod, 1995) who found that responses to choice tasks were more consistent and utility estimates more stable using partial-profile questions rather than full-profile questions. Further research was presented in this 2003 Sawtooth Software Conference supporting those conclusions (Patterson and Chrzan, 2003). An additional benefit is that including only a few attributes allows the concepts to be more clearly displayed on the computer screen.

Survey Section 3 - Nearly full-profile questions with "none" option

The third section of the survey is the second of the two choice sections. The tasks in this section are nearly full-profile: each concept in each question included various levels for the same set of 11 of the 13 attributes. Thus, each concept represented a specific hypothetical reenlistment package. Each question also included a "none" or "would not reenlist" option, but the questions varied in terms of the number of reenlistment packages from which respondents were asked to choose. Specifically, there were nine total questions in the section: three of them had one concept plus a none option, three had two concepts plus none, and three had three concepts plus none. A sample task with two reenlistment packages is shown in Figure 3.

These nearly full-profile tasks were used principally to estimate the "None" threshold parameter. Ideally, the reenlistment packages in these tasks would have included all 13

attributes, in which case, they would have been truly full-profile tasks. However, when we considered that people might have 640x480 resolution computer monitors, we decided that concepts with all 13 attributes just wouldn't be readable. We carefully deliberated which two attributes to leave out by considering which attributes might be less important and which, when left out, could most naturally be assumed to be held at an average level of desirability relative to the levels studied (if respondents consistently viewed these omitted attributes as “average” levels, then there would be no systematic effect on the none parameter).

Finally, in addition to serving the different roles described above, including two types of choice questions is also expected to increase the level of interest in what potentially could be a tedious questionnaire.

THE DATA

The survey was fielded via disk-by-mail, and was sent to approximately 9,000 Sailors who were within one year of a first reenlistment decision. Although the current trend is toward surveying on the internet or by e-mail, we were advised against internet-based delivery mechanisms because of access issues, especially for junior Sailors and Sailors on ships. In addition to the survey disks and the return mailer for the disks, the survey packets also included the following written documentation: a cover letter explaining the purpose of the survey; instructions for starting and completing the survey; and a list of all 13 job characteristics and their definitions.

Figure 3.
Sample task – Nearly full-profile question, with none

If you were facing your next reenlistment decision and these were the only two options available to you, which would you choose, or would you not reenlist? Please check only one box.

Reenlist, Package 1 ☐	**Reenlist, Package 2** ☐	**Don't Reenlist** ☐
PAY, BENEFITS, INCENTIVES, AND TERMS OF REENLISTMENT		Neither of these packages appeals to me; I would rather not reenlist for a second obligation.
3% basic pay increase	6% basic pay increase	
1-point increase in SRB multiplier	½-point increase in SRB multiplier	
50% of SRB paid up front, remainder in annual installments	75% of SRB paid up front, remainder in annual installments	
$50-per-month increase in sea pay	No increase in sea pay	
Match TSP up to 5% of basic pay	Match TSP up to 7% of basic pay	
3-year reenlistment obligation	5-year reenlistment obligation	
CAREER AND ASSIGNMENT PROCESS		
Location guarantee for next assignment	No location or duty guarantee for next assignment	
Spend 75% of your time using skills and training	Spend 30% of your time using skills and training	
Get promoted 6 months **sooner** than expected	Get promoted 6 months **later** than expected	
QUALITY OF LIFE		
10 hours per workweek for voluntary classes and study	3 hours per workweek for voluntary classes and study	
Live in 1- to 2-person barracks	Get BAH and live in civilian housing	

Following standard survey practice, we also sent notification letters approximately two weeks before and reminder letters approximately three weeks after the survey packets were mailed. Advance notice of the survey was also given through the media. Specifically, articles announcing that the survey would soon be mailed were published in The Navy Times and the European and Pacific Stars and Stripes, as well as on the Navy's own news website, www.news.navy.mil.

Finally, the survey was in the field for approximately 14 weeks, and the response rate was about 18 percent — just a few percentage points higher than the expected rate of 15

percent. After data cleaning,[2] the final sample size on which analysis was based was 1,519 respondents.

Approaches to Utility Estimation with Data from the Hybrid Survey

Modeling goals

The main goal of the research was to create an accurate choice simulator for studying how various changes in the work environment and offerings might affect reenlistment rates. The Navy had commissioned similar research in the past where the supplier used a discrete choice methodology and aggregate utilities (MNL). While the Navy had been generally pleased with the results, a major drawback of the aggregate simulator was the lack of estimated standard errors for the shares of preference, and thus confidence intervals around the predictions.

The aim of the research this time was to model many attributes (13) using a choice-based approach, and also to report standard errors for shares. We decided, therefore, to fit individual-level models, which when used in choice simulators yield useful estimates of standard errors.

As benchmarks of success, we felt the simulator should:

- Produce accurate individual-level models of preference,
- Demonstrate good accuracy overall in predicting aggregate shares of choice among different reenlistment offerings.

Three approaches for utility estimation

Recall that there were three main sections in our hybrid conjoint survey: 1) Self-explicated, 2) Partial-profile choice tasks (15 tasks), 3) Near-full profile choice tasks (9). The partial profile choice tasks did not feature a "Would not reenlist" (same role as "None") option. However, the "Would not reenlist" option was available in the final nine near-full profile choice tasks. Given these sources of information, we could take a number of paths to develop the part worths for the final choice simulator. The final simulator must reflect part worths for the 13 attributes x 4 levels each (52 total part worths) plus an appropriate "Would not reenlist" parameter estimate. Even though the near-full profile tasks featured 11 of the 13 attributes (because of screen real estate constraints), for ease of description we'll refer to them as full-profile.

We investigated three main avenues that have proven useful in previous research.

1. **ACA-Like "Optimal Weighting" Approach:** The self-explicated section was the same as employed in ACA software, and yielded a rough set of part worths. Using HB, we could also estimate a set of part worths from the partial profile choice tasks. Similar to the approach from a previous version of ACA, we could find "optimal" weights for the self-explicated and choice-based part worths to

[2] For data cleaning, we looked at time taken to complete the partial profile questions (i.e., section 2), and the degree to which there were patterned responses to these questions. After reviewing the data, we chose to eliminate any respondent who took fewer than two minutes to complete section 2 and any respondent who chose the same response or had the same pattern of responses on all 15 questions in the section. Based on these criteria, 34 respondents were dropped from the sample.

best fit the choices in the final full-profile choice section. Rather than use OLS (as does ACA), we could use HB to estimate optimal weights (constrained to be positive).

This approach involves a two-step HB estimation procedure. First, we used HB to estimate part worths using the partial-profile choice tasks. Then, given those part worths and those from the self-explicated section, we ran a second HB model using the choice tasks from the full-profile choice-based section. We estimated three parameters: 1) weight for self-explicated part worths, 2) weight for partial-profile choice part worths, 3) "would not reenlist" threshold. (Details regarding this are described in the appendix.)

As a point of comparison, we also fit a "self-explicated only" model, which used the model specification above, but estimated only a weight for the self-explicated part worths and the "would not reenlist" threshold.

2. **Choice Questions Only:** Using HB estimation, researchers have often been able to develop sound individual-level models using only a limited number of choice questions. These models often feature good individual-level predictions as well as exceptional aggregate share accuracy. We decided to investigate a model that ignored the self-explicated information altogether, and used only the choice-based questions. For each respondent, we combined both the partial-profile choice tasks and the full-profile choice tasks[3]. We used logit, latent class and HB estimation with this particular model specification.

   The interesting thing to note about this approach is that the full-profile questions are used not only to calibrate the "would not reenlist" threshold, but are also used to further refine the part worth estimates. In the "optimal weighting" model described above, the full-profile choice tasks are only used to find optimal weights for the previously determined part worths and to calibrate the "would not reenlist" threshold. (More details are provided in the appendix.)

3. **Constrained HB Estimation:** We also tried a type of HB estimation that constrains part worth estimates according to the ordinal relationships given in the self-explicated section of the interview. For example, if a respondent rated a particular level higher than another in the self-explicated section, we could constrain the final part worths to reflect this relationship. The approach we used for this HB estimation is called "Simultaneous Tying" (Johnson 2000). During estimation, two sets of individual-level estimates are maintained: an unconstrained and a constrained set. The estimates of the population means and covariances are based on the unconstrained part worths, but any out-of-order levels are tied at the individual-level prior to evaluating likelihoods. The model

[3] Previous researchers have pointed to potential problems when combining data from different formats of preference/conjoint questions, as the variance of the parameters may not be commensurate across multiple parts of a hybrid design (Green et al. 1991). Moreover, given other research suggesting the scale factor for partial- and full-profile CBC differs (Patterson and Chrzan, 2003), this seemed a potential concern. We resolved to test this concern empirically, examining the fit to holdouts to understand whether unacceptable levels of error were introduced.

specification described above (Choice Questions Only) was also used here, but with individualized, self-explicated constraints in force.

Comparing Modeling Approaches Based on Internal Validity Measures

Holdouts for internal validity[4]

To check the internal validity of our model, we held out both choice tasks *and* respondents.

Holdout Respondents: We randomly divided the sample into two halves, and used the model from one half to predict the holdout choices of the other half, and vice-versa.

Holdout Tasks: We held out three of the full-profile choice tasks (which included the "would not reenlist" alternative). The model developed using all the other preference information was used to predict the choices for these held out tasks.

Measures of internal validity

We gauged the models with two criteria of success: aggregate predictive accuracy (Mean Absolute Error, or MAE), and individual hit rates. An example of MAE is shown below.

Concept	Actual choices	Predicted choices	Absolute difference
Package 1	30%	24%	6
Package 2	50%	53%	3
Would not reenlist	20%	23%	3
		Total error:	12
		MAE:	12/3 = 4

If the actual choice frequencies for the three alternatives in a particular choice scenario and the predicted choice (using the model) were as given above, the MAE is 4. In other words, on average our predictions are within 4 absolute percentage points of the actual choices.

To estimate hit rates, we use the individual-level parameters to predict which alternative each respondent would be expected to choose. We compare predicted with actual choice. If prediction matches actual choice, we score a "hit." We summarize the percent of hits for the sample.

Comparative internal validity

We used Randomized First Choice (RFC) simulations to estimate probabilities of choices (shares) for the alternatives, tuning the exponent for best fit. A summary of the

[4] For a discussion of using holdout choice tasks and holdout respondents, see Elrod 2001.

performance for the different approaches is given in the table below, sorted by share prediction accuracy (MAE).

Model	Estimation Method	MAE	Hit Rate
Choice Questions Only	HB	1.7	65%
	Latent Class	2.3	58%
	Logit	2.5	51%
Optimal Weighting	2-stage HB	2.7	63%
Self-Explicated	N/A	3.3	60%
Constrained	HB	5.8	59%

The model with the best aggregate predictive accuracy (Choice Questions Only, HB) also had the highest hit rate. This is not always the case in practice, and if it occurs it is a very satisfying outcome. Our goals for achieving high group-level and individual-level predictive validity could be met using this model.

