PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

February 1999

FOREWORD

We are pleased to present the Proceedings of the Seventh Sawtooth Software Conference, held in San Diego, California in February 1999. As with previous conferences, the focus was quantitative methods in marketing research. This year saw many authors reporting on commercial and methodological studies involving Choice-Based Conjoint. Notably, four presentations used actual sales to validate their conjoint models, which is a welcomed practical contribution. In contrast to recent conferences, relatively more emphasis was placed on data collection, with a number of talks focussing on alternative data collection methods, particularly the Internet.

For the first time at a Sawtooth Software Conference, a “Most Valuable Presentation” award was voted on by conference attendees and awarded at the end of the final session. Joel Huber (co-authors Bryan Orme and Dick Miller) received the award for his presentation entitled “Dealing with Product Similarity in Conjoint Simulations.”

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. Many discussants have prepared written 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 indebted to the authors and discussants for making this conference a success and for advancing our collective knowledge in this exciting field.

Some of the papers presented at this and previous conferences are available in electronic form at our Technical Papers Library on our home page: http://www.sawtoothsoftware.com.

Sawtooth Software

April, 1999

CONTENTS

EVALUATING THE REPRESENTATIVENESS OF ONLINE CONVENIENCE SAMPLES 1

Karlan J. Witt, IntelliQuest, Inc.

CONJOINT ON THE WEB - LESSONS LEARNED 9

Michael Foytik, DSS Research

BUT WHY? PUTTING THE UNDERSTANDING INTO CONJOINT 23

Ray Poynter, Deux

Comment by *Sam Kingsley* 31

CONVENTION INTERVIEWING — CONVENIENCE OR REALITY? 33

Don Marshall, Isis Research

DISK-BASED MAIL SURVEYS: A LONGITUDINAL STUDY OF PRACTICES AND RESULTS 43

Arthur Saltzman, Para Technologies and *William H. MacElroy,* Socratic Technologies

WHAT WILL WORK OVER THERE? COMPUTER-BASED QUESTIONNAIRES
IN FOREIGN LANGUAGES 55

Brent Soo Hoo and *Lori Heckmann,* Griggs-Anderson Research

Comment by *Karlan J. Witt* 69

EFFICIENT FEE STRUCTURES FOR MUTUAL FUNDS 71

Ronald T. Wilcox, Carnegie Mellon University

Comment by *Michael G. Mulhern*.. 99

KNOWING WHEN TO FACTOR: SIMULATING THE TANDEM APPROACH TO CLUSTER ANALYSIS........ 101

Andrew Elder, IntelliQuest, Inc.

THE NUMBER OF LEVELS EFFECT: A PROPOSED SOLUTION 109

Dick McCullough, MACRO Consulting, Inc.

Comment by *Dick R. Wittink*........ 117

MATCHING CANDIDATES WITH JOB OPENINGS USING WEB-BASED ADAPTIVE CONJOINT 123

Man Jit Singh, Futurestep and *Sam Kingsley,* dataDirect

PREDICTING PRODUCT REGISTRATION CARD RESPONSE RATES WITH CONJOINT ANALYSIS 131

Paul Wollerman, The Polk Company

CONJOINT ANALYSIS ON THE INTERNET 145

Jill S. Johnson, The Bonham Group Market Research Company, *Tom Leone,* MediaOne, and *John Fiedler,* POPULUS, Inc.

TWO ORDINAL REGRESSION MODELS........ 149

Tony Babinec, SPSS, Inc.

FORECASTING SCANNER DATA BY CHOICE-BASED CONJOINT MODELS 169

Markus Feurstein and *Martin Natter,* Vienna University of Economics & BA, and *Leonhard Kehl,* Kehl Pricing Research & Consulting

PREDICTING ACTUAL SALES WITH CBC:
HOW CAPTURING HETEROGENEITY IMPROVES RESULTS........ 183

Bryan K. Orme, Sawtooth Software, Inc. and *Michael A. Heft,* Daymon Associates

USING SCANNER DATA TO VALIDATE CHOICE MODEL ESTIMATES........ 201

Jay L. Weiner, The NPD Group, Inc.

SHOULD CHOICE RESEARCHERS ALWAYS USE ‘PICK ONE’ RESPONDENT TASKS? 207

Jon Pinnell, MarketVision Research, Inc.

ASSESSING THE RELATIVE EFFICIENCY OF FIXED AND RANDOMIZED EXPERIMENTAL DESIGNS **225**

Michael G. Mulhern, Mulhern Consulting

Comment by *Richard M. Johnson* **233**

FULL VERSUS PARTIAL PROFILE CHOICE EXPERIMENTS: AGGREGATE AND DISAGGREGATE COMPARISONS **235**

Keith Chrzan, IntelliQuest, Inc.

Comment by *Bryan K. Orme* **249**

DEALING WITH PRODUCT SIMILARITY IN CONJOINT SIMULATIONS **253**

Joel Huber, Duke University, *Bryan K. Orme,* Sawtooth Software, Inc., and *Richard Miller,* Consumer Pulse

Comment by *Richard M. Johnson* **267**

A COMPARISON OF ALTERNATIVE SOLUTIONS TO THE NUMBER-OF-LEVELS EFFECT **269**

Dick R. Wittink, Yale University and *P. B. Seetharaman,* Washington University, St. Louis

Comment by *Joel Huber* **283**

USING LISREL AND PLS TO MEASURE CUSTOMER SATISFACTION **285**

Lynd D. Bacon, Lynd Bacon & Associates

Comment by *P. B. Seetharaman* **305**

PRODUCT MAPPING WITH PERCEPTIONS AND PREFERENCES **307**

Richard M. Johnson, Sawtooth Software, Inc.

SUMMARY OF FINDINGS

Evaluating the Representativeness of Online Convenience Samples (Karlan J. Witt): "The widespread adoption of the internet has provided a low-cost, rapid-turnaround alternative for conducting market research," Karlan explained. She warned against the notion that large sample sizes (in the thousands and tens of thousands) can alleviate problems of representativeness, citing the well-known erroneous prediction of 1936 presidential election results by *Literary Digest* due to a biased convenience sample.

Karlan compared the results of online convenience samples to those recruited through RDD (Random Digit Dialing). She found that the convenience sample differed significantly from RDD in terms of respondent demographics, online usage behavior, and brand share. She stressed that weighting the convenience sample to match known targets was not the solution. "In many cases it did make the data look more like the RDD data, but did not fully address the inherent non-response bias present in the online sample."

Karlan maintained that there are some applications in which online convenience samples may sometimes be appropriate: comparative studies, tracking studies where relative changes are the focus, attitudinal studies and segment identification and profiling studies.

Conjoint on the Web–Lessons Learned (Michael Foytik): Mike shared insights based on experience fielding both traditional full-profile (pairwise comparisons) and choice-based conjoint surveys over the Internet. Interestingly enough, Mike expressed that one of the greatest roadblocks to doing conjoint over the Internet was in convincing clients to accept a conjoint methodology.

Response rates to Internet surveys have exceeded response rates for mail surveys by 25% to 50%, though the novelty of taking web surveys will certainly wear off in the future and response rates will decline. Mike reported that the internal consistency and predictability of holdouts for Internet conjoint data has exceeded that of mail in split-sample tests.

For successful Internet conjoint surveys, Mike suggested making the survey adaptive in tailoring the consideration set or further customizing the conjoint tasks, assigning unique passwords, using clickable links to the survey in E-mail messages, providing a contact name and number on each Internet page, permitting restarts, and programming to the least common browser denominator.

But Why? Putting the Understanding into Conjoint (Ray Poynter): Ray pointed out that conjoint analysis has been a very useful technique for quantifying preferences for product features. Despite its popularity, Ray argued that this is often not enough: ". . . too often the researcher is then asked a question such as 'but why do 25% of the consumers prefer the orange, oval, tall jug?'. Failure to answer this seemingly straightforward question will frequently result in the client doubting the wisdom of both the researcher and of the executive who commissioned the project."

Ray described a simple BASIC program his firm developed for “replaying” a completed ACA interview. The computer displays the tradeoff questions again on the screen together with the respondent’s answer. A qualitative assessment is made together with the respondent regarding why he/she responded in a particular way. Ray explained how the replay feature can be used in designing the questionnaire to ensure that respondents understand the instrument and that the information gained will benefit the client.

Ray also advocated spending a good deal of time selecting the list of attributes, using pilot tests with self-explicated part worths to reduce the list, making sure respondents can understand the questionnaire, and weeding out redundancy across attributes. He concluded, “The Conjoint process is very powerful at determining preferences and quantities. However, it can only do that if it asks the right questions. The best way to ask the right questions is to have a thought through approach to questionnaire design and to utilise good qualitative inputs to that process.”

Convention Interviewing–Convenience or Reality? (Don Marshall): Don reviewed his experiences with conducting pharmaceutical research using conjoint research. Doctors, he claimed, are usually good conjoint subjects due to their scientific background and familiarity with making tradeoffs between such issues as safety and efficacy. However, they are difficult to recruit and expensive to interview. Physicians often attend medical conventions to learn about new developments. “These medical conventions clearly present a wonderful opportunity to complete a large number of physician interviews in a relatively short time frame,” Don explained, but “. . . they also have been criticized as providing a non-representative sample.” Central facility interviews have been proposed as a more representative technique.

Using a split-sample study, Don found no significant difference between the part worths resulting from convention versus central site recruitment. “Overall, these results indicate that convention interviewing is an efficient and economic alternative to central facility interviewing when interviewing doctors. . .” he determined.

Disk-Based Mail Surveys: A Longitudinal Study of Practices and Results (Arthur Saltzman and William H. MacElroy): The authors conveyed results from two surveys (1994, 1998) among market research professionals. The move toward “personalization”of DBM surveys was one of the most noticeable trends, they noted. More researchers are pre-notifying and qualifying respondents and sending personal outbound envelopes and cover letters. In spite of the increase in personalization and targeting, response rates to DBM surveys are falling. The reported average response rate to DBM surveys was 52% in 1994 and 33% in 1998.

Researchers participating in the 1998 research said that they are becoming more interested in on-line (Internet) interviewing rather than DBM, citing “ease of implementation, quicker response time, lower cost.” Other researchers noted the limitations and “newness” of the Web as obstacles and plan to continue using DBM in the future. The authors concluded: “Due to the rapid increase of Internet usage among the general population, online research is clearly the wave of the future . . . However, it is apparent that DBM still has its place, at least in the short run. This will be true as long as there remains a gap between PC users and online users . . . until PC usage and Internet usage are close to the same level, there will be a need for the DBM survey technique allowing for the broadest possible sample pool.”

What Will Work Over There? Computer-Based Questionnaires in Foreign Languages (Brent Soo Hoo and Lori Heckmann): The authors provided an overview of the different character representation and computing protocols in different countries. Most Anglo/European based alphabets are based on the ASCII set of codes. Standard English-based software systems (including systems from Sawtooth Software) can manage these languages. However, many countries use "multi-stroke" double-byte characters, such as in Japan. Custom programming is required when fielding studies with double-byte characters. Foreign language projects also involve a host of other complicating factors. The authors provided an extensive list of supporting articles, Website references, and books.

Efficient Fee Structures for Mutual Funds (Ronald T. Wilcox): Ron conveyed how choice-based conjoint analysis (and individual-level analysis under ICE) can be used to learn how consumers evaluate key attributes of a mutual fund. The weight consumers give to fees a fund charges can be used by fund managers to design the fee structures that will maximize utility for both the consumer and the fund manager.

Ron found that consumers place more weight on the past performance (particularly the 10-year horizon) of a fund than the finance literature suggests they should. Consumers, he maintains, are also irrationally more sensitive to loads than to expense ratios. Ron's research suggests new arenas in which conjoint analysis can be successfully employed. The findings can be useful for both government regulators of securities and mutual fund managers. Ron concluded, "We know so little, yet there could be truly substantial business practice and public policy implications resulting from an increased knowledge of consumers' decision processes in this marketplace."

Knowing When to Factor: Simulating the Tandem Approach to Cluster Analysis (Andrew Elder): Deciding how to pre-treat the input variables (standardization, centering, weighting or factor analyzed) is a critical first step to conducting any cluster analysis procedure**.** Andy reviewed some of the arguments made over the last few decades regarding whether to pre-process input variables through Principal Components analysis (the Tandem approach), or to use the variables in their raw form.

Andy pointed out the common argument that "if unequal numbers of raw variables load on the various factors, then clustering on the raw variables themselves . . . could bias the cluster solution in favor of over-represented factors." He also reiterated Rich Johnson's argument that using factor scores can "smooth out" the lumpiness of data that is useful in cluster analysis.