With the Optimal Weighting approach, we found that slightly less than 10 percent of the weight on average was given to the self-explicated part worths and 90 percent or more given to the choice-based part worths. For those readers with experience with Sawtooth Software's ACA, this is not usually the case. Our experience is that ACA often gives about 50 percent of the weight or slightly more to the self-explicated portion of the hybrid conjoint survey.

Even though the self-explicated information alone provided relatively accurate estimates of both shares and hit rates (though not as good as the best model), we were unable to find a way to use the self-explicated information to improve our overall fit to holdout choices. We were particularly surprised that the constrained estimation didn't turn out better. This method worked well in a study reported by Johnson *et al*. in this same volume.

After determining which model performed best, we added the three holdout tasks back into the data set for estimation. We re-estimated the model using all tasks (15 partial profile, plus 9 full-profile) and all respondents combined. This final model was delivered for modeling sailors' preferences.

CONCLUSIONS REGARDING MODELING

Summary of results

For this data set, 15 partial-profile choice tasks plus the 6 full-profile tasks provided enough information for relatively accurate modeling at both the individual and aggregate levels. Using the self-explicated information only degraded performance, where performance was defined in terms of predicting holdout choice sets. Whether the added self-explicated information would improve the model in terms of fitting actual sailors' choices is an important topic that is beyond the scope of this paper. As a final note, these findings may not generalize to other data sets.

Was the self-explicated section a waste?

These conclusions raise the obvious question as to whether including the self-explicated section was wasted effort. Of course, we didn't know prior to collecting the data whether the partial-profile choice tasks alone could provide enough information to stabilize individual-level estimates for the 52 part worth parameters — and stable individual estimates was an established goal at the onset. Jon Pinnell had raised questions regarding the stability of individual-level estimates from partial-profile choice in the 2000 and 2001 Sawtooth Software Conference Proceedings (Pinnell 2000, Pinnell 2001).

In hindsight, one might suggest that we would have been better off asking respondents additional choice tasks rather than spending time with the self-explicated exercise. However, self-explicated exercises can provide a systematic introduction to the various attributes and levels that may help respondents establish a more complete frame of reference prior to answering choice questions. After a self-explicated exercise, respondents are probably better able to quickly adopt reliable heuristic strategies for answering complex choice questions. It is hard to quantify the overall value of these self-explicated questions, given that we didn't include a group of respondents that didn't see the self-explicated task.

APPENDIX

TWO-STAGE HB ESTIMATION USING "OPTIMAL WEIGHTING"

1. Using the 15 partial-profile choice tasks (4 concepts per task, no "None") and CBC/HB software, estimate individual-level partial profile utilities for 13 attributes x 4 levels = 52 total levels. These are zero-centered utilities. Normalize these so that the average of the differences between best and worst levels across all attributes for each individual is equal to one.
2. Using the same technique as described in ACA software documentation, use the ratings for levels and importance ratings from the Priors section to develop self-explicated utilities. Normalize these also as described above.
3. Use six of the nine full-profile choice questions (which included a "None" concept) to find optimal weights for partial-profile choice utilities and self-explicated utilities to best predict choices, and additionally fit the None parameter. The X matrix consists of three columns: a) utility of that concept as predicted by partial profile utilities, b) utility of that concept as predicted by self-explicated utilities, c) dummy-code representing the "None" alternative (1 if present, 0 if absent). Use CBC/HB to estimate these effects. Constrain the first two coefficients to have positive sign.

CONCATENATED CHOICE TASKS MODEL

1. Recall that we have information from two separate choice sections: a) 15 partial profile choice tasks, 4 concepts each, with no "None", b) 6 full-profile choice tasks, with a "None."
2. We can code all the information as a single X matrix with associated Y variable (choices). The matrix has 40 total columns in the X matrix. The first 39 columns are all effects-coded parameters representing the 13 attributes (each with three effects-coded columns representing the four levels of each attribute). The final parameter in the X matrix is the dummy-coded "None" parameter (1 if a "None" alternative, 0 if not). For the first 60 rows of the design matrix (the partial profile choice tasks), the "None" is not available.
3. The full-profile choice tasks contribute the only information regarding the scaling of the None parameter relative to the other parameters in the model. They also contribute some information for estimating the other attribute levels.

REFERENCES

Elrod, Terry (2001), "Recommendations for Validation of Choice Models," Sawtooth Software Conference Proceedings, 225-43.

Green, P. E., Krieger, A. M., and Agarwal, M. K. (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," Journal of Marketing Research, 28 (May), 215-22.

Johnson, Richard M. (2000), "Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes," Sawtooth Software Technical Paper available at www.sawtoothsoftware.com.

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

Keith Chrzan and Terry Elrod,. "Choice-based Approach for Large Numbers of Attributes," Marketing News, vol. 29, no. 1, p. 20, January 2, 1995.

Matthew S. Goldberg (2001), *A Survey of Enlisted Retention: Models and Findings*, CNA Research Memorandum D0004085.A2.

Patterson, Michael and Keith Chrzan (2003), "Partial Profile Discrete Choice: What's the Optimal Number of Attributes?" Sawtooth Software Conference Proceedings.

Pinnell, Jon (2000), "Customized Choice Designs: Incorporating Prior Knowledge and Utility Balance in Choice Experiments," Sawtooth Software Conference Proceedings, 179-93.

Pinnell, Jon (2001), "The Effects of Disaggregation with Partial Profile Choice Experiments," Sawtooth Software Conference Proceedings, 151-65.

Sawtooth Software, Inc. The CBC System for Choice-Based Conjoint Analysis, January 1999.

CREATING A DYNAMIC MARKET SIMULATOR: BRIDGING CONJOINT ANALYSIS ACROSS RESPONDENTS

JON PINNELL AND LISA FRIDLEY
MARKETVISION RESEARCH

After completing a project, researchers often wish they had known something when designing the research that they learned from the research. While a combination of good planning, experience, industry expertise and pre-testing can eliminate many of these instances, some are inevitable. Researchers conducting a study with conjoint or discrete choice analysis are not immune to this predicament. It isn't unheard of that after reporting the results of a study (maybe a week later, maybe 12 months later), a technology that wasn't feasible at the time of study design becomes feasible. The newfound attribute, however, is not in the study. Other less fortunate scenarios also exist that can result in an attribute not being included in a specific study.

The researcher facing this scenario could react in a number of ways, along a continuum. This continuum, which we have named the denial continuum, is bounded on one side by denying the missing attribute exists and relying solely on the previous study; and is bounded on the other side by denying the previous study exists and conducting an entirely new study. We anticipate problems with either extreme.

We were interested if an approach could be developed that would allow an efficient methodology to update an existing conjoint/discrete choice study with incremental or additional information. Specifically, we were interested if a previously conducted study could be updated with information from a second study in an acceptable fashion.

Initially, we evaluated two competing approaches. The first, bridging, might be more common in the conjoint literature. The second idea was data fusion. Each is discussed in turn:

CONJOINT BRIDGING

The issue presented here has some similarity to bridging conjoint studies. Though discussed less frequently today, bridging was a mechanism used to deal with studies that included a large number of attributes. In this application, a single respondent would complete several sets of partial profile conjoint tasks. Each set would include unique attributes, save at least one that was required to be common across multiple sets. For example, the task in total might include 12 attributes, the first set of tasks included attributes 1-5, the second set included attribute 1 and 6-9, and the third set included attributes 1 and 10-12. Utilities would be estimated for each of the three sets separately and then combined back together. The common attribute allowed the studies to be bridged together. Subsequent to this original use, bridging designs also became used in applications dealing with pricing research. In this application, two conjoint studies were completed with a single respondent. The first study dealing with product features and the second dealing more specifically with price and maybe brand. Again, at least one

attribute was required to be common between the two studies. This approach was also known as dual conjoint or multistage conjoint (see Pinnell, 1994).

At some level, the situation at hand is like a multistage conjoint. However, multistage conjoint studies were typically conducted within subject. Alternatively, the design was blocked between subjects, but utilities were developed only at the aggregate (or maybe subgroup) level. In the scenario presented here, we couldn't be assured that we would be able to conduct the second study among the same respondents, and after stressing the importance of considering individual differences felt uncomfortable with an approach that did not produce individual level utilities. We therefore rejected bridging as an approach to solving the problem. As an alternative, we explored approaches more related to data fusion. We approach the topic initially as a data imputation problem.

DATA FUSION/IMPUTATION

A vast literature has developed over the past 40 years to deal with the issue of missing data (see Rubin or Little for extensive discussions). Data can be missing for several reasons — by design or not. When the data are missing by design — the more simple case — it is because the researcher chose not to collect the data. When the data are not missing by design it is typically a result of either item non-response or unit non-response. Unit non-response refers to a researcher's request for a potential respondent to complete a survey and the respondent (unit) not complying. Item non-response, on the other hand, refers to a respondent who participated in the survey, but did not provide answers to every question. In either case, the researcher has a series of choices to make. Several approaches to deal with various non-response problems have been developed. They include:

- Ignore missing data,
- Accept only complete records,
- Weight the data,
- Impute missing data.

We focus on imputation. In practice, imputation is known by many names, though the most common are imputation and ascription. The goal of imputation or ascription is to replace missing values (either missing items or missing units) with reasonable values. The standard of reasonableness is held to different interpretations based on the researchers and the application. In some cases, reasonable just means that the filled in values are within the range allowed. Other cases, however, require that the data be reasonable for the specific case being remedied. That is, maintaining internal consistency for the record.

To illustrate, we will present imputation approaches as a remedy for item non-response. Three specific methods are commonly used:

- Mean substitution,
- Hot deck,

- Model based.

Each is discussed in turn.

Mean Substitution

One common approach to determining a reasonable value to impute is to impute a mean. In the case of item non-response within a ratings battery, it is common to impute a row-wise mean. In other instances, a column-wise mean is imputed. In imputation applications, imputing a column-wise mean is more common than imputing a row-wise approach.

While mean substitution is commonly used, it will not maintain the marginal distribution of the variable, nor will it maintain the relationship between the imputed variable and other variables.

The mean substitution procedure can be improved by proceeding it with a step in which the data records are post-stratified and the conditional mean is imputed. This has been shown to improve the final (imputed) data, but it depends entirely on the strength of the relationship between the post stratifying variables and the imputed variables. In our experiences, the variables we are seeking to impute are only mildly related to common post-stratifying variables, such as demographics.

Hot Deck

In a hot deck imputation, the reasonable values are actual observed data. That is, a record (recipient case) that is missing a value on a specific variable is filled in with data from a record that includes a value for the variable (donor case).

The donor case can be selected in a number of ways. The least restrictive selection method is to select a case at random, though limits can be placed on the number of times a case can act as a donor. This random hot deck procedure will maintain the marginal frequency distribution of the data, but will likely dampen the relationship between the variables.

To better maintain the relationships between variables constraints are imposed on the donor record. As with the mean substitution routine, the data are post-stratified and the donor record is constrained to be in the same stratum as the recipient record. This is often referred to as a sequential hot deck, and seems more sensible than the random hot deck.

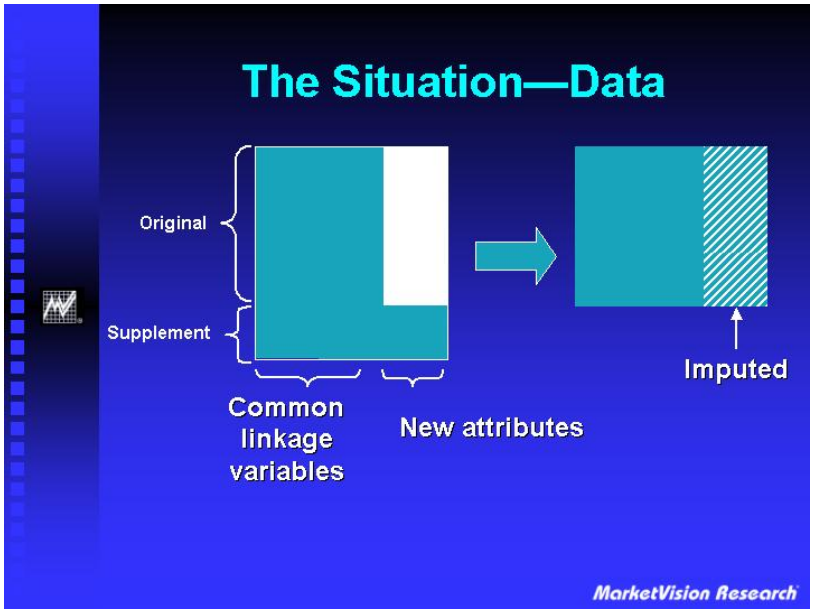
As its extreme, the post-stratification could continue until the strata are quite small. In this case, each donor is matched to only one possible recipient. This special case of hot deck imputation is referred to as nearest neighbor hot deck. Some authors distinguish nearest neighbor from hot deck procedures as the neighbors might not be exact matches but are the nearest (see Sande for further discussion).

By using either a sequential or nearest neighbor hot deck, the relationships between variables are maintained.

Model-Based

The model-based approach is a further extension of imputation where the values to be filled in are predicted based on the variables that are not missing. This approach, while theoretically very appealing, often encounters difficulty in practice as the patterns of missing data are unpredictable and make any model development difficult. Several Bayesian models have been suggested, but their practical use in commercial settings seems slow in adoption.

Each of these approaches is used for imputation. In the examples discussed above, we have used the case of item non-response to illustrate the three approaches. In the case at hand, though, the goal is not to impute a missing value as might come about from item non-response, but to impute the utility structure of attributes for respondents who didn't have those attributes in their study. The example can be graphically illustrated as follows:



In this undertaking it is probably worth stating what our goal is and isn't. Our goal is to make a valid inference about the population, including the heterogeneity in the population. Our goal is not to replace the original respondent's missing utilities. Specific requirements of the exercise include:

- Produce individual level utilities,
- Maintain results of first study,
- Secondary study must be conducted quickly and cost efficiently.

Given this, our application is much like a data fusion problem but we use techniques more commonly used for imputation.

Approach

We explore approaches that allow us to develop utility estimates for the unobserved variables. The approach must also allow us to test the reasonableness of the imputed values.

The proposed approach is to conduct a small-scale study that includes the new attributes as well as all of the original attributes and classification questions. The goal would be to complete a study approximately one-quarter of the size of the original among an independent sample of respondents. Then, from this smaller pilot study, make inferences about the utility structure among the original respondents, at the individual level.

With this as background, our topic might fit more clearly under the heading of data fusion rather than imputation. However the methods employed to fill in the unobserved data will more closely match the imputation methods discussed above.

The following methods will be used to estimate the unobserved utilities:

Linear Regression (LIN)

In linear regression, the common utilities are used to estimate the new utilities in the supplemental sample, and then the resulting parameters are used to develop estimates in the original sample.

Latent Class (LCA)

Much like the linear regression model, the latent class model predicts the missing utilities as a linear composite of the common utilities. However, heterogeneous parameters are allowed.

Nearest Neighbor (NN, 4NN)

The nearest neighbor methods involve a two-step procedure in which the donor case and recipient case that are most similar in their common utilities are identified and then the missing utilities are estimated from that nearest neighbor. This approach can either just use the donor's utilities as the estimate of the missing utility, or can rely on a regression model (as above) using only the near neighbor's utilities for model development. Nearest neighbors can be defined on the one nearest case (NN) or based on a set of near neighbors, such as a four nearest neighbors (4NN).

Bayesian Regression

Finally, a Bayesian regression model was used. The Bayesian approach should account for heterogeneity, as the LCA and NN methods do, but potentially with more stability.

To explore how well each method would work in our setting we simulated datasets matching the scenario. For each of four datasets, the utilities for three attributes were deleted for a portion of the respondents. Then each method was used to estimate the known (but deleted) utilities. The datasets were selected and ordered for presentation such that each progressive dataset provides a more rigorous test of the methods, so the

results of the methods from the fourth study should be weighed more heavily than those from the first study.

Each approach will be evaluated in the following criteria:

- Correlation with known utilities,
- Hit rate of actual choices,
- Mean absolute deviation of share results from simulation.

Of these, we are inclined to place the most weight on the error of share predictions. Several works support this supposition (see Elrod and Wittink). It is also important to keep in mind that the known utilities are measured with error and are themselves a fallible criterion.

EMPIRICAL FINDINGS

Study 1

	Linear	LCA	NN	4NN	Bayesian
Correlation	0.877	0.866	0.793	0.854	0.878
Hit Rates	0.791	0.781	0.774	0.787	0.791
MAD	2.26	2.04	1.85	1.41	2.32

Study 2

	Linear	LCA	NN	4NN	Bayesian
Correlation	0.762	0.755	0.664	0.746	0.761
Hit Rates	0.726	0.728	0.708	0.726	0.728
MAD	0.82	0.83	1.51	1.28	0.77

Study 3

	Linear	LCA	NN	4NN	Bayesian
Correlation	0.773	0.761	0.491	0.615	0.772
Hit Rates	0.803	0.797	0.740	0.776	0.804
MAD	2.08	2.12	1.11	1.85	1.97

Study 4

	Linear	LCA	NN	4NN	Bayesian
Correlation	0.235	0.234	0.490	0.622	0.754
Hit Rates	0.628	0.682	0.734	0.772	0.794
MAD	3.17	6.97	1.68	2.64	2.42

Summary of Empirical Findings

Below, we present a simple average of the five methods' results across the four studies.

	Linear	LCA	NN	4NN	Bayesian
Correlation	0.662	0.654	0.610	0.709	0.791
Hit Rates	0.737	0.747	0.739	0.765	0.779
MAD	2.08	2.99	1.54	1.79	1.87

It appears that the Bayesian approach outperforms on two of the three criteria, substantially for one and marginally for another. However, for the criterion in which we place the most credence, both the nearest neighbor and the four nearest neighbor methods outperform the Bayesian and other two approaches. This prompted us to explore this near neighbors solution further to see if either a two nearest neighbors or a three nearest neighbors outperform the one or four nearest neighbors, and both do on the key criterion, as shown in the following table:

	NN	2NN	3NN	4NN
Correlation	0.610	0.658	0.690	0.709
Hit Rates	0.739	0.755	0.770	0.765
MAD	1.54	1.38	1.51	1.79

We next explore if the two or three nearest neighbor methods could be improved with the additional step of defining neighbors using both demographics and utility structure, compared to just utility structures as done above. The addition of the demographics improved neither the two nearest nor three nearest neighbors' performance, and actually was consistently deleterious.

Finally, we explore if some combination rather than a simple average could improve on the resulting estimated utilities. Using the inverse of the squared Euclidian distances as a weight, we calculate a weighted average used to define the nearness of the neighbors. The following table shows the results of this weighting (W), which provide a further reduction in error of the key criterion.

	2NN	W2NN	3NN	W3NN
Correlation	0.658	0.661	0.690	0.689
Hit Rates	0.755	0.762	0.770	0.777
MAD	1.38	1.14	1.51	1.00

CONCLUSIONS

We set out to see if an approach could be developed that would allow an efficient methodology to update an existing conjoint/discrete choice study with incremental or additional information. We are left, happily, with the basic conclusion that the approach outlined above seems to work.

The Bayesian method outperforms on two of the three criteria, but the near neighbor methods outperform on the criterion in which we place the most weight. The two nearest or three nearest neighbor methods do consistently well, especially in the more complicated applications. The performance of the nearest neighbor methods did not improve with the inclusion of demographic data. However, the performance of the nearest neighbor methods did improve by using a weighted composite rather than a simple composite.

REFERENCES

Elrod, Terry (2001), Recommendations for Validation of Choice Models, in *Sawtooth Software Conference Proceedings*, Victoria, BC: Sawtooth Software, Inc., September, 225-243.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley.

Pinnell, Jon (1994), Multistage Conjoint Methods to Measure Price Sensitivity, Proceedings of AMA Advanced Research Techniques Forum; Beaver Creek, CO.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse Surveys*. John Wiley & Sons, Inc.

Sande, I. G. (1982) Imputation in surveys: coping with reality. *The American Statistician*, 36(3), 145-152.

Wittink, Dick (2000), Predictive Validation of Conjoint Analysis, in *Sawtooth Software Conference Proceedings*, Hilton Head, SC: Sawtooth Software, Inc., March, 221-237.

ADVANCED TECHNIQUES

USING GENETIC ALGORITHMS IN MARKETING RESEARCH

DAVID G. BAKKEN
HARRIS INTERACTIVE

The past decade or so has been witness to an explosion of new analytic techniques for identifying meaningful patterns in data gathered from and about customers. Neural networks, classification and regression trees (CART), mixture models for segmentation, and hierarchical Bayesian estimation of discrete choice models have led to significant advances in our ability to understand and predict customer behavior. Many of the new techniques have origins outside of market research, in areas including artificial intelligence, social sciences, applied statistics, and econometrics. More recently, a technique with roots in artificial intelligence, social science, and biology offers market researchers a new tool for gleaning insights from customer-based information.