Andy used a synthetic data set to test alternative pre-processing approaches. He found that when there was a great degree of imbalance in the loading of input variables on independent factors, the tandem approach excels. Otherwise, raw scores performed better when the variables provide relatively balanced representation across factors. Andy concluded that the decision to use the Tandem approach versus raw variables should be data driven and admonished researchers to investigate the structure and independence of the input variables through factor analysis before making that decision for a particular data set.

The Number of Levels Effect: A Proposed Solution (Dick McCullough): Dick reviewed past evidence on the Number of Levels (NOL) Effect: “The effect occurs when one attribute has more or fewer levels than other attributes. For example, if price were included in a study and defined to have five levels (holding the range of variation constant), price would appear more important than if price were defined to have two levels.”

Dick reported results from two studies. The first featured rank-order data for traditional conjoint, and confirmed the existence of the NOL effect. In the second study, Dick introduced a clever solution to the NOL problem involving a two-stage design. In the first half of the design, only the most and least desirable levels were included (customized per respondent, based on stated preferences). The second half used all levels in the full study design. Respondents first completed the equal-levels exercise, followed by the complete design. As Dick explained: “The full-level utility estimates can then be linearly scaled into the two-level estimates. The resulting utilities will exhibit the correct attribute relative importance and also maintain the relative positions of levels within each attribute.”

While Dick’s solution appears on the surface quite straightforward, the interview typically must be computerized so that the custom design can be built on-the-fly. He noted that problems occur when respondents’ stated exterior levels do not match the derived external levels from the full levels exercise.

Matching Candidates with Job Openings (Man Jit Singh and Sam Kingsley): The authors demonstrated how ACA is being used on the Web in matching potential job candidates with job openings. Singh and Kingsley are the first to develop a Web-enabled version of ACA based on Sawtooth Software’s underlying code. To date, hundreds of thousands of applicants have completed their ACA survey on job preferences on the Web.

According to the authors, “the job search and recruiting business has traditionally relied on the ‘telephone tree’”, followed by a detailed review by senior associates, follow-up interviews, offers and counter-offers. By collecting much information up-front in a standard survey, followed by a more detailed assessment of the tradeoffs (ACA) each applicant makes regarding compensation and other job-related elements, a large number of applicants can be sorted through efficiently.

Because the level of analysis is at the individual, part worth reversals can be problematic. While ACA generally results in fewer reversals than other conjoint methods, they still can occur. Simple averaging and “tieing” procedures are one suggested remedy.

Predicting Product Registration Card Response Rates with Conjoint Analysis (Paul Wollerman): Paul used conjoint analysis to predict response rates to the “warranty cards” that manufacturers place in their products. In the past, “live tests” had been conducted to measure the actual return rate for alternative versions of warranty cards. “Implementing these tests,” Paul explained, “has always proved to be logistically difficult; plus, it takes a long time to obtain usable results.”

A conjoint survey that included “mocked-up” product registration cards and pictures of the products in which those cards were supposedly packed were mailed to respondents. Respondents indicated how likely they were to return each card, had they purchased the displayed appliance or product.

The conjoint predictions provided at face-value a reasonable *relative* fit with actual measured response rates for the same profiles from live tests. More importantly, Paul suggested that "the rank order and magnitude of the attributes' importance and the utility scores of the individual attributes are a good fit with our experience." Paul also noted that respondents grossly over-estimated their likelihood of returning cards relative to actual return rates. He concluded that "conjoint analysis can be used to measure consumers' preferences for product registration card design features, and predict how a given card will perform relative to another."

Conjoint Analysis on the Internet (Jill Johnson, Tom Leone, and John Fiedler): The authors provided a case study involving conjoint analysis over the Internet. The client was a major telecommunications company that wanted to measure preferences for a high-speed Internet access product. Johnson and Fiedler used Sawtooth Software's CVA Internet Module.

Respondents were recruited by telephone, and provided email addresses during the phone interview. The authors stressed how important it was to train telephone interviewers to exactly record the email addresses. Passwords were assigned for each respondent, and an email package called MailKing was used to send personalized email messages to each respondent.

The phone-recruit phase of the internet study was a key to the success of the project. According to the authors: "some at the company were skeptical of online methodology (due to concerns about representativeness). The initial telephone recruit combined with 48% response rate served to allay any concerns."

Two Ordinal Regression Models (Tony Babinec): Tony explained that ordinal variables abound in marketing research. Purchase likelihood, satisfaction scores, and income groups are common examples. "It is probably a safe guess," noted Tony, "that many if not most researchers scale these variables using sequential integer scores . . . and then analyze them using conventional linear regression. Doing so involves the implicit assumption that the intervals between adjacent categories are equal."

Tony argued that there can be "serious statistical problems" if one uses OLS regression with ordinal data. Two methods proposed to deal with ordinal responses are the "cumulative logit model" and the "adjacent category logit model." Using a sample data set, Tony showed that "the adjacent-categories logit model can work well in situations where the cumulative logit model does not fit the data." "It is hoped," Tony concluded, "that this paper will help spur the wider use of this new modeling approach."

Forecasting Scanner Data by Choice-Based Conjoint Models (Markus Feurstein, Martin Natter, and Leonhard Kehl): The authors present solid evidence that correctly tuned CBC models can accurately predict actual sales (as reported by scanner data). They also found that ignoring the existence of heterogeneity in choice-based conjoint modeling "leads to biased forecasts."

Alternative methods for capturing heterogeneity are tested: *a priori* segments: regional segmentation and usage frequency; K-means segments based on demographics; latent class; and individual-level logit estimation. The Latent Class model performed the best in predicting actual sales, followed by individual estimation. The other methods of recognizing heterogeneity only marginally improved predictions relative to the aggregate model, which performed the worst.

The authors found that it was necessary to include external effects for accurately predicting market shares. When quarterly corrections were made (on longitudinal weekly sales data), share predictions for the remainder of the quarter were quite accurate.

Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results (Bryan K. Orme and Michael A. Heft): The authors presented evidence that, under proper conditions, conjoint analysis can accurately predict what buyers do in the real world. Their results were based on CBC interviews conducted in grocery stores, where the CBC results were used to predict actual sales for three product categories of packaged goods from those same stores with good success.

Additionally, Orme and Heft showed that capturing heterogeneity (reflecting differences in preference between groups or individuals) with Latent Class or ICE can improve predictions. Many complex effects (substitution, cross-effects and interactions) can be accounted for nearly "automatically" with disaggregate Main Effect models. They noted that complex terms can be built into large aggregate logit models, but that such models risk overfitting. Moreover, that approach places a great deal of responsibility on the analyst to choose the right combination of terms.

Using Scanner Data to Validate Choice Model Estimates (Jay L. Weiner): The main purpose of Jay's paper was to "compare three forms of multinomial logit model estimation . . . main-effects only, main-effects with interactions and brand-specific parameter estimation." The key advantage to the latter approach would be to "allow the model to deal with the Independence of Irrelevant Alternatives (IIA) problem." Additionally Jay addressed whether using a constant sum allocation (next 10 purchases) versus a First Choice approach significantly affected results in his discrete choice study.

Having IRI scanner data as a criterion for predictive validity was a strong component of Jay's paper. Jay found that "The main-effects only model seems to be quite robust and efficient for fast-moving consumer package goods." The choice results did a good job in predicting actual market shares, with a Mean Absolute Error of between 2 to 4%. "It is clear, however, that there are different price curves for each brand," he cautioned. "To gain key insight into these effects, estimating interactions works well." Jay also reported success with the constant-sum allocation model, noting that "the additional degrees of freedom gained from using 'next 10' purchases enhance the predictability of the model." Modeling brand-specific effects also slightly improved the predictive validity of the models, relative to main-effects and main-effects plus interactions models.

Should Choice Researchers Always Use "Pick One" Respondent Tasks? (Jon Pinnell): Jon presented results from a commercial study in which different respondents received four alternative types of choice tasks. He tested two non-metric approaches: First Choice and Full Rank Order; and two metric approaches: Constant Sum Allocation and "Scaling". (Under the "Scaling" option, respondents identified the best and worst concepts in the task. The best was assigned a 100, and the worst 0, and respondents positioned the remaining concepts within that frame.)

Jon reported that the First Choice task took the least time for respondents to complete, followed by Rank Order, Allocation and Scaling. Respondents could answer ten first choice tasks in the time it took to complete just short of six Allocation exercises. All methods resulted in similar utilities and inferred importances, but Jon found that the non-metric methods were more reliable than the metric approaches. He concluded: “The metric methods (especially Scaling) appear very expensive relative to first choice. For the time required, they offer apparently little information above the first choice utilities.” Jon also was “intrigued by what appears to be the processing differences between First Choice and the Rank Order tasks” and speculated that respondents might “make better first choices if they know that they will also be asked to make second choices as well.”

Assessing the Relative Efficiency of Fixed and Randomized Experimental Designs (Michael G. Mulhern): Much research has been presented regarding design efficiency in choice analysis, particularly for fixed designs. Mike presented results comparing fixed and randomized designs. The main question he focused on was the relative gains in efficiency as more versions of fixed designs were added to the pool. Mike commented that “There is some evidence that the total number of choice sets required may differ for symmetric and asymmetric designs. A symmetric design is an experimental design where each attribute contains the same number of levels. An asymmetric design contains attributes with varying numbers of levels.”

Using computer-generated data, Mike assessed the relative efficiency of different designs. He found, “For the large symmetrical choice experiment, the randomized design is 95% as efficient as the optimal fixed design with approximately 190 choice sets. For the large asymmetrical choice experiment investigated in this study, it was found that the randomized design is approximately 14% more efficient with 980 choice sets than the fixed asymmetric design with 49 choice tasks.” The conclusion that randomized designs can actually be more efficient than fixed orthogonal designs for asymmetric designs is an important finding that lends even greater credibility to the use of the easy-to-implement randomized designs in CBC.

Full versus Partial Profile Choice Experiments: Aggregate and Disaggregate Comparisons (Keith Chrzan): In contrast to full-profile (FP) experiments in which concepts contain every attribute, partial-profile experiments (PP) show only a subset of the attributes (typically five) at a time. A key benefit, Keith argued, of PP choice designs is the ability to measure many more attributes than is traditionally thought prudent with FP choice designs.

Keith conducted a split-sample choice study over the Internet using an Internet panel. Half of the respondents received a FP experiment, and the other half received a PP. Fifteen choice tasks for FP and 18 tasks for PP (with two concepts each) were included, followed by holdout choices. Nine attributes were studied.

In contrast to previous studies which have reported very little difference in the logit part-worths, he found some statistically significant differences between full- and partial-profile choice effects. Despite the differences, both full- and partial-profile simulations provided roughly equally accurate predictions of holdout choice shares, no matter the method of analysis.

Keith analyzed his data using three alternative methods: aggregate logit, ICE and Hierarchical Bayes. He found that utility estimates from aggregate logit and HB predicted aggregate holdout choice shares with reasonable accuracy, though neither method prevailed. ICE did not perform as well, leading Keith to speculate that “the relative sparseness of the partial-profile choice experiment data set could have been expected to hamper the operation of ICE.” Keith reported that ICE predictions for the full-profile cell also were poor. ICE’s poor showing, he explains “ . . . probably reflects ICE’s instability when too few observations are collected from each individual respondent." An exciting implication of his study was that useful individual-level utilities may be possible from partial profile choice experiments.

Dealing with Product Similarity in Conjoint Simulations (Joel Huber, Bryan K. Orme and Richard Miller): The authors tackled a long-standing problem by offering a new simulation approach for dealing with product similarity. They explained that traditional conjoint simulators based on the BTL or logit model have suffered from IIA problems. Similar or identical products placed in IIA simulators tend to result in "share inflation." The first choice model, while not susceptible to the IIA difficulty of unrealistic share inflation for similar offerings, typically has produced shares of preference that are too extreme relative to real world behavior. Also, first choice models are inappropriate for use with logit or latent class models. In the family of Sawtooth Software products, a Model 3 "Correction for Product Similarity" has been offered to deal with problems stemming from product similarity. However, this model is often too simplistic to accurately reflect real world behavior.

The authors proposed a new method called "Randomized First Choice (RFC)" for tuning market simulators to real world behavior. RFC adds random variation to both attribute part-worths and to the product utility, and simulates respondent choices under the first choice rule. They demonstrated how RFC can be tuned to reflect any similar product substitution behavior between the extreme first choice rule and the IIA-grounded logit rule. They also showed that RFC improved predictions of holdout choice tasks (reflecting severe differences in product similarity) for logit, latent class, ICE and hierarchical Bayes. The greatest gains were for the aggregate methods. The disaggregate methods, while less in need of corrections for product similarity, still benefitted from RFC. This paper was voted "Most Valuable Presentation" by attendees at the conference.

A Comparison of Alternative Solutions to the Number-of-Levels Effect (Dick R. Wittink and P.B. Seetharaman): The authors restated the danger of the Number of Levels (NOL) effect: ". . . the distance between the largest and smallest part worths of an attribute is a positive function of the number of intermediate levels, holding other things constant." They reported on a ratings-based full profile study which showed that the insertion of two interior levels (holding the range constant) produced a greater effect on derived attribute importances than an expansion of the range to three times the original range (holding the number of levels constant). This demonstrated that the NOL effect is potentially huge, especially in full-profile data (whereas in ACA 3.0 it seems to be only half the magnitude).

They also reported on a methodological study involving both ACA (version 3) and CBC. The NOL effect was found to be more prevalent in CBC. A customized ACA approach was proposed in which the number of levels describing an attribute was related to a respondent's self-explicated importance score. They demonstrate that the customized version reduced the NOL effect and improved ACA version 3 results, though the authors noted that the NOL effect "in ACA 4.0 is much smaller than it is in ACA 3.0 which we used here."

Using LISREL and PLS to Measure Customer Satisfaction (Lynd D. Bacon): CSM (covariance structure models; most commonly modeled with software called LISREL) and PLS (Partial Least Squares) are two SEM methods that can be used to understand the drivers of customer satisfaction and to quantify their importance.

Lynd articulated the pros and cons of each technique, and concluded that CSM and PLS can provide distinct advantages as methods for analyzing satisfaction data. They both provide a means of estimating measurement error and reducing its biasing effects on the estimation of other quantities. Each method provides a means of explicitly modeling multicollinearity so that its deleterious effects on estimation can be reduced. They also differ from each other in important ways. The CSM approach typically requires continuous data and assumptions about their distribution.

Ensuring that CSM model parameters are identified can be difficult. PLS doesn't require distributional assumptions and parameter identification is less of a problem, but it produces biased estimates due to the estimation procedure. PLS was developed for applications in which little theory is available, and predictive accuracy is of paramount importance. CSM, on the other hand, was developed for confirmatory modeling, and will provide the most useful results when the data are collected with using it in mind. When uncertain about which method should be used, Lynd suggested applying both procedures to study their discrepancies, and then deciding which is more useful for your application.

Product Mapping with Perceptions and Preferences (Richard M. Johnson): Rich reviewed the history of perceptual mapping as it relates to marketing research. He noted that approaches have been developed to map products or objects based on preferences and on perceptions, but seldom have both elements been combined in product mapping. “Maps based on perceptions are easy to interpret and good at conveying insights, but they are often less good at predicting individual preferences,” Rich explained. “Maps based on preferences are better at accounting for preferences, but their dimensions are sometimes hard to interpret.”

Rich presented a new method that he termed “Composite Product Mapping” that combines both perceptions of brands on attributes and preferences among brands. The perceptual information results from attribute ratings for brands, and the preference information can come from a variety of sources, including pairwise judgments or conjoint part worths.

He demonstrated that the composite methods often result in maps that closely resemble discriminant-based perceptual maps, but that the attribute vectors and product positions are better linked to preferences. Rich also illustrated that composite mapping can result in different (and more useful) maps than discriminant analysis if the variables that drive discrimination are not important to preferences. In this case, variables unimportant with respect to preference are relegated to less important, higher dimensions with his new approach. Additionally, contours representing “density of demand” can be added to the maps indicating areas of relative preference.

Rich concluded, “All in all, there seems to be no downside to using composite mapping methods, and the benefit of possibly improved interpretation and prediction can be great.”

EVALUATING THE REPRESENTATIVENESS OF ONLINE CONVENIENCE SAMPLES

Karlan J. Witt
IntelliQuest, Inc.

INTRODUCTION

The widespread adoption of the Internet has provided a low-cost, rapid-turnaround alternative for conducting market research. Anyone with a web site and basic HTML programming skills can create their own online studies on an as-needed basis. The results of these studies are often reported in marketing presentations, with sample sizes in the thousands, and all the comfort levels typically associated with large, robust market studies.

However, the old adage that quality is more important than quantity certainly proves true regarding Internet surveys. Using a traditional example, in the 1936 presidential election, the *Literary Digest* achieved a sample of over 2 million to predict 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. The purpose of this paper is to explore the use of online surveys, describe the challenges that exist, and evaluate some suggestions for overcoming them.

This paper is organized into five sections, beginning with a brief background and history of online surveys and a set of definitions of terms to be used in the paper. The paper will then describe the methodologies of the two studies used for the basis of comparison, compare the results obtained through direct comparison as well as some statistical models, and finally provide conclusions and some recommendations for using online surveys.

BACKGROUND

Online surveys are growing in popularity. There are more tools to create surveys and collect data, more mechanisms for obtaining research sample, and more types of individuals creating the surveys. When data is distributed in a report or presentation, it takes on a life of its own, as if committing it to paper makes it valid and reliable. Based on IntelliQuest's experience conducting online surveys, we concur that in many cases online surveys did indeed provide a faster, cheaper way to collect data. Our challenge is quantifying when and when not to use them.

DEFINITIONS

"Online survey" is a term that is applied to several types of surveys:

- Traditional recruit (e.g., using RDD calling) with a follow-up invitation to visit a web site to take a survey
- A pop-up survey on your own web site
- A pre-recruited, web-enabled panel

- Banner ads placed in several places on the Internet inviting respondents to participate in a web-based survey

It is the banner ads that we are calling an "Internet convenience sample". These are very common. The use of these samples by both market researchers and other client-side individuals is growing in popularity, so we undertook studies to validate the use of Internet convenience samples.

SURVEY METHODOLOGIES

This paper will compare the results of two studies. As background for interpreting those results, this section describes the survey methodologies for each. The first is IntelliQuest's Worldwide Internet/Online Tracking Service (WWITS™), and the second is a study conducted at one point in time in order to make the comparisons described in this paper. We will call the latter the "online snapshot study".

WWITS

IntelliQuest's WWITS study is a quarterly tracking study that is projectable to the U.S. population aged 16 or older. It is intended as a sizing and profiling tool, so the methodology is extremely rigorous. The topic of the survey is disguised so that people who don't use, aren't familiar with, or have little interest in the Internet do not disproportionately terminate as refusals. The members of the household are rostered and randomized to avoid any bias based on the person most likely to answer the phone. Call-back attempts are spread out over time, and up to 20 attempts are made to reach the target respondent. These efforts result in a response rate of greater than 80%. For this paper, we are using the Q1 1998 data. We obtained 722 completed interviews for use in the analysis. The data shown for WWITS is weighted to the U.S. census.

Online Snapshot Study

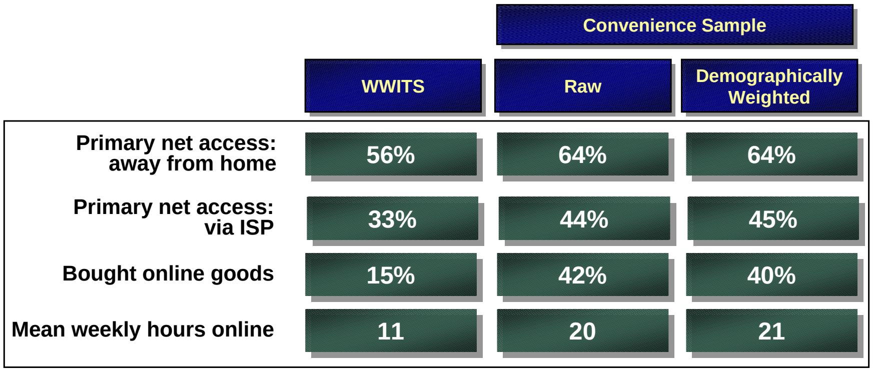
This study was also conducted in Q1 1998. The sample for the survey was recruited using banner ads placed on several sites on the Internet. These included news, travel, and technology web sites. We obtained 1260 respondents for use in the analysis, which reflects a 0.0021% response rate; a rate not atypical for banner ad recruiting.

COMPARISON OF RESULTS

The primary variables collected for comparison included respondent demographics, Internet usage variables and brand shares. In our experience using convenience samples, these are the types of variables where differences occur. We will compare individual variables, as well as several multivariate techniques, to see if the variables and the relationships between variables are similar in the two studies. The specific hypotheses we are testing are as follows:

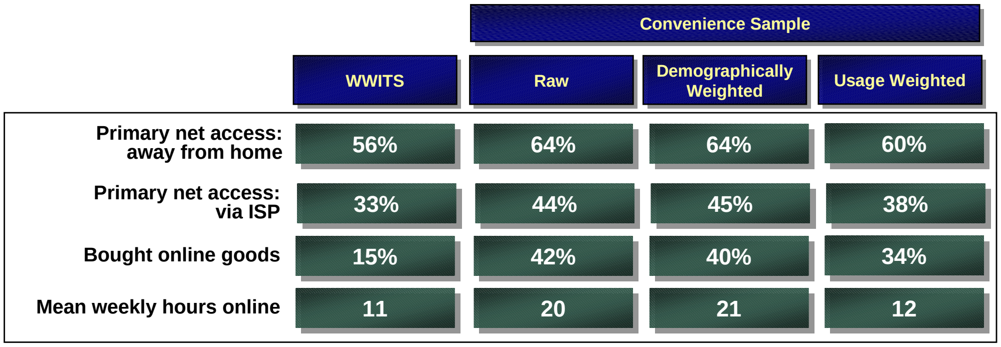
The two sample sources will produce similar results for the demographics, usage variables and brand share data.

The two samples will not produce statistically significantly different statistical models.



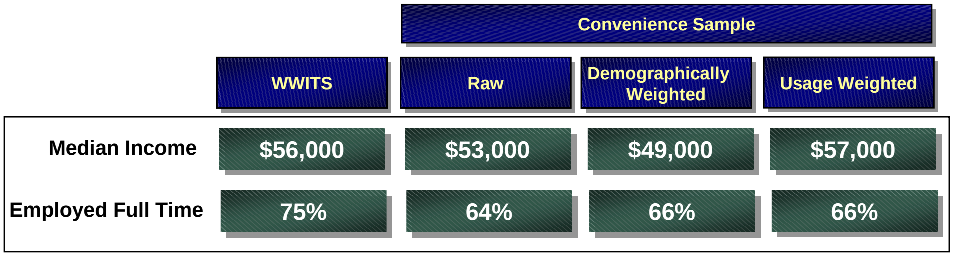
		Convenience Sample	
	WWITS	Raw	Demographically Weighted
Primary net access: away from home	56%	64%	64%
Primary net access: via ISP	33%	44%	45%
Bought online goods	15%	42%	40%
Mean weekly hours online	11	20	21

We see from the data that the convenience sample is more technologically-savvy than the RDD sample. Using the demographic variables of gender, age and education, we then demographically-weighted the data to determine the ability to weight the results to better reflect the RDD results. With regard to the Internet usage variables in the above table, there was little improvement, we therefore constructed a second weighting scheme to account for the usage variables.



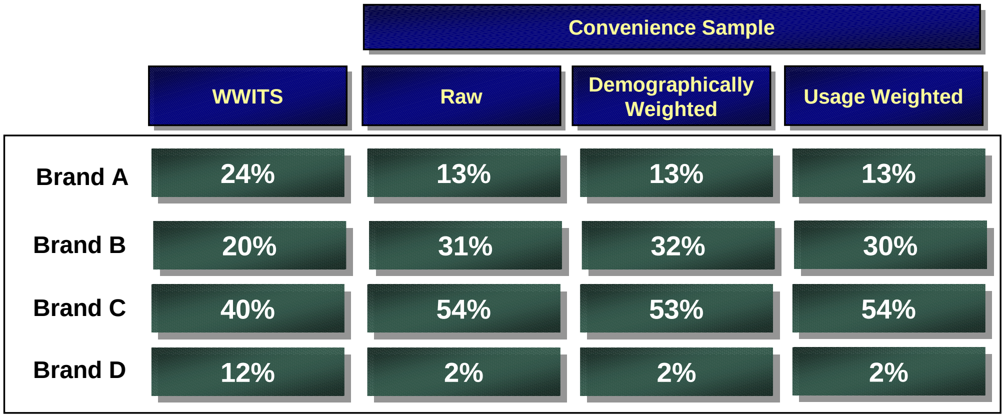
		Convenience Sample		
	WWITS	Raw	Demographically Weighted	Usage Weighted
Primary net access: away from home	56%	64%	64%	60%
Primary net access: via ISP	33%	44%	45%	38%
Bought online goods	15%	42%	40%	34%
Mean weekly hours online	11	20	21	12

The online usage weighting improved the results with regard to several variables, but was still off on others, as shown for employment below.