Genetic algorithms (GAs) were invented by John Holland in the 1960's as a way to use computers to study the phenomenon of biological adaptation. Holland was also interested in the possibility of importing this form of adaptation into computer systems. Like neural networks, GAs have their origins in biology. Neural networks are based on a theory of how the brain works, and genetic algorithms are based on the theory of evolution by natural selection. Holland (1975) presents the genetic algorithm as a representation of the molecular process governing biological evolution. Since their inception, GAs have found application in many areas beyond the study of biological adaptation. These include optimization problems in operations research, the study of cooperation and competition (in a game-theoretic framework), and automatic programming (evolving computer programs to perform specific tasks more efficiently).

HOW GENETIC ALGORITHMS WORK

Genetic algorithms have three basic components that are analogous to elements in biological evolution. Each candidate solution to a GA problem is represented (typically) as a *chromosome* comprised of a string of ones and zeros. A single "gene" might consist of one position on this chromosome or more than one (in the case of dummy variables, where a gene might be expressed in more than two "phenotypes"). A *selection operator* determines which chromosomes "survive" from one generation to the next. Finally, *genetic operators* such as mutation and crossover introduce the variation in the chromosomes that leads to evolution of the candidate solutions.

Figures 1 and 2 illustrate the biological model for genetic algorithms. Here, each chromosome is a string of upper and lower case letters. The upper case letters represent one form or "level" of a particular gene, and the lower case letters represent a different form. In a "diploid" organism with genes present on corresponding pairs of chromosomes, the *phenotype*, or physical manifestation of the gene, is the result of the influences of both genes. Gregor Mendel discovered the way in which two genes can give rise to multiple phenotypes through the dominance of one level over another.

Figure 1.
Diploid "Chromosome"

A B C D E f g H I J K L M

a B c d E f G h I J K L M

In Figure 1, upper case letters are dominant over their lower case counterparts, so the organisms with either "A/A" or "A/a" in the first position (locus) would have the same phenotype, while an organism with the "recessive" combination, "a/a," would display a different phenotype. This simple form of inheritance explains a number of specific traits. In general, however, phenotypes result from the influence of several different genes. Even in cases where both a dominant and recessive allele are paired, the true phenotype may be a mixture of the traits determined by each allele. Sickle cell anemia is a good example of this. An individual with genotype S/S does not have sickle cell anemia (a condition in which red blood cells are misshapen, in the form of a sickle). An individual with genotype s/s possesses sickle cell anemia, a generally fatal condition. An individual with genotype S/s will have a mixture of normal and sickle-shaped red blood cells. Because sickle-shaped cells are resistant to malarial infections, this combination, which is not as generally fatal, is found in African and African-American populations.

Figure 2 illustrates two important genetic operators: crossover and mutation. In the course of cell division for sexual reproduction, portions of chromosome pairs may switch places.

Figure 2.
Crossover and Mutation

a B c d E f g H i J K L M

A B C D E f G h I J K L M

In Figure 2, a break has occurred between the 5th and 6th loci, and the separated portions of each chromosome have reattached to the opposite member of the pair. In nature, the frequency of crossover varies with the length of the chromosome and the location (longer chromosomes are more likely to show crossover, and crossover is more likely to occur near the tips than near the center of each chromosome).

Mutation occurs when the gene at a single locus spontaneously changes its value. In Figure 2, a mutation has occurred at the 9th locus.

While all sexually reproducing organisms have diploid chromosomes, most genetic algorithms employ "haploid" chromosomes, with all the genetic information for an individual carried in one string.

SPECIFYING A GENETIC ALGORITHM

While genetic algorithms vary greatly in complexity, the following simple algorithm reflects the general form of most GAs:

1. Randomly generate a set of *n* chromosomes (bit-strings) of length *l*
2. Define a *selection operator* (objective function) and determine the "fitness" of each chromosome in this population
3. Select a pair of "parent" chromosomes, with probability of selection proportional to fitness (multiple matings are allowed)
4. With probability $\boldsymbol{p}_{\text{crossover}}$, crossover the pair at a randomly chosen point
5. If no crossover occurs, form two offspring that are exact copies of each parent
6. Mutate the offspring at each locus with probability $\boldsymbol{p}_{\text{mutation}}$
7. Replace the current population with the new population
8. Calculate the fitness of each chromosome in this new population
9. Repeat steps 2 through 8

The National Public Radio program "Car Talk" recently provided a problem that can be used to illustrate the specification of a genetic algorithm. The following problem was presented in the "Puzzler" section of the program:

> Dogs cost $15 each, cats cost $1, and mice are $.25 apiece. Determine the number of dogs, cats and mice such that the total number of animals is 100, the total amount spent to purchase them is $100, and the final "solution" must include at least one individual of each species.

While this is a rather trivial problem that can be easily solved analytically, we can use the problem to illustrate the way a genetic algorithm works. First, we must encode a 100-bit chromosome such that, at any one position, we represent a dog, cat, or mouse. For example, we could create the following string:

DCMMMCMMMMMMDMMMMMMMCMCMM…..M

In this chromosome, "D" stands for dog, "C" for cat, and "M" for mouse. We could also represent these alternatives as numbers (1,2,3) or as two dummy-coded variables. The specific encoding may depend on the objective function.

Next, we must determine the objective function that will be used to evaluate the fitness of each individual chromosome. One of the constraints in our problem, that the total number of animals must equal 100, is captured in the specification of the 100-bit chromosome. Every candidate solution, by definition, will have 100 animals. The second requirement, a total cost of $100, is used for the fitness function. If we replace the letters in the chromosome above with the prices for each species (D=$15, C=$1, and M=$.25), the objective function is the sum of the string, and the goal is to find a string such that the sum equals $100.

The problem includes one additional constraint: the solution must contain at least one dog, one cat, and one mouse. While it may not be necessary to include this constraint for the GA to find the correct solution to this problem, without this constraint it is possible that some of the chromosomes in any population will be completely outside the solution space. As a result, the GA may take longer to find the solution to the problem. We can introduce this constraint by fixing the first three positions (e.g., as "DCM" or 15, 1, and 0.25) and allowing the GA to vary only the remaining 97 positions.

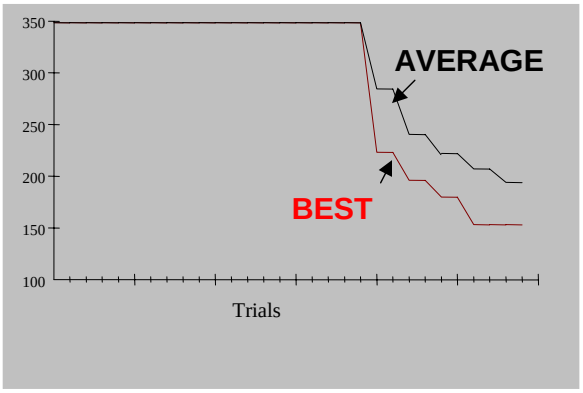
Figure 3 illustrates the results of one run of the GA for this problem, which arrived at the correct solution after 1800 generations. In the first panel (trials 0-500), the "best" solution in the initial population had a total price of about $350[1]. Because our fitness criteria is a minimizing function, the lower line represents the best solution found so far, and the upper line represents the average fitness function value for all solutions so far. The third panel (trials 1001-1800) reveal a fairly common pattern we have observed. The GA has reached a solution very close to the target, but requires many additional generations to find the final solution (3 dogs, 41 cats, 56 mice). This occurs because we have defined a continuous fitness function (i.e., the closer to $100, the better)[2]. As the members of a population approach the maximum of the fitness function, through selection, the probability that each solution will be kept in succeeding generations increases. Changes in fitness are distributed asymmetrically as the population approaches the maximum, with decreases in fitness more likely to occur at that point than increases[3]. As the solution set approaches the maximum fitness, the most fit individuals become more similar. Because crossover tends to maintain groups of genes, crossover becomes less effective as the maximum fitness of the most fit individuals increases, and more of the improvement burden falls on the mutation operator.

[1] The software used to run this GA (Evolver), does not plot improvements for the first 100 trials.

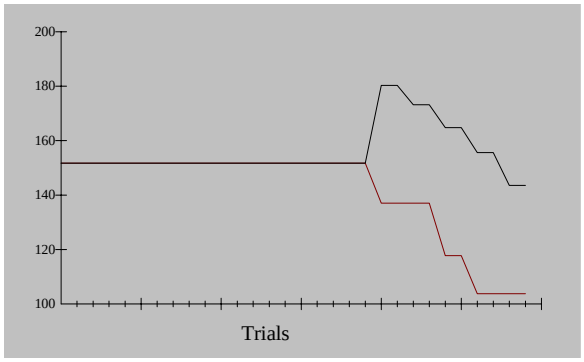
[2] If the fitness function was discrete, such that all chromosomes that did not satisfy the "sum=$100" criteria had mating probability of 0, the GA would not work.

[3] At this point, there are far more candidate solutions on the "less fit" side of the fitness distribution, so any new individual has a higher probability of being less fit than the most fit individuals in the current population.

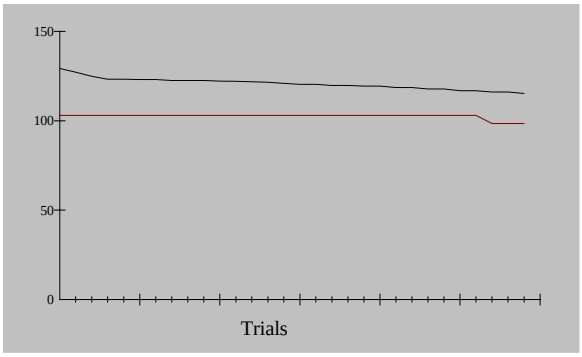
Figure 3.
Example GA Run for Dog/Cat/Mouse Problem



Trials 1-500



Trials 501-1000



Trials 1001-1800

For the GA solution depicted in Figure 3, the population size was 50. Varying population size for this problem has little impact on either the solution time or the ability to reach a solution. Solution times varied from a few seconds to several minutes, and from a few generations (11) to almost 2000 generations. In a few instances, the GA failed to arrive at the exact solution within a reasonable time or number of generations, with fitness stabilizing within $1 of the target amount. This is one of the potential drawbacks of GAs — when there is an exact solution to the problem, GAs may not converge on the exact solution.

GENETIC ALGORITHMS VERSUS OTHER SEARCH METHODS

As noted above, the dog, cat, and mouse problem is easily solved analytically. In fact, we can write an algebraic expression just for this problem that allows us to substitute values for one or more variables to find a solution. This algebraic expression represents a "strong" method for searching the solution space. Strong methods are procedures designed to work with specific problems. "Weak" methods, on the other hand, can search the solution space for a wide variety of problems. Genetic algorithms are weak methods. Weak or general methods are generally superior at solving problems where the search space is very large, where the search space is "lumpy" — with multiple peaks and valleys, or where the search space is not well understood.