		Convenience Sample		
	WWITS	Raw	Demographically Weighted	Usage Weighted
Median Income	$56,000	$53,000	$49,000	$57,000
Employed Full Time	75%	64%	66%	66%

IntelliQuest then carried both weighting schemes through our remaining analyses to compare the results for the brand share data.



		Convenience Sample		
	WWITS	Raw	Demographically Weighted	Usage Weighted
Brand A	24%	13%	13%	13%
Brand B	20%	31%	32%	30%
Brand C	40%	54%	53%	54%
Brand D	12%	2%	2%	2%

Brand share data was the next variable we examined. In the data from WWITS and the online snapshot study (both weighted and unweighted), we see that Brand C is a strong leader in the market. After that, however, the rank orders vary. In the WWITS study, Brand A is second, followed by Brand B. In the online snapshot data, the order is reversed. Brands A and B had differing penetration rates within the techno-savvy and novice user categories, suggesting that when the composition of our sample shifted, so did the resulting brand shares obtained. However, the criteria we often hear from clients is "Would I make a different business decision using this data?" To answer that question, we needed to take a different approach.

The next step was to develop several statistical models to see if the comparative relationships between the variables were consistent, even if the variables themselves were not. The techniques used included:

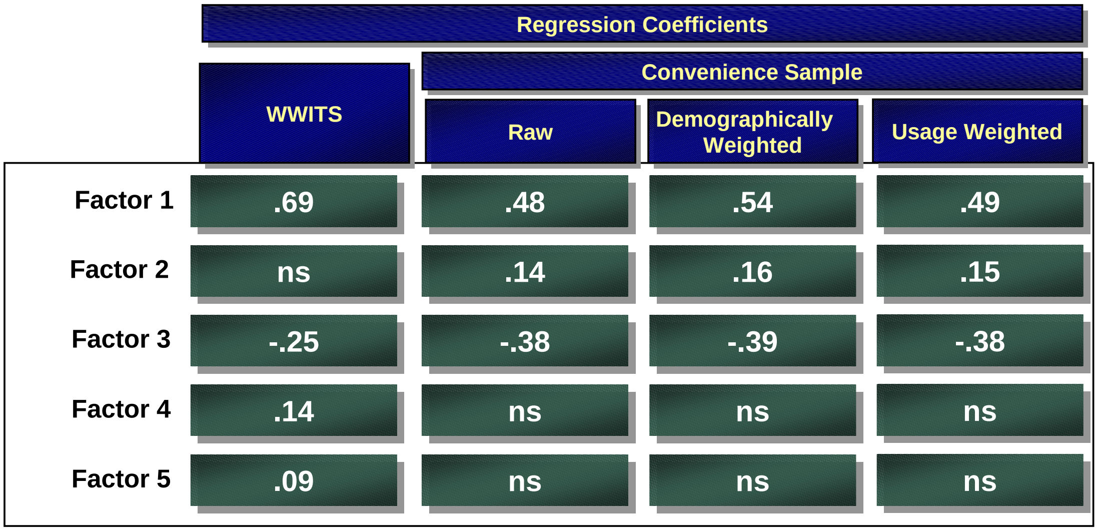
- Factor analysis
- Multiple regression
- Segmentation

Factor Analysis

A factor analysis of 13 attributes representing respondents' attitudes toward ecommerce was performed. A five-factor solution appears common to both samples. We used a Procrustean factor analysis that allows us to analyze both data sets at once, and provides a goodness of fit metric, where 1.0 is perfectly good and –1.0 is perfectly bad. On these data sets, our index of fit was 0.897, which is really good. It says that the structure or the interrelationship between the 13 attributes is well represented in both data sets.

Multiple Regression

For this analysis, we took the five factors that we identified above and used them to predict respondents' propensity to purchase online.



	Regression Coefficients			
		Convenience Sample		
	WWITS	Raw	Demographically Weighted	Usage Weighted
Factor 1	.69	.48	.54	.49
Factor 2	ns	.14	.16	.15
Factor 3	-.25	-.38	-.39	-.38
Factor 4	.14	ns	ns	ns
Factor 5	.09	ns	ns	ns

ns = coefficient is not significantly different from 0

This table shows that the top two factors are the same, but their magnitudes differ. Factors 4 and 5, significant to the WWITS data set, are not significant in the online snapshot study. Similarly, factor 2 is significant in the online snapshot data, but is not in the WWITS data set. Overall this analysis yielded a mixed bag of results, despite numerous weighting attempts.

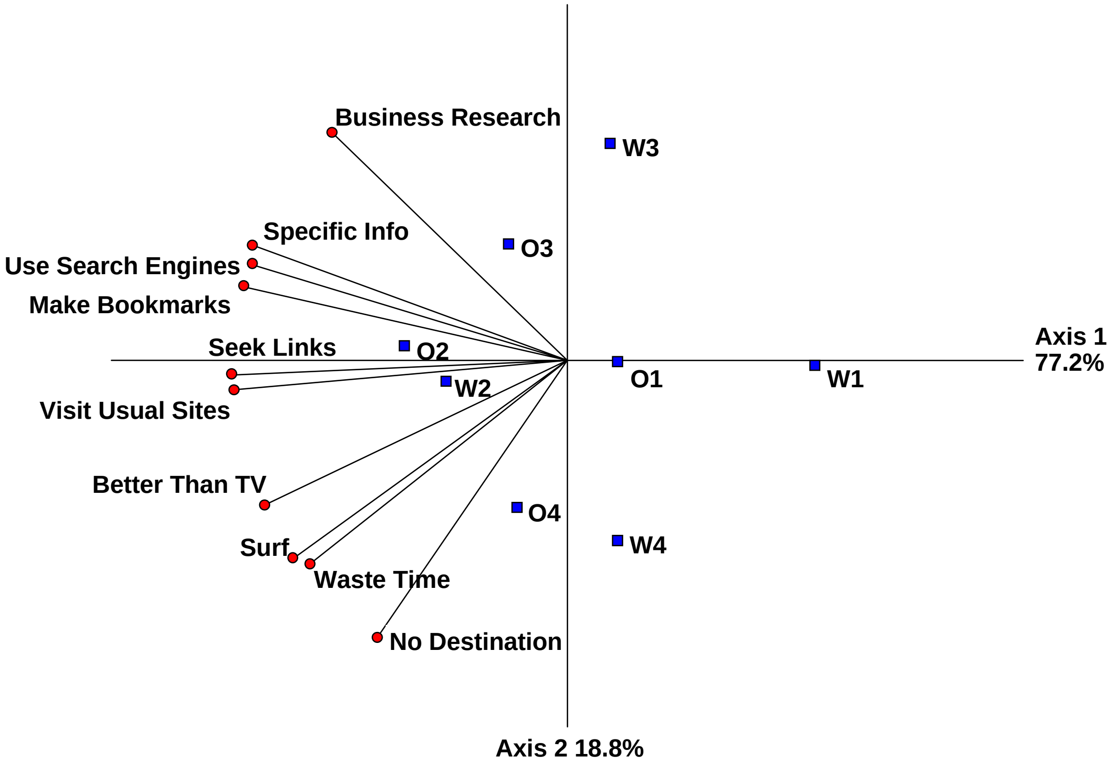
Segmentation

The final analysis was a segmentation analysis. For several years, we have performed a k-means cluster analysis on the WWITS data and have found the same segments every time. The sizes of the segments have grown or contracted over time, as the complexion of the online audience has changed, but overall the WWITS segments are very robust.

To replicate the segmentation using the online snapshot data, we used two approaches. First, we allowed the segments to find their own centroids, simply specifying that there should be four segments. The second approach was to provide the coordinates for the WWITS centroids, and compare the segment sizes and composition.

	Segment Sizes		
		Convenience Sample	
	WWITS	Free Centroid	Fixed Centroid
Segment 1	20%	34%	2%
Segment 2	25%	17%	61%
Segment 3	31%	19%	19%
Segment 4	24%	30%	18%

What we found was that the fixed centroid method yields a solution with very different segment sizes. The free centroid method produced segments where the sizes were in line, then we needed to check the identity.



On the above map, the lines represent the 13 ecommerce attributes measured. The four WWITS segments are identified as w1, w2, w3, and w4. The online snapshot segments are identified as o1, o2, o3, and o4. The interesting thing about the online snapshot segments is that they appear to be just left-shifted from the WWITS segments. The relationships between the segments are very similar. If we remember that the online snapshot sample is biased in terms of technology users, it makes sense.

CONCLUSIONS

Our Internet convenience sample varies in terms of respondent demographics, online usage behavior and brand share, regardless of weighting schemes. The statistical models were more encouraging in their descriptions of the two samples, but disturbingly, our predictive model, the regression analysis, was not robust.

Weighting was not the solution. In many cases it did make the data look more like the RDD data, but it did not fully address the inherent non-response bias present in the online sample.

However, through this and other experiences, IntelliQuest believes online samples have their use in a market researcher’s arsenal. We offer the following advice for using such convenience samples:

- Comparative studies, especially those using test and control stimuli
- Tracking studies where wave-to-wave changes are the focus
- Psychometric studies which focus upon the structure of respondents’ beliefs, values or attitudes
- Segment identification and profiling studies

Based on the findings for this study, we recommend further research be conducted to compare Internet convenience samples on additional metrics and to compare other types of popular online methodologies to more traditional approaches.

CONJOINT ON THE WEB—LESSONS LEARNED

Michael Foytik
DSS Research

In the past year, we have conducted several conjoint studies using the Internet and World Wide Web (Web). We have executed choice-based conjoint, pairwise comparisons (graded pairs) and full profile conjoint surveys via the Web. We will try to impart what we have learned over the last two years in this paper.

BACKGROUND AND EXPERIENCE

DSS Research started using tradeoff matrices and then full-profile card sorts in the late-1970’s as we learned to understand and put conjoint to use for market research. In 1989, we happily adopted Sawtooth’s Adaptive Conjoint Analysis (ACA) program to implement many of our conjoint surveys. Over the years, we have conducted many disk-by-mail conjoint surveys with targeted groups using ACA, Conjoint Value Analysis (CVA) and Choice Based Conjoint (CBC) programs from Sawtooth. Disk-by-mail surveys can be considered the precursor to Internet surveys in many ways, as we will discuss later.

Like many companies, we started seeing the value of the Internet and the Web in 1995. DSS purchased a domain (dssresearch.com) and server space in late-1995. We completed our first Internet survey a couple of months after we went online, but we did not have an opportunity to conduct our first Internet conjoint survey until January 1997. By the beginning of 1997, the Mosaic browser had gone through several generations of improvement in the hands of Netscape and others. At this time, several factors came together to fortify the Internet’s position as a new medium for collecting information. In January 1997, there were approximately 25 million Internet users in the United States. Both Microsoft and Netscape were offering good Internet browsers free of charge. Internet server software was available from numerous vendors for UNIX and PC-based computers. And, several hypertext markup language (HTML) tools like HotMetal and Visual InterDev (now called Visual Studio) became available to make producing Internet content easier.

INTERNET CONJOINT PROJECTS

Since January 1997, we have conducted conjoint analysis projects in several different industries, using pairwise comparisons and choice-based conjoint, and involving different constituencies. Some examples of our Internet projects include:

- Pairwise comparisons with 15 attributes comprising health insurance plans (consumers).
- Choice-based exercise using 15 healthcare attributes (employers).
- Choice-based exercise using 15 healthcare attributes (repeat of above exercise with 3 different consumer segments).

- Two stage conjoint survey using pairwise comparisons and choice-based conjoint to evaluate new designs for commercial helicopters (businesses/wealthy individuals).
- Choice-based conjoint for personal communications services (consumers).
- Choice-based exercise to evaluate survey incentive options for online surveys (Internet users).

One of the biggest roadblocks we have faced in convincing clients to use the Internet for conjoint analysis has nothing to do with the Internet itself, but a lack of familiarity with conjoint in general. Even after 25 years, many corporate market researchers in healthcare and high technology companies have had little or no exposure to conjoint. In many cases we have been forced to focus on gaining client acceptance of conjoint analysis itself, ignoring potential Internet applications of this technique except where it is clearly the best choice.

We are fortunate to have several clients where Internet conjoint surveys have made good business sense. In a couple of studies, we simultaneously collected data using Internet and traditional methodologies because our clients' customers were reachable from both approaches. We have also funded a couple of Internet conjoint projects out of our own pocket to learn more about Internet users and to develop knowledge bases. These studies have given us the opportunity to compare and contrast Internet conjoint research under different situations.