Other general methods for searching a solution space include hill climbing and simulated annealing. A "steepest ascent" hill climbing algorithm could be implemented as follows:

1. Create a random candidate solution (encoded as a bit string)
2. Systematically change each bit in the string, one at a time
3. Evaluate the objective function ("fitness") for each change in the bit string
4. If any of the changes result in increased fitness, reset the bit string to the solution generating the highest fitness level and return to step two
5. If there is no fitness increase, return to step two, implementing different mutations for each bit in the string, etc.

APPLICATIONS FOR GENETIC ALGORITHMS IN MARKETING RESEARCH

Genetic algorithms ultimately may find a wide variety of uses in marketing research. Four applications are described in the following sections:

- Conjoint-based combinatorial optimization for single products
- Conjoint-based combinatorial optimization for a multi-product line
- TURF and TURF-like combinatorial optimization
- Simulation of market evolution

Other potential applications in marketing research include adaptive questionnaire design and predictive models to improve targeting.

CONJOINT-BASED COMBINATORIAL OPTIMIZATION

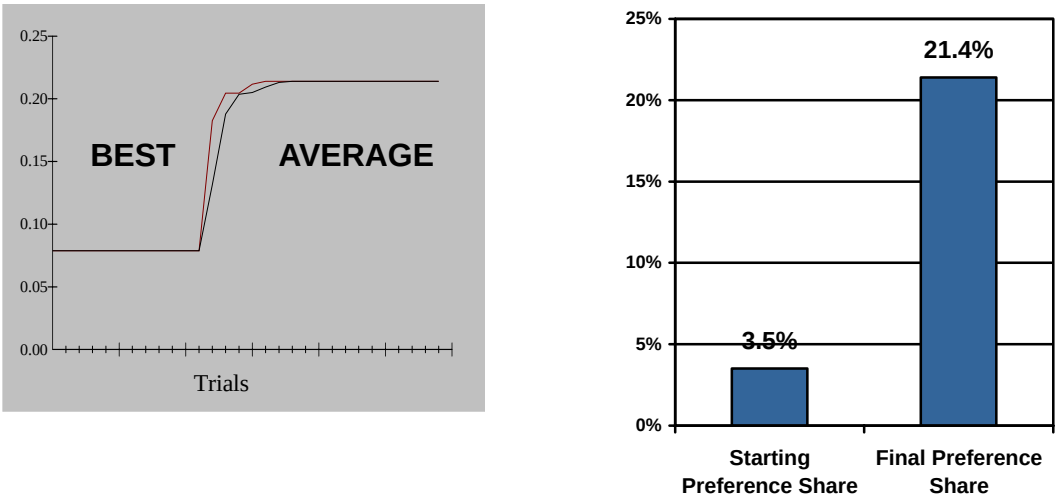
Conjoint methods are one of the most popular methods for identifying optimal product or service configurations. Utility values are estimated for attributes that differentiate the alternatives between existing and potential offerings in a particular product or service category. Optimization based on overall utility is straightforward: for any individual, the "best" product is comprised of the levels for each feature that have the highest utility. Optimization that incorporates the marketer's "loss function" — marginal cost or profit margin, for instance — is more complex. Brute force optimizers are useful for problems where we wish to optimize only one alternative in a competitive set, but such optimization is computer-intensive.

The product profiles in a typical conjoint study are analogous to the chromosomes in a genetic algorithm. Each attribute corresponds to a gene, and each level represents an "allele" — or value that the gene can take on. The genetic algorithm begins by generating several (perhaps 20-100) random product configurations and evaluating each against the "fitness" criterion used by the selection operator. The fitness function can be any value that can be calculated using a market simulator: total utility, preference share, expected revenue per unit (preference share X selling price), and so forth. The chromosomes ("products") in each generation then are allowed to mate (usually, as noted above, with probability proportional to fitness) and the crossover and mutation operators are applied to create the offspring.

The way in which the variables or product attributes are encoded should be considered carefully in setting up a GA for product optimization. For example, since a single product cannot usually include more than one level of each attribute, encoding an attribute as a string of zeros and ones (as in dummy variable coding) will necessitate additional rules so that only one level of each feature is present in each chromosome. It may be simpler in many cases to encode the levels of each attribute as a single integer (1, 2, 3, etc.). The position of the attributes in the chromosome may be important as well, since the crossover operator tends to preserve short segments of the chromosome, creating linkage between attributes.

Figure 4 shows the results of applying a GA to the optimization of a new auto model. A total of twelve attributes were varied. In this case the fitness or objective measure was preference share. The GA was instructed to maximize the preference share for this new vehicle against a specific set of competing vehicles. The starting value for the fitness measure was the manufacturer's "base case" specification.

Figure 4.
Single Product Optimization



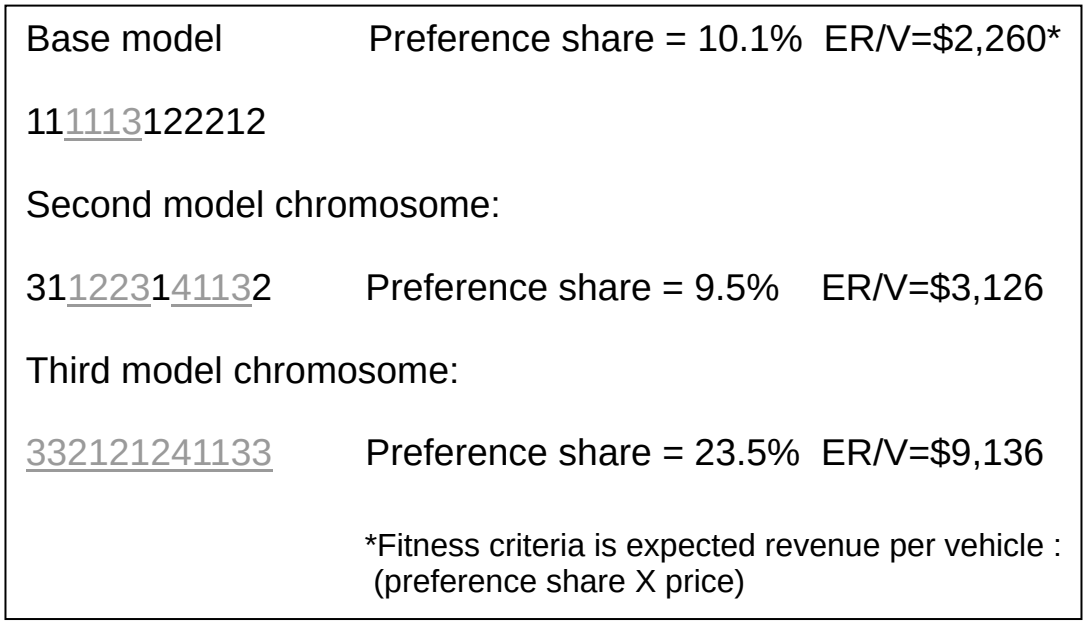
In this example, the GA ran for 1000 generations or trials (taking 8 minutes 35 seconds). The optimized product represents a significant improvement over management's "best guess" about the product to introduce. In all, 10 of the 12 attributes in the model changed levels from the base case to the "best" case. Because, as noted previously, GAs may not converge on the absolute best solution, manually altering some or all of the features of the final solution, one at a time, may reveal a slightly better solution.

PRODUCT LINE OPTIMIZATION

Identifying optimal product lines consisting of two or more similar products, such as different "trim levels" for a line of automobiles, is a more complex problem, since there are many more possible combinations. However, with an appropriately designed market simulator (i.e., one that is not subject to IIA), product line optimization using a GA is fairly simple. The fitness function can be combined across the variants in the product line — for example, total preference share, or total expected revenue. To ensure that the GA does not produce two or more identical products, it is necessary to specify at least one fixed difference between the variants.

Figure 5 expands our automotive example from a single vehicle to a three vehicle line-up. We start with a single vehicle optimization. In this case, the starting point is the best "naïve" product — for each attribute, the level that has the highest average part-worth is included. Because the goal of the product line optimization is to identify a base model and two alternatives, we allow only four features to vary for the base model. These are the underlined features. For the second model, four more features are allowed to vary, and for the third, all twelve features were varied in the optimization.

Figure 5.
Product Line Optimization



With only a single product in the line-up, maximum preference share was 17%, and the expected revenue per vehicle was $3,852. Adding two additional vehicles with differing optional equipment raises the total preference share for the make to 43% and expected revenue per vehicle increases to $14,522.

TURF AND TURF-LIKE COMBINATORIAL OPTIMIZATION

TURF (for "total unduplicated reach and frequency") analysis usually refers to a "brute force" (complete iteration) method for solving the n-combinatorial problem of finding the set of product features (or flavors, colors, etc.) that will lead to the greatest possible *reach* or penetration for a product. TURF analysis is suited to problems where the taste variation among the target market is characterized by a large proportion of the population liking a few flavors in common, such as chocolate, vanilla and strawberry, and many smaller subgroups with one or more idiosyncratic preferences. The object of the analysis is to find a combination of items, optional features, or flavors, that minimizes the overlap between the preferences, so that as many customers as possible find a flavor that satisfies their preferences.

Genetic algorithms offer a faster computational alternative to complete iteration TURF analysis. The critical step in applying a GA is the definition of the fitness measure. Examples of appropriate fitness measures include the simple match rate as well as a match rate weighted for either self-explicated or derived importances. Consider, for example, data from a "design your own product" exercise. Typically, survey respondents are asked to configure a product from a list of standard and optional features. The total problem space can be quite large. Even a small problem might have as many as 100,000 or more possible combinations. Design your own product questions may include some

mutually exclusive options (6 vs. 8-cylinder engine, for example). This creates a problem for traditional TURF analysis, since the solution with minimal overlap by definition will contain both engine options.

A more typical problem might be the selection of a set of prepared menu items for a "mini-mart." The task is to identify a small set of items to offer that will maximize the "satisfaction" of mini-mart customers who purchase prepared menu items in this "fast service" environment. The managerial goal is to reduce the cost of providing prepared food items. Survey respondents were asked their purchase intent (on a five point scale) for several menu items. The data were transformed so that an item with a top two box (definitely or probably purchase) response was coded as 1, and items with bottom three box responses were coded as 0. The fitness function was the percent of respondents with "matches" to the candidate solution on a specified number of items — three, four, or five items, for example. In this particular case, we wanted to find the one item to add to the three most popular items. The best gain in total reach for this problem, using the genetic algorithm, was 2%. The three items that generated the greatest unduplicated reach before adding a fourth item were all beverages and had, on average, lower overlap with each other than with any of the other items. This made it difficult to find an item that would have a noticeable impact on total reach.

With importance or appeal data for each feature, an importance "weight" can be factored into the analysis. For a given string, fitness is determined by adding up the importance weights of the features that match those selected by each individual respondent. Those solutions with higher importance-weighted match rates survive to reproduce in the next generation, until no better solution can be found.

Because the match rate is calculated for each individual, the GA makes segment-level analysis fairly straightforward. Additional functions, such as feature marginal cost or other loss functions, are easily incorporated into the fitness measure. Finally, even for large problems, genetic algorithms arrive at solutions very quickly, making it possible to test several feature sets of different size, for example, or different fitness measures.

SIMULATING MARKET EVOLUTION

Most conjoint-based market simulations are based on a static competitive environment. At best, some of the competitive characteristics might be changed to see what the impact of a specific response, such as a price reduction, will have on the predicted share of a new product. However, markets evolve, and the introduction of a new product may trigger a number of reactions among both competitors and customers. Moreover, the competitive reactions that persist over time will be those that have the greatest fitness.

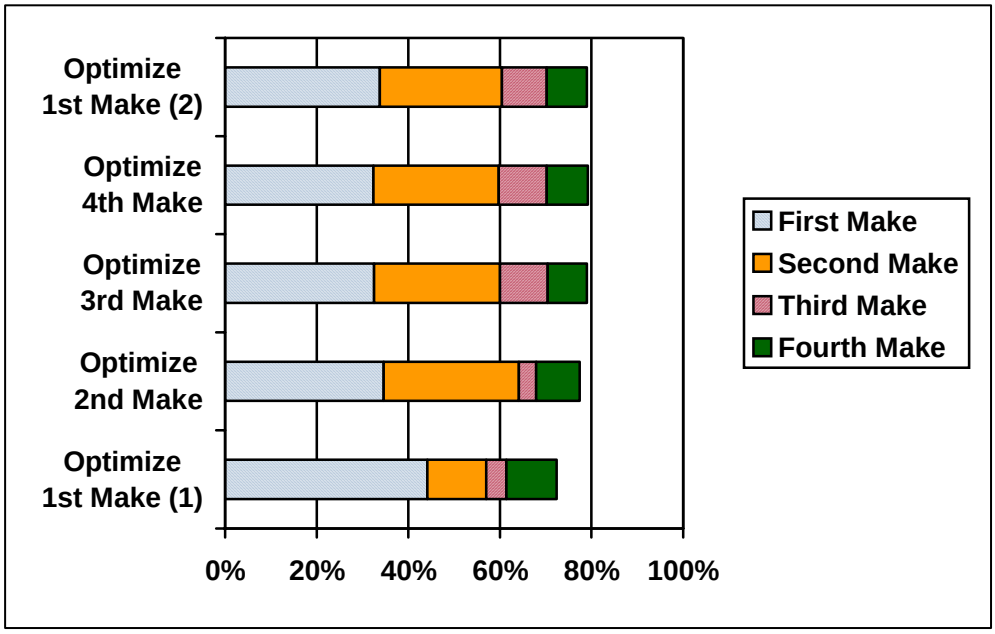
Genetic algorithms can be used to simulate potential competitive reactions over time. Rather than predetermine those reactions, as we do in static simulation, GAs will find the responses that have the greatest impact on the competitor's performance.

Returning to the automotive example used for product optimization, the GA was used to evolve several competing products sequentially. First, the new model was introduced into the current market and the best configuration identified. Next, a closely competing model was allowed to evolve in response to the new model. A second competitor was

next allowed to evolve, and then a third. Finally, the new model that was introduced in the first step was "re-optimized" against this new competitive context. In the actual marketplace, we would not expect purely sequential evolution of the competitors. A GA could be set up to allow simultaneous evolution of the competitors, but for this simple demonstration, sequential evolution was employed. It's important to note that we could also allow customer preferences to evolve over time, perhaps allowing them to become more price sensitive.

Figure 6 shows the results of the sequential evolution of the market. When the new product enters, it achieves preference share of almost 45%. The first competitor responds, gaining significantly. The third competitor manages to double its preference share, but the fourth competitor never fully recovers from the loss suffered when the new model entered the market. The market simulator (and the underlying choice model) was designed so that the total vehicle price depended on the features included in the vehicle. Therefore, changes in preference share are due primarily to changes in the included features, rather than changes in pricing for a fixed set of features.

Figure 6.
Simulating Market Evolution



WHEN TO USE GENETIC ALGORITHMS

Genetic algorithms may be the best method for optimization under certain conditions. These include:

- When the search space is large (having many possible solutions). For problems with a "small" search space, exhaustive methods will yield an exact solution.
- When the search space is "lumpy," with multiple peaks and valleys. For search spaces that are "smooth" with a continuously increasing objective function — to a global maximum — hill climbing methods may be more efficient.

- When the search space is not well understood. For problems where the solution space is well understood, domain-specific heuristics will outperform genetic algorithms.
- When the fitness function is noisy, or a "good enough" solution is acceptable (in lieu of a global maximum). For many optimization problems in marketing, the difference in the objective function will typically be very small between the best solution and one that is almost as good.

IMPLEMENTATION GUIDELINES

- Following a few simple guidelines will increase the effectiveness of genetic algorithms for marketing optimization problems.
- Carefully consider the encoding and sequencing of the genes on the chromosome. In conjoint optimizations, for example, attributes that are adjacent are more likely to remain linked, especially if the probability of crossover is relatively low.
- Start with reasonable population sizes, such as 50 individuals. GAs exploit randomness, so very small populations may have insufficient variability, while large populations will take longer to run, especially for complex problems.
- Run the GA optimization several times. GAs appear to be very good at getting close to the optimal solution for conjoint simulations. However, due to attributes with low importance (and feature part-worths near zero), the GA may get stuck on a near optimal solution. Incorporating a financial component into the fitness function may help avoid this situation.

ADDITIONAL RESOURCES

R. Axelrod, 1984. *The Evolution of Cooperation*, Basic Books.

J. H. Holland, 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press (2nd edition, MIT Press, 1992).

M. Mitchell, 1996. *An Introduction to Genetic Algorithms*, MIT Press.

GALib — a library of GAs in C+++. (lancet.mit.edu/ga/).

Evolver: Genetic Algorithm Solver for Microsoft Excel, Palisade Corporation (www.palisade.com).

COMMENT ON BAKKEN

RICH JOHNSON
SAWTOOTH SOFTWARE

I'd like to thank David Bakken for a clear and useful introduction to GAs and their usefulness in marketing research. We at Sawtooth Software have had quite a lot of experience with search algorithms during the past year, and our experience supports his conclusions.

I would like to make one general point, which is that there are many search algorithms that can be useful in marketing research for optimizing products or product portfolios. While GAs are certainly among those, others are also effective.

In some recent work we used a data set containing conjoint partworths for 546 respondents on 13 attributes having a total of 83 levels. We assumed a market of six existing products, and sought specifications for a seventh product that would maximize its market share as estimated by RFC simulations. In addition to GAs we tested three other search methods, all of which were hill-climbing methods that iteratively attempt to improve a single estimate, rather than maintaining a population of estimates. We ran each of the four methods seven times.

The three hill-climbing methods all got the same answer every time, which was also the best solution found by any method. The GA never got that answer, though it came close. There were also large differences in speed, with hill-climbing methods requiring only about a tenth the time of GA.

My colleague Bryan Orme has also reported another set of runs comparing GAs with a hill-climbing method using several data sets and several different scenarios, with 10 replications in each case. He found that both the GA and the hill-climbing method got the same apparently optimal answer at least once in each case. For the more simple problems, both techniques always found the optimal answer. He also found the GA took about ten times as long as the hill-climbing method. He found that the GA more consistently found "good" answers, but with a similar investment in computer time, for example if the hill-climbing method was run more times from different random starting points, the results were comparable.

Although other methods seem competitive with GAs in these comparisons, I agree with David that there is a definite role for GAs in marketing research, and I agree with his advice about when to use them: when the search space is large, lumpy, or not well understood, or when "close" is good enough. And those conditions characterize much of what we do in marketing research.

ADAPTIVE CHOICE-BASED CONJOINT

RICH JOHNSON
SAWTOOTH SOFTWARE
JOEL HUBER
DUKE UNIVERSITY
LYND BACON
NFO WORLDGROUP

A critical aspect of marketing research is asking people questions that will help managers make better decisions. Adaptive marketing research questionnaires involve making those questions responsive to what has been learned before. Such adaptation enables us to use the information we know to make our questions more efficient and less tedious. Adaptive conjoint processes for understanding what a person wants have been around for 20 years, the most notable example being Sawtooth Software's Adaptive Conjoint Analysis, ACA (Sawtooth Software 1991). ACA asks respondents to evaluate attribute levels directly, and then to assess the importance of level differences, and finally to make paired comparisons between profile descriptions. ACA is adaptive in two important respects. First, when it asks for attribute importances it can frame this question in terms of the difference between the most and least valued levels as expressed by that respondent. Second, the paired comparisons are utility balanced based on the respondent's previously expressed values. This balancing avoids pairs in which one alternative is much better than the other, thereby engaging the respondent in more challenging questions.

ACA revolutionized conjoint analysis, as we know it, replacing the fixed full profile designs that had been the historic mainstay of the business. Currently, ratings-based conjoint methods are themselves being displaced by choice-based methods, where instead of evaluations of product concepts, respondents make a series of hypothetical choices (Huber 1997). Choice-based conjoint is advantageous in that it mimics what we do in the market place. We rarely rate a concept prior to choice, we simply choose. Further, even though choices contain less information per unit of interview time than ratings or rankings, with hierarchical Bayes we are now able to estimate individual-level utility functions.

The design issue in choice-based conjoint is determining which alternatives should be included in the choice sets. Currently, most choice designs are not adaptive, and the particular choice sets individuals receive are independent of anything known about them. What we seek to answer in this paper is whether information about an individual's attribute evaluations can enable us to ask better choice questions. This turns out to be a difficult thing to do. We will describe a method, Adaptive Choice-Based Conjoint (ACBC) and a study that tests it against other methods.

WHAT MAKES A GOOD CHOICE DESIGN?

A good design is one in which the estimation error for the parameters is as small as possible. The error theory for choice designs was developed in seminal work by Dan McFadden (1974). For an individual respondent, or for an aggregation of respondents whose parameters can be assumed to be homogeneous, the variance-covariance matrix of errors for the parameters has a closed form:

$$\sum_{\beta} = (Z'Z)^{-1}$$

Where Z's have elements: $$z_{jn} = P_{jn}^{1/2}(x_{jn} - \sum_{i=1}^{J_n} x_{in}P_{in})$$

The $\boldsymbol{z_{jn}}$ are derived from the original design matrix in which $\boldsymbol{x_{jn}}$ is a vector of features of alternative *j* in choice set *n,* and $\boldsymbol{P_{jn}}$ is the predicted probability of choosing alternative *j* in choice set *n*. These somewhat daunting equations have been derived in Huber and Zwerina (1997), and have a simple and compelling intuition.