COMPARISONS OF INTERNET AND MAIL METHODOLOGIES

We will ignore telephone research, because only the most simplistic of conjoint studies can be successfully implemented over the telephone. Intuitively, mail surveys have many of the same characteristics as Internet surveys: no interviewer intervention, self-paced, similar perceptions of anonymity, and respondents record their own answers.

- ***Email versus mail.*** Email very closely resembles typical mail surveys, including most of the negative characteristics. There is no interaction with respondents, skip patterns must be handled with written instructions, responses can not be validated as they are entered and it is difficult to even verify receipt of the survey. We have conducted email surveys, but have never tried conducting conjoint surveys via email. We don't recommend email for conjoint surveys, unless it is a trivial choice-based or full profile exercise. At least with printed mail surveys, you can control the layout of the document to attractively present a large number of attributes in a limited amount of space. There are several factors which limit what you can do with email surveys:
    - ♦ *Every email client displays messages differently.*
    - ♦ *Many email clients can't handle embedded HTML formatting of the document.*
    - ♦ *It's more difficult to mark responses to email surveys because they must be typed directly on the message.*
    - ♦ *Email users are accustomed to quickly scanning their messages, increasing the likelihood that hurried users will delete long conjoint messages.*

- ***Web versus mail.*** Web-based surveys resemble interactive versions of the typical mail survey. The strengths of web conjoint surveys are:
  - ♦ *Almost as much control over look and presentation of questions as you get with layout of mail surveys, with less effort.*
  - ♦ *Ability to dynamically change conjoint tasks in reaction to respondents' inputs (e.g. pairwise comparisons).*
  - ♦ *Ability to validate respondents' inputs as they are entered.*
  - ♦ *Ability to present one task at a time (e.g. only one choice set or one graded pair is displayed at a time) to keep respondents from looking ahead or getting discouraged about survey length.*
  - ♦ *You can advertise or invite participation in an Internet survey using many different methods (e.g. email, banner ads, postcards, newspapers, magazines, etc.) without sending out an expensive survey.*
- ***Comparing results.*** We will focus on one particular methodology comparison to illustrate what we consistently find when comparing Internet and mail conjoint results.
  - ♦ *Respondents.* In a consumer healthcare study, we screened potential respondents via telephone and then invited them to participate in a conjoint survey. Respondents were allowed to select whether they completed the survey via mail or the Web. Households were randomly selected using random digit dialing. The only qualifications for participation in the study were that the person be a decision-maker for health insurance and that the person has commercial health insurance or Medicare. Among the households who qualified for the survey, 41.7% had access to the Internet either at home or at work and 69.2% of those people chose to do the survey via the Internet.
  - ♦ *Exercise.* Choice-based conjoint was used to collect the necessary data. The project consisted of 15 attributes and four choice options per set. Respondents were required to make a choice among the four options in each set. Using a randomized block design, each person evaluated 10 choice sets among the 80 choice sets included in the design. Each person also evaluated one holdout choice. Each choice set was carefully formatted to fit on an 8-1/2" x 11" page. The Internet choice sets were formatted using standard HTML tables. We could not control text wrapping or make precise text placement on the Internet survey. The large number of attributes meant that every respondent had to scroll up and down to see the complete descriptions of all four options. Despite our reluctance to present these 60 cell tables (15 attributes x 4 options) to Internet participants, we found no ill effects from the complexity of the task presented or the inability of users to view the entire exercise on one screen. Results compared favorably to the cleanly formatted mail surveys. Response rates and the predictive accuracy of the conjoint model from the Internet data were equivalent to a similar Internet conjoint survey we conducted that used only five attributes.

- ♦ *Incentives.* Before asking for participation in the conjoint survey, we explained that each participant would receive a $5 cash payment for completing the survey. They were also told their name would be entered in a drawing for one of four $250 gift certificates to be given away at the conclusion of the study.
- ♦ *Results.* It has been our experience on Internet surveys (conjoint and non-conjoint exercises) that participants give greater consideration to the questions and the answers they provide. Cronbach's Alpha and the Guttman Split-Half tests show the Internet responses to have greater internal consistency. Predictive modeling using satisfaction scores or stated intentions as the dependent variables has shown that the Internet samples produce more accurate predictions than their mail counterparts. In our limited comparisons, the Internet samples have produced conjoint models that provide a better fit of the conjoint data and these models more accurately predict holdout choices. Even if the consistency and predictive accuracy of these Internet conjoint models is due more to the novelty of the data collection methodology, it still illustrates the ability of researchers to accurately collect complex information via the Internet.

RESPONSE RATES

We have no reason to believe that Internet response rates will exceed those of mail surveys, over the long-term. But, in the short-term, we find response rates 25% to 50% higher than those of comparable mail surveys. For example, on the consumer conjoint study described above, 81.5% of those who selected the Internet survey completed the survey within the timeframe given. Among those who opted for the mail survey, 59.8% returned the survey within the allotted time. Both groups received two notifications concerning the conjoint survey. Internet participants received two email messages about 10 days apart. Mail participants received two surveys in the mail, approximately 14 days apart. Although the mail survey achieved a very high response (responses typically range from 45% to 60% on similar mail projects), the Internet survey did 36% better.

This differential between response rates of mail and Internet surveys is likely to narrow or even disappear at some point. We experienced very high response rates with the first disk-by-mail projects we conducted. Those response rates have dropped considerably in the last 5 years, but they are still somewhat higher than for regular mail surveys. We might expect a similar effect for Internet response rates, but in the short-term there are three good reasons for the high response rates on Internet surveys:

- ***Novelty.*** Internet surveys are still a novelty. Many people are likely to try an Internet survey just to see how it works or to see what can be done. The novelty factor will be the first to disappear as more and more surveys are conducted over the Internet.
- ***Ease of use.*** When taking advantage of the "computer-assisted" aspects of Internet applications, these surveys can be easy to use. In some cases, Internet surveys may be easier than mail surveys. There are no complex skip patterns to try and describe to respondents. Just let the survey application make the appropriate decisions. Feedback or help can be provided directly to users by linking to explanatory diagrams or text throughout the Internet survey. In some cases, tailoring the survey to each individual's interests can make the

make the Internet survey easier (or at least, more relevant). The set of brand names used in the conjoint exercise can be dynamically changed based on the consideration set of each individual. Similar modifications can be made to any Internet survey to address the participant's interests or concerns.

- ***Immediacy.*** Internet surveys have a big advantage over mail surveys in that they can be accessed immediately after alerting the potential respondent of their existence. Over the phone, you can read a simple Internet address (URL) to a qualified respondent and they can go straight to the web site. Email notification messages can be sent within minutes of obtaining a valid email address for someone. The receiver of an email message simply clicks on the hyperlink to jump straight to the Internet survey. You can get the survey in someone's hands while their commitment to complete the survey is still fresh in their mind. When compensating participants or registering them for prize drawings, Internet surveys also have the advantage of providing immediate feedback that their entry was received. Even with overnight delivery, mail surveys require several days to go from notification to returned response.

NOTIFYING POTENTIAL INTERNET RESPONDENTS

There are several ways of reaching out to Internet users to get them to participate in an Internet survey. This has evolved from a single choice a few years ago (soliciting responses from your own web site) to a number of internal and external (commercially) available solutions. We anticipate that pseudo-random Internet samples, equivalent to listed telephone samples, and highly targeted email lists will be available very soon.

Currently, the most common options for contacting Internet users are:

- ***Customer lists.*** Many companies are beginning to collect email addresses as part of their normal customer registration. You are likely to see a place for email addresses on home appliance warranty cards, church attendance sheets, or magazine subscriptions. Sales reps are now asking for an email address when people call in with questions or requests for information. Web sites are asking visitors to provide email address or sign up for online newsletters by giving their email address.

- ***Purchased email lists.*** Several companies are selling email addresses for one-time or multiple mailings. Although lacking the range of choices available for mail lists, you can now request email addresses in specific parts of the country, or by basic demographic characteristics like age and gender. You can also purchase email lists from or have your message included in a regularly scheduled newsletter from online web sites and online publications that target your audience. One word of warning about buying email addresses, make sure you are dealing with a reputable firm that only supplies "*opt-in*" email addresses. "*Opt-in*" means that those on the list explicitly asked to be included or they granted permission for the owner to make their email address available to other companies (usually restricted to companies offering something of interest). Any other kind of email list will be considered SPAM. Being a willing party to SPAM attacks can get you banned from most ISPs, results in bad publicity for your company, can lead to your web site being hacked into, or see your email server brought to its knees with a denial of service or other type of attack. SPAM a large group of people and you will quickly become public enemy number one. Currently, quality opt-in lists are very expensive with many charging at least $250 per thousand names for a one-time mailing. Like everything else, these prices should come down as competition heats up.

- ***Online panels.*** Just like their mail counterparts, online panels are now widely used to solicit opinions from people with known characteristics and known interests. The major mail panel companies are also offering online panels and several new companies have jumped in to fill the need for online panels. Because it is relatively inexpensive to develop an online panel, several companies have already developed lists that rival those of the largest mail panels. You may want to consider developing a specialized panel for your own research needs. We obtained demographic characteristics and email addresses for over 6,000 Internet users in 14 days. We did this by holding an online contest and asking people who were interested in future online surveys to check a box on our form and then give us their email address. Approximately 70% of those who saw our contest and filled out the demographic questions also joined our online panel. Those who agreed to join the online panel were found to be slightly more knowledgeable about Internet technologies, more comfortable using the Internet, more likely to buy over the Internet and they more closely resemble traditional upscale shoppers (female, 30-40 years old, and married).

- ***Online promotions.*** You can start by placing survey notifications on your own web site or the web site of your client. You can also reach your target audience through online promotions like banner ads on selected web sites, newsgroup postings, search engine placement, site sponsorship, etc. We caution the use of these methods because they may miss your target audience completely and hurried web surfers often overlook them. If you conduct online advertising, we recommend finding a site(s) which most closely target the topic you wish to survey. Commonly quoted figures for banner click through rates (the percentage of people who actually click on the banner when it appears on the page) hover around 1% to 2% for a general interest site. Click through rates typically do not exceed 5% even for well designed banner ads on highly targeted sites. From our own experience, the last banner ad we ran for an Internet survey cost us $4,000 for 170,000 impressions on Excite's group of search engines and related sites. The banner ad was an animated gif image that prominently mentioned $500 in prizes to be given away for answering a short survey. Our click through rate was right at 1% and only 68% of those who viewed the first page of our survey completed the entire survey.

- ***Traditional contacts and promotions.*** If your client only has mailing addresses or telephone numbers for their customers, you can use a traditional mail or telephone notification. If the customer base has a high probability of being connected to the Internet, send postcards inviting them to participate in a survey or include a notice in the next newsletter or outbound correspondence. When pre-screening of participants is necessary, short telephone calls offer the best opportunity to collect the necessary screening information and then provide a personal invitation to those who qualify. Printed publications (newspapers, magazines, etc.) are a possible source, but we would not expect a high response rate from any of these publications.

HANDLING EMAIL INVITATIONS

If you conduct online surveys, you are likely to need email to notify potential respondents where to go to find your survey or to invite people to participate in a survey they have yet to hear of. Although emails are simple to execute, we offer a few tips that might help:

- ***Use a persuasive subject.*** The subject text in an email message is like the title of a new book or the tag line on a new ad campaign. It has to grab the readers' attention in a moment and encourage them to read further. Even if these people are your customers or they have already agreed to participate, make sure your email subject text is persuasive. If addressees have already agreed to complete the survey, remind them of this fact. If a prize drawing or cash incentive is involved, mention this in the subject heading. If this is a blind invitation, give people a reason to at least read your email. Mention how the info will be used, describe direct benefits to the reader, or discuss a topic of interest to the receiver. You must walk a fine line between sensationalism and positive promotion. Many email users scan incoming email messages or use the filtering capabilities of their email program to eliminate junk emails. Filters often look for words and phrases like "$$$", "guaranteed", "can't miss", "money", "!!!", "get rich quick", etc. so avoid any phrases which could trigger a filter that automatically deletes these messages without them ever being seen. An example of a simple subject heading that worked well for us is:

  The healthcare survey we called you about

  A reminder that the deadline is nearing might look like:

  Thanks for agreeing to complete our survey – respond by 10/23/98 to enter drawing for $500.

- ***Make your point quickly.*** OK, so our persuasive subject heading encouraged everyone to open our email messages. We need to get to the point quickly to maintain their attention. Apply the same principles of good advertising copy and use anything that has worked well for you in the past with mail survey cover letters (but, keep it shorter than a typical cover letter). Briefly tell the reader what the survey is about and particularly what each reader can expect to get out of it (incentives, prize drawing, etc.). If appropriate, remind readers of their earlier commitments to complete the survey. Then provide a call to action like "*click on the link below to begin the survey*."