The **Z-**transformation centers each attribute around its expected (probability weighted) value. Once centered, then the alternatives are weighted by the square roots of their probabilities of being chosen. Thus the transformation involves a within-set probability- centered and weighted design matrix. Probability centering attributes and weighting alternatives by the square roots of their probabilities of being chosen within each choice set lead to important implications about the four requirements of a good choice design. The first implication is that the only information that comes from a choice experiment derives from contrasts within its choice sets. This leads to the idea of *minimal overlap,* that each choice set should have as much variation in attribute levels as possible. The second principle that can be derived is that of *level balancing,* the idea that levels within attributes should be represented equally. For example, if I have four brands, the design will be more accurate if each of the four appear equally often in the choice design. The third principle is *utility balance,* specifying that each alternative in the set have approximately equal probability of being chosen. Utility balance follows from the fact that each alternative is weighted by the square root of its probability of being chosen. At the extreme, if one alternative was never chosen within a choice set, then its weight would be zero and the experiment would not contribute to an understanding of the value of its attributes. The final principle is *orthogonality,* which says that the correlation of the columns across the **Z** matrix should be as close to zero as possible.

While these principles are useful in helping us to understand what makes a good choice set, they are less useful in designing an actual choice experiment because they inherently conflict. Consider, for example, the conflict between orthogonality and utility balance. If one were able to devise a questionnaire in which probabilities of choice were exactly equal within each choice set, then the covariance matrix would be singular because each column of the Z matrix would be equal to a linear combination of its other columns. Generally speaking, there do not exist choice sets that simultaneously satisfy

all four principles, so a search method is needed to find one that minimizes a global criterion.

The global criterion most used is the determinant of the variance-covariance matrix of the estimated parameters. Minimizing this determinant is equivalent to minimizing the volume of the ellipsoid defining the estimation errors around the parameters. The determinant also has facile analytical properties (e.g. decomposability, invertability and continuous derivatives) that make it particularly suitable as an optimization measure. Efficient search routines have made the process of finding an optimal design much easier (Zwerina, Huber and Kuhfeld 1996). These applications tend to be used in a context where one is looking for a good design across people (Arora and Huber 2001; Sandor and Wedel 2001) that works well across relatively homogeneous respondents.

Adaptive CBC, by contrast, takes individual-level prior information and uses it to construct efficient designs on the fly. The process by which it accomplishes this feat is detailed in the next section.

ADAPTIVE CBC'S CHOICE DESIGN PROCESS

Adaptive CBC (ACBC) exploits properties of the determinant of the expected covariance matrix that enable it to find quickly and efficiently the next in a sequence of customized choice sets. Instead of minimizing the determinant of the inverse of the Z'Z matrix, ACBC performs the mathematically equivalent operation of maximizing the determinant of Z'Z, the Fisher information matrix. The determinant of Z'Z can be decomposed as the product of the characteristic roots of Z'Z, each of which has an associated characteristic vector. Therefore, if we want to maximize this determinant, increasing the sizes of the smallest roots can make the largest improvement. This, in turn, can be done by choosing choice sets with design vectors similar to the characteristic vectors corresponding to those smallest roots.

In an attempt to provide modest utility balance, the characteristic vectors are further modified so as to be orthogonal to the respondent's partworths. After being converted to zeros and ones most of the resulting utility balance is lost, but this means one should rarely see dominated choices, an advantage for choice experiments.

ACBC begins with self-explicated partworths, similar to those used by ACA, constructed from ranking of levels within attributes and judgments of importance for each attribute (see Sawtooth Software 1991). It uses these to develop prior estimates of the individual's value parameters. The first choice set is random, subject to requiring only minimum overlap among the attribute levels represented. The information matrix for that choice set is then calculated and its smallest few characteristic roots are computed, as well as the corresponding characteristic vectors. Then each alternative for the next choice set is constructed based on the elements of one of those characteristic vectors.

Once we have a characteristic vector from which we want to create a design vector describing an alternative in the proposed choice set, the next job is to choose a (0-1) design vector that best approximates that characteristic vector. Within each attribute, we assign a 1 to the level with the highest value, indicating that that item will be present in the design. An example is given in the figure below.

Figure 1
Building a Choice to Correspond to a Characteristic Vector

	Attribute 1			Attribute 2		
	L1	L2	L3	L1	L2	L3
Char. Vector	-.03	.73	-.70	.93	-.67	-.23
Design Vector	0	1	0	1	0	0

This relatively simple process results in choice sets that are focused on information from attribute levels that are least represented so far.

Although Adaptive CBC is likely to be able to find good choice sets, there are several reasons why these may not be optimal. These will be just listed below, and then elaborated upon after the results are presented.

1. The priors themselves have error. While work by Huber and Zwerina (1996) indicates that approximate priors work quite well, poor priors could result in less rather than more efficient designs.
2. The translation from continuous characteristic vector to a categorical design vector adds another source of error.
3. D-error is designed for pooled logit, not for hierarchical Bayes logit with its ability to accommodate heterogeneous values across individuals.
4. The Adaptive CBC process assumes that human error does not depend on the particular choice set. Customized designs, particularly those that increase utility balance, may increase respondent error level. If so, the increased error level may counterbalance any gains from greater statistical efficiency.

In all, the cascading impact of these various sources of error may lead ACBC to be less successful than standard CBC. The result of the predictive comparisons below will test whether this occurs.

AN EXPERIMENT TO TEST ADAPTIVE CBC

We had several criteria in developing a test for ACBC. First, it is valuable to test it in a realistic conjoint setting, with respondents, product attributes and complexity being similar to those of a commercial study. Second, it is important to have enough respondents so that measures of choice share and hit rate accuracy can differentiate among the methods. Finally, we want a design where we can project not only to within the same sample, but also to be able to predict to an independent sample, a far more difficult predictive test.

Knowledge Networks conducted the study, implementing various design strategies among approximately 1000 allergy suffers who were part of their web-based panel. Respondents made choices within sets of three unbranded antihistamines having

attributes shown in Table 1. There were 9 product attributes, of which 5 had three levels and 4 had two levels. Notice the two potentially conflicting price measures, cost per day and cost per bottle. We presented price information to all respondents both ways, but within each choice task only one of these price attributes appeared. We did not provide the option of "None," so altogether there were a total of 14 independent parameters to be estimated for each respondent.

Table 1
Attributes and Levels Used To Define the Choice Alternatives

Attribute	Level 1	Level 2	Level 3
1. Cost/day	$1.35	$.90	$.45
2. Cost/100x 24 dose	$10.80	$7.20	$3.60
3. Begins working in	60 minutes	30 minutes	15 minutes
4. Symptoms relieved	Nasal congestion	Nasal Congestion and Headache	Nasal, Chest Congestion and Headache
5. Form	Tablet	Coated tablet	Liquid capsule
6. Interacts with Monoamine Oxidase Inhibitors?	Don't take with MOI's	May take with MOI's	
7. Interacts with antidepressants	Don't take with antidepressants	May take with antidepressants	
8. Interacts with hypertension medication	No	Yes	
9. Drowsiness	Causes drowsiness	Does not cause drowsiness	

The form of the exercise was identical for all respondents with the only difference being the particular alternatives in the 12 calibration choice tasks. To begin, all respondents completed an ACA-like section in which they answered desirability and importance questions that could be used to provide information for computing "prior" self-explicated partworths. We were confident of the rank order of desirability of levels within 8 of the 9 attributes, so we only asked for the desirability of levels only for one attribute, tablet/capsule form. We asked about attribute importance for all attributes.

- Next respondents answered 21 identically formatted choice tasks.
- The first was used as a "warm-up" task and its answers were discarded.
- The next 4 were used as holdout tasks for assessing predictive validity. All respondents received the same choice sets.
- The next 12 were used to estimate partworths for each respondent, and were unique for each respondent.
- The final 4 were used as additional holdout tasks. They were identical to the initial 4 holdout tasks, except that the order of alternatives was rotated.

The respondents were randomly allocated to five experimental conditions each containing about 200 people whose calibration sets were determined by different choice design strategies. The first group received standard CBC questionnaires. CBC provides designs with good orthogonality, level balance, and minimal overlap, but it takes no account of respondents' values in designing its questions, and so makes no attempt at adaptive design. The second group saw choice sets designed by ACBC. It does not directly seek utility balance, although it does take account of estimated partworths, designing questions that provide information lacking in previous questions. The third group also received questions designed by the adaptive algorithm, but one "swap" was made additionally in each choice set, exchanging levels of one attribute between two alternatives to create more utility balance. The fourth group was identical to the third, except their choices had two utility-balancing swaps. Finally, a fifth group also received questions designed by the adaptive algorithm, but based on aggregate partworths estimated from a small pilot study. This group was not of direct interest in the present comparison, and will not be reported further, although its holdout choices were included in the test of predictive validity.

RESULTS

Before assessing predictive accuracy, it is useful to explore other ways the experimental conditions did and did not create differences on other measures. In particular, groups did not differ with respect to the reliability of the holdouts, with the choice consistency for all groups within one percentage point of 76%. Also, pre-and-post holdout choices did not differ with respect to choice shares. If the first holdouts had been used to predict the shares of the second holdouts, the average mean error would be 2.07 share points. These reliability numbers are useful in that they indicate how well any model might predict.

The different design strategies did differ substantially with respect to the utility balance of their choice sets. Using their final estimated utility values, we examined the difference in the utility of the most and least preferred alternatives in each of the choice sets. If we set this range at 1.0 for CBC it drops to .81 for ACBC, then to .32 for ACBC with one swap and .17 for ACBC with two swaps. Thus ACBC appears to inject moderate utility balance compared with the CBC, while each stage of swapping then creates substantially more utility balance. This utility balance has implications for interview time. While regular CBC and ACBC took around 9.15 minutes, adding one swap added another 15 seconds and two swaps another 25 seconds. Thus, the greater difficulty in the choices had some impact on the time to take the study, but less than 10%.

In making our estimates of individual utility functions, we used a special version of Sawtooth Software's hierarchical Bayes routine that contains the option of including within-attribute prior rankings as constraints in the estimation. This option constrains final partworths to match orders of each respondent's initial ordering of the attribute levels. We also tested using the prior attribute importance measures, but found them to degrade prediction. This result is consistent with work showing that self-explicated importance weights are less useful in stabilizing partworth values (van der Lans, Wittink, Huber and Vriens 1992). Since within-attribute priors do help, their impact on prediction is presented below.

In comparing models it is appropriate to consider both hit rates and share predictions as different measures of accuracy. Hit rates reflect a method's ability to use the 12 choices from each person to predict 8 holdout choices. Hit rates are important if the conjoint is used at the individual level, for example to segment customers for a given mailing. Share predictions, by contrast, test the ability of the models to predict choice share for holdout choices. Share predictions are most important when the managerial task is to estimate choice shares for new products. Hit rates are very sensitive to the reliability of individuals' choices, which in this case hovers around 76%. We measure success of share predictions with Mean Absolute Error (MAE). MAEs are sensitive mostly to bias, since unreliability at the individual level is minimized by aggregation across independent respondents.