- ***Keep it short.*** Not only must the message be short, but the subject header must also be short. Besides the obvious need to maintain readers' attention, there is another reason for keeping messages short. Some email editors have message limits that prevent them from receiving large text messages. This primarily affects long email surveys, but it can have an effect on an extremely wordy survey invitation. More importantly, the subject headers should be no longer than 100 or 150 characters, because many email editors will truncate subjects beyond those limits or the subject matter will not display completely in the amount of space allowed for displaying the message subjects.

- ***Use clickable URLs.*** Anytime you invite someone to complete an Internet survey, you provide an address or URL to access the survey. In email messages, make sure the URL is a clickable link by prefacing the address with the *http://* protocol designation. Although the vast majority of email clients now support embedded hyperlinks, it's still worthwhile to provide a brief explanation on how to cut and paste this URL into a browser.

- ***Embed any IDs or passwords into clickable URL.*** If you wish to uniquely identify each invited participant or restrict access through passwords and user authentication, embed this information directly into the URL that you provide. Users notoriously have difficulty entering the unnatural sequences of letters and numbers that characterize most computer generated passwords. Long strings of numbers are also mistyped as ID numbers. Make the login process as painless as possible by embedding this information in the survey hyperlink. This will help you avoid many calls and emails from upset participants complaining that the password or ID you provided is invalid. Most email merge programs will easily allow you to pull this information out of a database and insert it in the appropriate place in the document. An example or an email URL with embedded ID and password is:

  http://www.dssresearch.com/startsurvey.asp?ID=123456789&password=secret

- ***Use an email merge program.*** Be prepared to send your custom emails in large volume. There are a number of free or low cost programs which can handle merging unique information into each email message the same way word processing programs create mail merges for printed documents. Special characters are used as placeholders for information like name, email address, ID and password to be inserted. Some email programs like Pegasus have built-in email merge capabilities that are adequate for small volume and simple email lists. We use a program called WorldMerge which reads standard dBase files and allows you to perform basic filtering to select specific records. However, this program is not particularly user friendly in handling the creation and editing of text for the message body. Find something that you are comfortable using, then be prepared to use it often.

- ***Send frequent messages.*** Take advantage of the Internet medium. Email messages are the cheapest way of contacting people throughout the world, so send your messages often. If you pre-qualify individuals for an Internet survey, we recommend sending email invitations as soon as possible. Preferably, send the email messages at the end of each interviewing shift or the next morning. We usually follow-up the first round of email messages with a reminder message four to six days later. A third message should be sent to non-responders within 14 days of the first message. Depending upon the target population and the subject matter, a fourth or fifth message might be sent. Just be careful not to abuse your subjects.

- ***Carefully record email addresses.*** Be prepared to instruct interviewers on the proper way to record email addresses. Unlike the recording of verbatim comments, one mistyped character can render the email address worthless. Make sure interviewers record the three components of any email address: the unique name to the left of the "@" sign, the second level domain name, and the first level domain name preceded by the "." as in ".com". Familiarize each interviewer with the separator characters ("@", ".") that appear in every email address. Instruct them that spaces are never allowed in email addresses. Warn interviewers that some email addresses contain third or fourth level domain names (e.g. mail.bna.bellsouth.net). For now, interviewers can be taught that all email addresses end with ".com", ".net", ".org", ".edu", ".gov", or ".mil", but this may soon change. Finally, make it company policy to carefully read back each email address character by character to insure accuracy. If you don't provide proper instruction to those collecting email addresses, you are likely to end up with something like "John Doe at all dot com", instead of "johndoe@aol.com."

SURVEY INCENTIVES

Incentives for mail and telephone surveys have long been studied and debated. We have no reason to believe that Internet participants will differ significantly from mail and telephone participants with regard to incentives. However, we do believe there are potential differences due to the way users typically scan Internet sites rather than carefully reading them. We took advantage of the ease with which Internet conjoint studies can be implemented to conduct a simple study on Internet survey incentives. We created a simple, choice-based conjoint exercise with three attributes. Participants were asked to imagine they received an offer to complete a survey over the Internet and must now decide whether they wish to participate or not. The survey was described by these three attributes:

- ***Length of survey.*** The estimated length of time to complete the Internet survey was shown. The options were 5 minutes, 15 minutes and 30 minutes long.
- ***Cash incentive.*** Respondents were told that surveys could offer cash incentives of $10, $5 or not pay cash to survey participants.
- ***Prize drawing.*** Survey participants would be eligible for prizes ranging from two $500 gift certificates to one $100 gift certificate or no prize giveaway at all. There were a total of nine possible prize values, including no prize. We quoted odds of winning each prize at 1 in 10,000 for each prize given (e.g. 4 prizes would have an odds of 4 / 10,000 or 1/2,500).

We used a randomized block design with six blocks of five choice sets each. Each choice set contained three potential Internet surveys (composed of the three attributes) and an option to choose "none of these." Each survey participant was randomly assigned to one of the six block groups based on the random ID assigned as they entered the web site.

We recruited Internet users to take the conjoint survey. Each person examined the choice sets and responded to each by indicating the survey they would most prefer to participate in, or they indicated that none of the surveys shown in that choice set were acceptable. We received over 7,000 responses during the four week period this conjoint survey was available on the Internet.

The incentives were found to be very attractive. Only 3.2% of all responses were for the "*none of these*" choice. Measuring the overall importance of each attribute by the taking the difference between the highest and lowest valued levels of each attribute, shows *cash payments* edging out *prize drawings* as the most valuable feature. Respondents placed six times more value on the $10 cash gratuity than they did on a $5 cash gratuity. By comparison, doubling the number of gift certificates to be given away in a prize drawing only increases its utility by a

factor of
2 to 2.5.

The takeaways we get from this simplistic and unscientific survey are not totally surprising, but they are probably worth repeating:

- ***Respondents prefer the sure thing.*** Respondents prefer a guaranteed $10 cash payment to a prize drawing offering the chance to win one of two $500 gift certificates, even when the odds of winning are 1 in 2,500. However, a $5 cash payment is equivalent to a prize drawing with a single $250 gift certificate or a prize drawing with three $100 gift certificates. This is an important statement because the cost of paying $5 to 4,000 or even 400 Internet participants far exceeds the cost of setting up a drawing to give away $250 to $500 in prizes.

- ***Keep surveys under 15 minutes.*** The utility for a 5 minute Internet survey is 6 times greater than it is for a 15 minute survey. From there the curve gets even steeper. The utility for a 15 minute survey is more than 10 times greater than the utility for a 30 minute survey.

- ***Offer some form of incentive to any Internet survey of more than a few questions.*** Unless you have a captive audience who have a high degree of interest or attachment to the subject matter you wish to discuss, we recommend the use of some type of incentive on all but the most trivial Internet surveys. Even a 5 minute survey looks unattractive to busy Internet surfers unless there is the equivalent of at least $200 in prizes available to those who participate.

- ***Find non-cash incentives.*** Be creative. Incentives do not always have to involve cash or monetary prizes. You might allow survey participants to access private information not normally made available to the public that is owned by you or your client. You may give away company products if the study already identifies you as the sponsor. On a general consumer survey, you might solicit sponsorship from organizations that would like to access the information you wish to collect. Sponsors could provide money, products or links from a popular web site to your online survey. We experienced success offering access to a health risk assessment survey that gave users detailed information on their health status just for volunteering to complete a survey on healthcare issues.

- ***Avoid offering cash payments.*** For general-purpose surveys, we do not recommend offering cash payments directly to survey participants despite their obvious value as an incentive. Currently, there are no standards or prevalent systems that allow you to conduct electronic transactions with Internet users. For instance, you can't pass someone the equivalent of $5 in an email or through a server cookie unless both parties subscribe to a proprietary electronic commerce system (like eCash, Microsoft Wallet, etc.). Therefore, the expense of sending the gratuity can greatly increase the total cost of the incentive program. For online panels, difficult to reach respondents, or targeted email and customer lists, cash incentives may still be a good choice. However, plan on using regular mail to deliver the gratuities to each person.

- ***Turn cost savings into better incentives.*** On several of our Internet conjoint surveys, we have used cost savings from conducting the research over the Internet to fund higher incentives for respondents. If you were planning to offer a $5 incentive to mail survey respondents who returned their survey, you might consider offering $10 to those who will do it over the Internet. When you add up the costs of mailing two 10 to 20 page surveys and possibly using a postcard reminder between mailing #1 and mailing #2, you can easily spend more than $5 in materials, postage and handling. Getting at least some of those people to use an Internet survey is to your benefit.

THINGS TO DO TO HAVE A SUCCESSFUL INTERNET CONJOINT SURVEY

Here are a few things to keep in mind on your next Internet Conjoint survey:

- ***Make it interactive.*** Take advantage of the Internet's capabilities whenever feasible. One way to do that is to make the conjoint survey as interactive as possible. You can do this by using a few questions at the beginning of the survey to tailor the consideration set for price or brand name in a choice-based conjoint exercise. You can use respondent specific pairwise comparisons (like ACA) or dynamically create choice-based conjoint designs the way Sawtooth's CBC program does. The only negative to using these types of conjoint surveys is that, for the time being, you must have the programming capabilities to implement the interactivity for yourself. But this too should change as conjoint becomes more prevalent on the Internet.

- ***Database driven exercise.*** If you develop a fixed design for a choice-based exercise, we recommend placing this design into a database. The database can then be queried to recall all the information needed for each choice set. We create our databases with fields for block number, set number, column number and then a field for each attribute in the conjoint exercise. If four options will be shown in each choice set, a query for a particular choice set will return four records (one for each option to display). It is a simple programming step to loop through those records and format each option into a column of an HTML table. We even pull the attribute descriptions straight from the field names in the database and display them to the left of the choices.

- ***Embed respondent specific information in ID.*** When conducting a conjoint exercise with a randomized block design, we generally use a modulus operator on the ID number to determine which block group each respondent belongs in. Other information can be keyed from the respondent's ID or can be looked up in a user reference table based on the ID given.

- ***Assign unique passwords.*** When working with pre-screened individuals or customer/purchased lists, we always assign a sequential ID and a unique password to each respondent. This gives us the ability to keep outsiders from stumbling into our survey and it prevents respondents from easily guessing the ID number/password combinations of other respondents. Passwords should be at least 6 characters long. When using randomly generated passwords, only use numbers and upper or lower case letters. Punctuation marks, dashes and other special characters are difficult for potential respondents to reenter. Do not mix upper and lower case letters because potential respondents are not that careful when hand-entering these passwords. You can write a simple program, Excel macro or database procedure to randomly generate these passwords to any length desired. As a final check, run a query to make sure that your passwords are unique.

- ***Use clickable links to your survey.*** We discussed this earlier, but it is worth mentioning again. When sending email notifications or invitations, use a clickable link to your survey in each email that carries the unique ID and password for each individual.

- ***Provide a contact point.*** Make sure every Internet survey has a phone number and email address of a contact person *on every page*. Problems are most likely to occur during login or somewhere in the middle of the survey. If you only provide contact information at the beginning of the survey, you will leave stranded those people who have problems halfway through the survey. Many people still don't grasp the concept that the Internet is nothing but a network of computers and communications devices. Any momentary lapse in the chain of connections can lead to delays or temporary error messages. In most cases, a respondent's problem can be corrected by asking the person to resubmit the last page they were on, but some people never think to try this on their own.

- ***Allow users to restart surveys where they left off.*** This point is related to the one above. Conjoint surveys can be long and have extended periods of inactivity between each response. If an error does occur or the person has to discontinue the survey for any reason, we advise you to allow the person to restart the survey in the future at the point where they left off. This requires extra programming to make it function properly. One issue is tracking the last question submitted so you can immediately return to that location. The second issue is programming the survey to look in the database for previously saved
information whenever a restart flag is set or the necessary information is not submitted as a query string.

- ***Program for lowest common denominator.*** Unless you are surveying within a particular company where hardware and software requirements are fixed, you must consider what the typical Internet user has in his or her possession. Programming for the lowest common denominator insures that at least 90% of all Internet users will be able to access and complete your conjoint survey. We currently define the lowest common denominator as a low-end Pentium class machine with a 640 x 480 resolution monitor, 256 color display, and running Netscape 3.0 or Microsoft Explorer 3.0. Depending on whose numbers you believe, 10% to 15% of all Internet users still use a version 3.0 browser. Even with the proliferation of 17 inch monitors, over 30% of the people who responded to our most recent Internet survey said their monitor is set at 640 x 480 resolution. Also keep in mind that WebTV users have even less screen resolution than the typical computer user and there are now handheld devices for surfing the Internet. AOL's ancient browser (version 3.0) does not even handle frames and it has numerous other limitations. Running Netscape's or Microsoft's browser within AOL limits their functionality, as well. We already take advantage of some 4.0 browser features like dynamic mouseovers when designing our surveys, but we make sure all survey pages are compatible with the 3.0 browsers to avoid locking out potential users. We anticipate relying primarily on 4.0 browser features by second quarter of 1999.

- ***Test your surveys.*** We test every survey under Netscape 3.0, Netscape 4.0, Explorer 3.0, Explorer 4.0 and AOL's 3.0 browser. This will help you cover at least 90% of all Internet users. At a minimum, test your surveys under the latest versions of Netscape and Explorer.

PITFALLS TO AVOID WITH YOUR INTERNET CONJOINT SURVEYS

There are some specific issues to avoid when creating and managing your Internet conjoint projects:

- ***Don't use latest technologies.*** We concluded the previous section talking about what technologies to consider. We start this section by discussing the technologies to avoid if you are trying to reach a mass audience of Internet users. Although Java and Dynamic HTML (DHTML) have been around for a couple of years, there are still many who can not or will not use them. Some people set their browsers to ignore Java applets and many people still use the 3.0 browsers that do not recognize DHTML code. WebTV users are precluded from using most downloadable technologies. You also have to consider the corporate firewalls that block executable programs like Java and ActiveX. Programs like RealAudio are in widespread use, but many technologies like Macromedia's Flash are in their infancy. We have found that many Internet users have no experience with Adobe Acrobat for reading PDF files, even though this program has been freely available for years. If it doesn't come standard with Netscape or Explorer 3.0, you will find limited acceptance.

- ***Don't rely exclusively on client-side scripting.*** Client-side scripting is very helpful in taking some of the burden of validating user inputs off your Internet server. However, do not make the mistake of relying exclusively on this for all your validation. Double-check any critical information when you process each survey page on your server. To work under both browsers, write your client-side scripts in JavaScript and test extensively under both browsers. VBScript only works with Microsoft's browsers.

- ***Carefully store survey data.*** Unless you want to expose your survey results to outsiders, don't save your survey data in the same directory where the survey forms reside or in a publicly visible directory. It's easy for people to stumble on this information accidentally or otherwise, if it is not safeguarded.

- ***Be careful using automated procedures.*** Most of the latest HTML tools offer wizards or other automated programs to help you create survey forms, auto-respond to emails and handle other common functions. Carefully examine the output of these programs before using them. For example, the default functionality of Microsoft FrontPage's webbot for creating survey forms stores the survey results in a text file and lists the path of the file in the HTML code. If not changed, any survey participant can glance at the source of the HTML page from their browser, cut and paste the URL, then watch their browser download your survey results. In addition, these automation routines are not always bug free, so test these just as you would test hand-written code.

- ***Save user responses frequently.*** Don’t try to carry too much respondent information forward from page to page during a survey. Doing so can actually slow down processing if the information is being sent back to the respondent’s browser with each new page. It also prevents you from allowing respondents to return and restart an incomplete survey. Since you don’t know when the connection might be broken, you don’t know when to save the existing response for later use. Failure to save survey responses as they are received also increases the likelihood of loosing some or all of the survey responses. If a query string gets too long for the server’s buffer (the maximum length depends on the particular server), the data is truncated and fields or portions of a field can be lost. Any problems during communications between the respondent’s browser and your Internet server can also result in losing all information previously entered. We always save responses as each survey page is submitted. Any delays while waiting to access a server database have proven to be minimal.

BUT WHY? PUTTING THE UNDERSTANDING INTO CONJOINT

Ray Poynter
Deux

INTRODUCTION

This paper focuses on understanding why people make the trade-off decisions they do. Over the years, Conjoint Analysis in its many forms and guises has been very powerful in showing us what decisions people will make. Armed with our Conjoint methodology, we can determine whether a particular consumer is more likely to prefer the short, red, round jug or the blue, square, tall jug. We can question a sample of consumers and find out how many of them like each feature, and the extent to which they like each feature. From this sample we can draw inferences about the population. From these inferences we can design products to meet consumers' needs.

Many times when the data are analysed and presented, the knowledge of what is preferred, and by how much, is sufficient to answer the client's needs. Often, the researcher's own understanding of the market will fill in the missing gaps. However, too often the researcher is then asked a question such as "but why do 25% of the consumers prefer the orange, oval, tall jug?". Failure to answer this seemingly straightforward question will frequently result in the client doubting the wisdom of both the researcher and of the executive who commissioned the project.

This paper illustrates a simple method to collect the necessary information to answer this potentially awkward question. In addition the method can be expanded to help design better questionnaires. Indeed the paper starts with the process of questionnaire design and then moves on to the data collection process.

DESIGNING THE QUESTIONNAIRE

It is widely accepted, and frequently ignored, that in order to obtain high quality results we need high quality questionnaires. Too often questionnaires are deigned by client and researcher either brainstorming or simply re-using the questions they have used in other surveys.

Failure to ask the right question can be catastrophic. Anybody doubting this should read, or re-read, the books and articles on the Coca-Cola debacle of launching New Coke. In the reviews the then head of Coca-Cola's research identifies the key problem as one of asking the wrong questions.

In Conjoint studies we, at Deux, try to adopt a structured approach which is based on two phases (Generation and Refinement). Each phase comprises two or more steps. The next section of this paper outlines these phases and steps. However, it must be noted that the process detailed here is our target and is not achieved in every project!

Generating the Attributes and Levels

In our structured approach the first step is to compile a long list of attributes and levels. These features will then be refined down and then subjected to a Piloting process.

Repertory Grids

This standard element in the researcher's toolbox is conducted with clients, researchers, and respondents. The idea is to form a battery of attributes that differentiate products. In particular the techniques require the client and researcher to consider the whole picture, as opposed to the 'main objective' of the research project. In the form we use this technique we follow the following three steps.

1. Show groups of three brands/items/products;
2. The subject groups two of them together and describes why they are similar to each other and different from the third;
3. Select three more items and ask the subject to repeat the exercise but to come up with a new reason to group and differentiate.

Expert Views

The client will usually know the questions they need the answers to. In addition, a researcher with experience will know the sort of questions which probably ought to be added. One classic element here is adding a nil or none option to the client's list of levels.

Reducing/Refining the Attributes

At the end of the generation phase of the questionnaire development the researcher normally has too many attributes, too many levels, and concerns about their usability and appropriateness. The Refinement phase is conducted by applying a range of filters.

Usability Filter

Sometimes the generation will provide attributes that are salient to respondents but are not useful to the client. An example of this might be country of origin. The client has limited ability to change their position on this attribute. Since the information is not usable the strategy depends on a simple question.

If the feature is not likely to be important to respondents it should be left out of the study. If it is likely to be important then it should be asked in its simplest form (for example you might ask domestic versus foreign). This will leave more room for other attributes and levels.

Self-Explicated Filter

At this stage the researcher normally has too many features. One method to reduce them is to conduct a self-explication exercise with a small number of respondents. We achieve this by conducting an interview modelled on the ranking and importance section of the ACA interview. Any attribute that is unimportant to all respondents can be dropped. However, if the attribute is 'important' to the client you can either use this finding as a cause for re-drafting the attribute or for developing a two-stage conjoint.

When this filter produces too few candidates for deletion then the next step is to remove those attributes that are not important for at least one respondent.

Competence Filter

Can respondents answer the question? Clients can be very optimistic about the knowledge of respondents. Attributes and levels should be tested with typical consumers to see if they make sense. If the attribute cannot be readily understood then it should be deleted unless it is central to the research. If it is central to the research then methods must be found to 'educate' the respondent.

In finance related studies it is often necessary to spend 10 to 30 minutes going through service descriptions, showing brochures, and mocking up a prospectus. These steps are necessary to ensure that the respondent can answer the questions in a meaningful way.

Redundancy Filter

Do you have two or more attributes that measure the same thing? We see attributes such as *well trained reception staff* and *well trained retail staff*. In your market these attributes may always move together. The first step in looking for redundancy is simply to critically read your attribute list. The second step is to create a grid of products and to score all of the products on the levels. Then using a spreadsheet simply run the correlations for each attribute with all other attributes.

Every high attribute correlation should be reviewed to see if, in this market, at this time, these two attributes tend to be linked in the market place.

THE REPLAY PROGRAM

At this stage it is useful to include a description of the Replay program. The program is used in questionnaire design and in interpreting the final results. The use of the program will be described later in the paper. This section simply explains what the program is.

The Replay program is a simple program, written in Basic, which can be used to replay an ACA interview. The program simply takes the most recent ACA interview on the disk and shows what was on the screen during the interview and the options the respondent selected.

At its most simple the Replay program reads in the completed interview and shows each screen and the choices made. In this simple usage, the interviewer asks the respondent why they made each decision at this point and notes down the key comments.

In its more sophisticated guise the Replay program can be configured to:

1. Only launch when specific criteria are met;
2. Only display part of the interview;
3. Show a text window to type the open-ended text directly into the machine.

A BETTER PILOT

Once the basic questionnaire has been designed it should be pilot tested. This is true of almost all questionnaires in almost all forms of research. In this section the paper reviews how we use the Replay program to improve our pilot testing of ACA Conjoint interviews.

The draft interview is administered to a small number of test consumers. Teething problems are usually apparent in watching them attempting to complete the interview. The frames may not convey the intended meaning. Some of the levels may not be intelligible to a typical consumer. This is the normal pilot process.

After the interview is completed we run the Replay program to take respondents back through the interview, showing them each option they saw and the decision they made. At each stage, we can ask respondents to say why they made their choices. Many times we will find our research instrument works.

However, if the instructions are not clear we might find that the pilot consumers have not done as we expected. For example, they may have selected levels in ascending order of priority rather than descending. In the pairs section we may find that they selected one option because they did not understand the other. They may be confused because they could not bring themselves to believe that the other option was possible. For example, we have recently found problems convincing some respondents to accept that a product could be high in fat and delicious; a couple of decades ago we similarly found difficulty in testing the idea that a product could be low in fat and delicious!

As part of our pilot interviews we tend to remove all *a priori* ordering. Occasionally the researcher can be surprised by the order the respondents put on the statements. For example an English bigot might prefer a service to be in English only as opposed to say English and minority languages.

GENERAL USE OF REPLAY

The main use of the Replay program is in the context of providing additional information about the underlying reasons that cause consumers to make the choices that they make.

In the context of the real interview we will, when timing and budgets allow, run a pilot of say 10 to 20 interviews. These interviews will be recruited and conducted in accordance with our normal procedures. At the end of the interview the Replay program is run and the respondent is asked to state why they made the choices they did. Since the questioning follows the completion of the interview, the data can still be safely added to the rest of the study.

When we schedule a project we try to phase the data collection so we can review the first wave of data. If we have used Replay in the first wave we might discover the need to modify the interviewer instructions, or at worst the interview. More often we will review the first wave of data and decide to add Replay to some of the subsequent interviews. This tends to be a reactive process to questions that arise out of the initial review of the data.

Specific Cases

In this section the paper will draw on Deux's experiences with the Replay program. The examples are taken from several countries.

When More is Less

In general we assume that more is better. Indeed the Conjoint paradigm is based on the fact that the utility of a product is the sum of its parts. So the more parts we have the higher we expect the sum to be. However, the respondents do not always perceive it this way. The best place to see this is when the initial ranking section produces ‘surprising results’.

One study (in 1998) referred to a range of specific banking options. The customers were offered an attribute which had the following levels: counter service only, telephone banking only, Internet banking only, counter and telephone, counter and Internet, telephone and Internet, and counter plus telephone plus Internet. The client’s view was that this attribute was largely ordered but not sufficiently to just assume *a priori* ordering. The opening assumptions were:

1. The three options with only one feature would mostly be worst;
2. The three options with two features would mostly be preferred to single feature options;
3. The option with all three features would be universally preferred.

When we received the first wave of data we noticed this was far from the case. So we set about capturing some Replay information. We modified the Replay program to just query part of the ranking section. The interviews were specifically triggered by respondents who did not rank the data in ways we would expect. We found two interesting groups.

There was a small but significant group of people who preferred counter service only to counter plus one or more features. In analysing their data we could see that they felt that a service which offered more than just counter would offer a counter option which was less good. For example the counter service might have shorter hours, the bank costs might be higher because of the unwanted telephone option, or they might receive a hard sell to move over to the telephone or Internet service.