Hit rates, shown in Table 2, demonstrate two interesting tendencies, (although none of the differences within columns is statistically significant). With unconstrained estimation there is some evidence that two swaps reduce accuracy. However, when constraints are used in estimation, hit rates improve for all groups, with the greatest improvement for the group with most utility balance. We will return to the reason for this main effect and significant interaction after noting a similar effect on the accuracy of the designs with respect to predicting choice share.

Table 2
Accuracy Predicting Choices
Percent of Holdouts Correctly Predicted by Different Design Strategies

Design Strategy	No HB Constraints	Within Attribute Constraints
Regular CBC	75%	77%
Adaptive CBC	74%	76%
Adaptive CBC + 1 Swap	73%	77%
Adaptive CBC + 2 Swaps	69%	79%

To generate expected choice shares we used Sawtooth Software's Randomized First Choice simulation method (Orme and Huber 2000). Randomized First Choice finds the level of error that when added to the fixed portion of utility best predicts holdout choice shares. It does this by taking 1000 random draws from each individual after perturbing the partworths with different levels of variation. The process finds the level of variation that will best predict the choice shares of that group's holdout choices. Since such a procedure may result in overfitting choice shares within the group, in this study we use the partworths within each group to predict the combined choice shares from the other groups.

Table 3 provides the mean absolute error in the choice share predictions of the four design strategies. For example, regular CBC had mean absolute error of 3.15 percentage points, 20% worse than the MAE of 2.61 for ACBC. Without constraints, the new ACBC method was the clear winner. When the solutions were constrained by within-attribute information, all methods improved and again, as with hit rates, groups with greatest utility balance improved the most. With constraints, as without constraints, ACBC remained the winner. We are not aware of a statistical test for MAEs, so we cannot make statements about statistical significance, but it is noteworthy that ACBC with constraints has an MAE almost half that of regular CBC (without constraints).

Table 3
Error Predicting Share
Mean Absolute Error Projecting Choice Shares for Different Design Strategies

Design Strategy	No HB Constraints	Within Attribute Constraints
Regular CBC	3.15*	2.28
Adaptive CBC	2.61	1.66
Adaptive CBC + 1 Swap	5.23	2.05
Adaptive CBC + 2 Swaps	7.11	3.06

* Read: Regular CBC had an average absolute error in predicting choice shares for different respondents of 3.15 percentage points.

Why were constraints more effective in improving the designs that included swaps? We believe this occurred because swaps make the choice model less accurate by removing necessary information from the design. In our case, the information used to do the balancing came from the prior estimates involving the rank orders of levels within attributes. Thus, this rank-order information used to make the swaps needs to be added back in the estimation process.

An analogy might make this idea more intuitive. Suppose in a basketball league teams were handicapped by being balanced with respect to height, so that swaps made the average height of the basketball players approximately the same at each game. The result might make the games closer and more entertaining, and could even provide greater opportunity to evaluate the relative contribution of individual players. However, such height-balanced games would provide very little information on the value of height *per se*, since that is always balanced between the playing teams.

In the same way, balancing choice sets with prior information about partworths appears to make the individual utility estimates of partworths less precise. We tested this explanation by examining the relationship between utility balance and the correlation between priors and the final partworths. For (unbalanced) CBC, the correlation is 0.61, while for Adaptive CBC with two swaps, this correlation drops to 0.36, with groups having intermediate balancing showing intermediate correlations. Bringing back this prior information in the estimation stage raises the correlation for all the methods increase to a consistent 0.68.

SUMMARY AND CONCLUSIONS

In this paper we presented a new method using a characteristic-roots-and-vectors decomposition of the Fisher information matrix to develop efficient individual choice designs. We tested this new method against Sawtooth Software's CBC and against Adaptive CBC designs that had additional swaps for utility balance. The conclusions relate to the general effectiveness of the new method, the value of swapping and the benefit from including priors in the estimation stage.

Adaptive Choice-Based Conjoint

For those performing choice-based conjoint, the relevant question is whether the new adaptive method provides a benefit over standard CBC, which does not alter its design

strategy depending on characteristics of the individual respondent. We find that the two techniques take about the same respondent time. In terms of accuracy, there are no significant differences in predicting individual choice measured by hit rates. However, the new method appears to be more effective at predicting aggregate choice shares, although we are not able to test the statistical significance of this difference.

While we take comfort in the fact that this new prototype is clearly no worse than standard CBC, an examination of the four points of potential slippage discussed earlier offers suggestions as to ways Adaptive CBC might be improved. The first issue is whether the priors are sufficiently accurate in themselves to be able to appropriately guide the design. The second issue arises from the imprecision of approximating continuous characteristic vectors with zero-one design vectors. The third issue focuses on the appropriateness of minimizing D-error, a criterion built around pooled analysis, for the individual estimates from hierarchical Bayes. The final issue is whether human error from more difficult (e.g. utility balanced) choices counteracts any efficiency gains. Below we discuss each of these issues.

The first issue, basing a design on unreliable prior estimates, suggests a context in which the adaptive procedure will do well relative to standard CBC. In particular, suppose there is relatively low variability in the partworths across subjects. In that case, the HB procedure will do a fine job of approximating the relatively minor differences in values across respondents. However, where there are substantial differences in value across respondents then even the approximate adjustment of the choice design to reflect those differences is likely to be beneficial in being able to differentiate respondents with very different values from the average.

The second issue relates to additional error imparted by fitting the continuous characteristic vectors into a categorical design vector. The current algorithm constructs design vectors from characteristic vectors on a one-to-one basis. However, what we really need is a set of design vectors that "span the space" of the characteristic vectors, but which could be derived from any linear transformation of them. Just as a varimax procedure can rotate a principal components solution to have values closest to zero or one, it may be possible to rotate the characteristic vectors to define a choice set that is best approximated by the zeros and ones of the design vectors.

The third issue relates to the application of D-error in a hierarchical Bayes estimation, particularly one allowing for constraints. While the appropriateness of the determinant is well established as an aggregate measure of dispersion, in hierarchical Bayes one needs choice sets that permit one to discriminate a person's values from the average, with less emphasis on the precision of the average *per se*. Certainly more simulations will be needed to differentiate a strategy that minimizes aggregate D-error from one that minimizes error in the posterior estimates of individual value.

The final issue involves increasing possible human error brought about by the adaptive designs and particularly by their utility balance. As evidence for greater task difficulty we found that the utility balanced designs took longer, but by less than 10%. Notice, however, that the increased time taken may not compensate for difficulty in the task. It is possible the error around individual's choices also increases with greater utility

balance. That said, it will be difficult to determine the extent of such increased errors, but they will clearly limit the effectiveness of the utility balance aspect of adaptive designs.

Utility Balance

Along with orthogonality, minimal overlap and level balance, utility balance is one of the factors contributing to efficiency of choice designs. We tested the impact of utility balance by including one or two utility-balancing swaps to the Adaptive CBC choice sets. Unless constraints were used in the estimation process, two swaps degraded both hit rate and MAE accuracy. It is likely that the decay in orthogonality induced by the second swap combined with greater individual error to limit accuracy. However, if too much utility balance is a bad thing, a little (one swap) seems relatively benign. Particularly if constraints are used, then it appears that one swap does well by both the hit rate and the choice share criteria.

The general problem with utility balance is that it is easy to characterize, but it is hard to determine an optimal level. Some utility balance is good, but too much quickly cuts into overall efficiency. D-error is one way to trade off these goals, but limiting the number of swaps may not be generally appropriate. For example, for a choice design with relatively few attributes (say 3 or 4), one swap should have a greater impact than in our study with nine attributes. Further, the benefit of balancing generally depends on the accuracy of the information used to do the balancing. The important point here is that while any general rule of thumb recommending one but not two swaps may generally work, but it will certainly not apply over all circumstances.

Using Prior Attribute Orders as Constraints

Using individual priors of attribute level orders as constraints in the hierarchical Bayes analysis improved both hit rates and share predictions. It is relevant to note that using prior importance weights did not help; people appear not to be able to state consistently what is important to them. However, the effectiveness of using the rankings of levels within attributes suggests that choices do depend importantly on this information.

Using this within-attribute information had particular value in counteracting the negative impact of utility balancing. Utility balancing results in less precision with respect to the information used to do that balancing. Thus it becomes important to add back this information in the analysis aspect that was lost in the choice design.

In the current study most of the attributes were such that respondents agreed on the order of levels. Sometimes these are called “vector attributes,” for which people agree that more of the attribute is better. Examples of vector attributes for antihistamines include speed of action, low price and lack of side effects. By contrast, there are attributes, such as brand, or type of pill or bottle size, on which people may reasonably disagree with respect to their ordering. Where there is substantial heterogeneity in value from non-vector attributes we expect that optimizing design on this individual-level information and using priors as constraints should have even greater impact than occurred in the current study.

In conclusion, the current study gives reason to be optimistic about the effectiveness of adaptive choice design. It is likely that future research will both improve the process by which prior information guides choice design and guide changes in design strategies that adjust to different product class contexts.

REFERENCES

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization of Choice Experiments," *Journal of Consumer Research, 26:2* (September) 273-283.

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

Huber, Joel (1997) "What We Have Learned from 20 Years of Conjoint Research: When to Use Self-Explicated, Graded Pairs, Full Profiles or Choice Experiments," Sawtooth Software Proceedings 1997: Available at http://www.sawtoothsoftware.com/download/techpap/whatlrnd.pdf.

McFadden, Daniel (1974) "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics,* P. Zaremka, ed. New York, Academic Press, 105-142.

Orme, Bryan and Joel Huber (2000), "Improving the Value of Conjoint Simulations," *Marketing Research*, 12 (Winter), 12-21.

Sandor, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Manager's Prior Beliefs," *Journal of Marketing Research, 38* (November), 430-44.

Sawtooth Software (1991), "ACA System: Adaptive Conjoint Analysis," Available at http://www.sawtoothsoftware.com/download/techpap/acatech.pdf

Sawtooth Software (1999), "Choice-Based Conjoint (CBC)," Available at http://www.sawtoothsoftware.com/download/techpap/cbctech.pdf

van der Lans, Ivo A., Dick Wittink, Joel Huber and Marco Vriens (1992), "Within- and Across-Attribute Constraints in ACA and Full Profile Conjoint Analysis," Available at http://www.sawtoothsoftware.com/download/techpap/acaconst.pdf

Zwerina, Klaus, Joel Huber and Warren Kuhfeld (1996), "A General Method for Constructing Efficient Choice Designs," Available at http://support.sas.com/techsup/technote/ts677/ts677d.pdf