There was a somewhat larger minority who were happy to rate telephone plus counter as superior to counter or telephone but who rated any option including the Internet as worse than the same option without the Internet. Here the main motivator was fear. Many respondents said that they thought a bank which offered Internet services was less reliable, more prone to be ‘hacked’, and generally less trustworthy.

Armed with this explanatory detail we were able to go back to the client with options in the modelling which provided options in interpretation. They were also able to review their communication so that telephone banking was seen as a universal plus (for example it reduced queues at the counter, reduced overall bank charges, and helped ensure counter staff had more time to deal with the people who still wanted to visit the office). The Internet option was promoted as being very separate from the rest of the service, only promoted to likely users, and a heavy stress was placed on the security provisions.

Pictures into Words

In many Conjoint studies a non-text element is added. These stimuli may be pictures, mock-ups, or multimedia clips. In this more complicated environment the researcher needs more than just which option is preferred and by how much.

For example, in one case we were researching an alcohol re-packaging study. The options included five price points, three bottle colours, two bottle shapes, three label designs, and five propositions. The main objective of this stage of the research was to design three products to be created for an extended in-home product placement test. From the start of the process we knew that just knowing what options were preferred would not be sufficient to please the client. This study was run with a percentage of the respondents going through the Replay process.

The analysis of the data showed us a number of important features that helped design the three products for the final test. For example the most preferred colour was originally chosen because of an historical link with the product and the place of its manufacture. This link was not stressed in the communication since the company assumed it was widely known. The data showed the respondents liked the glass but that the intended product link went completely unnoticed!

How Green is My Product

In a large number of Conjoint tests with White Goods and Consumer Durables across Europe we have seen considerable segments who prefer the 'Green' products. For example there is a sector which prefers products that use less water, less detergent, and fewer additives. In several of these studies we have used a Replay approach. One interesting finding was that in most cases the respondents were choosing the Green options because they cost less to run. The comments made by these respondents were stressing the reduced cost of lower power consumption and the reduced cost of detergents and/or fuel. Clearly, the lesson is that most Green buyers expect a premium in terms of running costs, even if they are prepared to pay more for the original purchase.

Fat is the Enemy

In 1997 we conducted a Conjoint study into a range of 'healthy' yoghurts. The key attributes included calories, grams of fat, size, type (e.g. set versus live). The grams attribute went in units from 10 grams to 0.1 grams (0.1 grams is about 0.00353 ounces). The first set of the data surprised most of the research team when it showed that the utility difference between 1 grams and 0.1 grams was large, for most respondents.

In order to investigate the study more fully some of the remaining interviews were conducted using a Replay option. The text data made it clear that most respondents had no 'real' understanding of how little a gram was. Very few realised that there was no effective difference between the low values for fat. Indeed the reference point the respondents were using was the set that the interview had created.

By contrast the value ascribed to calories was very low, many people saying that calories do not matter it is only the fat that affects you! Interestingly the same research team had conducted a similar study ten years before. At the earlier data calories were almost the only important feature.

This feature can be generalised to any market where respondents may not understand the units, no matter how simple they seem. For example, in a washing machine study the respondents will often take their cue on spin speeds by the RPM shown in the interview.

CONCLUSIONS

The Conjoint process is very powerful at determining preferences and quantities. However, it can only do that if it asks the right questions. The best way to ask the right questions is to have a thought through approach to questionnaire design and to utilise good qualitative inputs to that process.

One weakness of traditional Conjoint studies is that it can be weak on explaining why. This paper has demonstrated that this weakness can be addressed. The paper also points out that gaining a qualitative understanding is easier if the scheduling of the project is such that the data is delivered in waves, rather than all at once.

COMMENT ON POYNTER

Sam Kingsley
dataDirect

This paper should alert conjoint users to the consequences of inadequate preparation for conjoint data collection. The author suggests several useful techniques to systematically address questions of (1) whether the right attributes and levels are being specified; (2) whether respondents understand the conjoint questions the way they were intended; and, (3) how respondents' answers should be interpreted. The author candidly states that his self-prescriptions are completely followed in only a small percent of projects.

These ideas can be extended into greater practice by:

1. Including in conjoint interviewing programs methods to automatically ask respondents to explain their answers. In particular, responses that differ from the predicted ones would be valuable to have respondents explain. This extension is currently being considered by our company for inclusion in ACQNET, a service which provides Web-based ACA interviewing.
2. Giving respondents feedback about their conjoint choices — and what that means about their values — is an important direction in the future of conjoint. Once respondents receive such feedback, they sometimes want to say "But, what I meant to say was ...." Building into conjoint interview procedures an opportunity for respondents to explain their choices in the context of feedback is an opportunity to increase the value of the interview to both the respondent and the interviewer. It may increase our opportunity as researchers to "leave the respondent whole."

In addition to asking "But why?" during pre-testing, if the techniques can be routinely included in conjoint interview programs, then "But why?" can be asked of all survey respondents.

CONVENTION INTERVIEWING—CONVENIENCE OR REALITY?

Don Marshall
Isis Research

BACKGROUND

My work in the pharmaceutical industry has frequently involved evaluating the market potential offered by a new product that is currently in clinical development. The objective of these market evaluations is most commonly to help determine whether the compound in question offers enough sales potential to justify the investment of an additional $100-$200 million or more in clinical development costs. This marketing research is obviously critical in deciding whether to continue the development program, or in deciding whether to in-license the rights to a compound. Additionally, for compounds in the later stages of development it is frequently desirable to carefully evaluate the emerging product profile in order to determine which components of the product profile are really critical to market acceptance of the product to insure that there is adequate clinical support for those claims. Frequently the attributes that are critical to market acceptance are very different from those that are required to obtain regulatory approval to market a product.

By way of background for those not familiar with the pharmaceutical industry, the FDA regulates the industry in an effort to prevent the marketing of ineffective products or the use of unjustified product claims in the promotion of marketed products. Federal statutes require that a product has to be proven to be both safe and effective before it can be approved for marketing. This is accomplished through a series of pre-clinical and clinical trials over a period of many years prior to submitting the NDA (New Drug Application) to the FDA for approval. Pre-clinical testing involves testing the compound in a series of laboratory tests in tissue and animal models to determine if it really seems to work in the intended fashion. Following this, clinical trials are begun in humans. Phase I clinical trials are relatively small trials conducted in healthy volunteers to evaluate the safety of the compound as well as to begin developing information on dosing. Following the successful completion of Phase I trials the compound then moves into Phase II trials. Phase II trials are larger and involve patients with the disease or condition that the compound is targeting. This step involves more precise dose ranging as well as a preliminary evaluation of the safety and efficacy of the compound in treating the target disease. Phase III trials are much larger, and are designed to document in a rigorous statistical process the safety and efficacy of the product. With rare exceptions, Phase III trials are double-blind, placebo controlled studies (neither the patient nor the physician treating that patient know whether that patient is receiving a placebo or an active compound. Nobody knows until after the code has been broken at the completion of the trial). Each stage in this development process gets progressively more complex and expensive, and the vast majority of compounds synthesized fall by the wayside. In fact, the National Cancer Institute and the FDA estimated in 1995 that, on average, it took the synthesis of 5,000 compounds to get one drug approved. The Pharmaceutical Research and Manufacturers Association (PhRMA) has estimated that the average cost to develop one new pharmaceutical product is $500 million. It has also been estimated (PhRMA) that only 3 out of 10 marketed Rx drugs recover their development costs. Understandably, the companies funding this research would like a reliable, unbiased means of estimating the true market potential for a compound in development so that they can make the best resource allocation decisions for the different compounds in their pipeline. Over the years, conjoint analysis has proven to be a very effective tool to aid in this process.

As you know, conjoint analysis involves identifying the relevant product attributes that contribute to the decision to use the new Rx medicine, as well as the different levels of each attribute. These are then combined in different fashions to develop a series of hypothetical product profiles. These combinations are themselves determined in a rigorous statistical algorithm. By presenting these hypothetical profiles to respondents and asking them to indicate their relative preference for each, it is possible to statistically infer the importance or weight that they attach to each level of each attribute. Physicians would seem to be ideal respondents for a technique like conjoint since they have been rigorously trained to evaluate the different attributes of the medicines that they are considering prescribing for their patients. They are very comfortable trading off the different levels and types of safety and efficacy of the different drugs that they prescribe, as well as balancing the various convenience attributes for the same products with the other aspects of the product profile. Convenience attributes can include such things as dosing frequency, tablet size, and the availability of multiple forms (solid, liquid, IM, IV). The scientific basis of medicine and their daily practice seem tailor-made to make physicians ideal conjoint respondents. There are, however, some significant problems with doctors as survey respondents – they are difficult to recruit and expensive to interview.

The practice of medicine is obviously very complex and is also evolving quite rapidly. While it is very important for physicians to keep up with the latest advances in medical practice, the daily demands of their practice make it very difficult for them to fit this time for learning into their busy schedule. For this reason, physicians frequently attend medical conventions where they can devote three or four days to learning about recent developments in medical practice, including recently introduced and soon to be approved drugs. Within the pharmaceutical industry, interviewing doctors at these conventions has long been touted as a very efficient and economical means of interviewing a large number of physicians in a relatively short time. Each medical specialty will have an annual meeting of their society. Depending on the specialty, there may also be state and regional meetings that also offer the opportunity to interview (relatively) large numbers of physicians. Some of the more common specialties (GP/FPs, IMs, OBGs, surgeons, etc.) will have thousands of physicians attending their annual meeting. If the timing is convenient, these medical conventions clearly present a wonderful opportunity to complete a large number of physician interviews in a relatively short time frame (most medical conventions are three days long).

There is a potential drawback to conducting pharmaceutical marketing research at medical conventions however. For as long as medical conventions have been touted as an efficient venue to interview physicians, they have also been criticized as providing a non-representative sample. These critics have claimed that doctors who attend medical conventions are different in some key fashion from those physicians who do not attend medical conventions, and that convention interviewing merely provides a convenience sample, but not a representative or projectable sample.

As a result, convention interviewing has endured a somewhat cyclical popularity over at least the last 20 years. There have been periods of heavier reliance on convention interviewing to capitalize on the increased efficiency that it allows, followed by periods of diminished use as the industry becomes more concerned with the potential lack of representativeness provided by convention goers.

The study discussed in the current paper was conducted in the early 90's to evaluate the market opportunity for a potential new product to treat Alzheimer's Disease. While the primary objective of this study was to evaluate the market potential for this compound, it also presented an ideal opportunity to evaluate the difference between convention interviews and traditional face-to-face interviews conducted in a central facility.

METHODOLOGY

Since the product was in the early stages of development, its product profile was not yet very clear, and was likely to change many times as development continued and new clinical data was reported. In this situation conjoint analysis would seem to be the ideal methodology due to the tremendous flexibility offered by the simulation models that can be developed from this technique. We elected to use preference based conjoint for this project, specifically ACA. This was a fairly typical conjoint study in many ways. A list of relevant attributes and levels was developed based on a series of focus groups with physicians from the appropriate specialties (GP/FP, IM, and neurologist) as well as from secondary research on competitive products in the pipeline and interviews with our own clinicians. This resulted in an attribute list with nine different attributes and a total of 29 different levels. These attributes and levels were then programmed into ACA. The ACA questionnaire was embedded in a Ci3 interview that contained a series of questions on relevant background and practice issues. The completed questionnaire was then pre-tested with a small number of physicians to ensure that the wording was clear, that the question flow was appropriate, that nothing superfluous had been included, and that nothing important had been omitted. Following the pre-test, the final questionnaire went to the field for central facility interviews. Central facility interviews were conducted in six different cities spread around the country. The sample size that resulted from this phase of the research was 229 usable interviews.

These central facility interviews occurred in October, shortly before the annual convention of the Southern Medical Association. The SMA is a general medical conference, not limited to any one specialty. This allowed us the opportunity to interview both GP/FPs and IMs (there is generally not adequate representation of the smaller specialties like neurology at general medical conventions like the SMA to develop a reasonable study sample). We rented a booth at the SMA and furnished it with carpet, tables, chairs and six PCs to conduct the interviews. At the end of the convention, we had interviewed 56 primary care physicians.

Immediately following the completion of all central facility interviews, the pooled data was analysed in a fairly standard fashion – a market simulation model was developed incorporating a base case or most likely profile for the compound in question. Then this base case profile was systematically modified on one attribute at a time and the model was re-run to determine how sensitive the estimated product use was to each attribute in the profile. After these results and the accompanying forecast and recommendations had been presented to management we then went back to the raw data to conduct our own marketing research evaluation of the comparability of the two different data collection techniques.

CENTRAL FACILITY VS. CONVENTION INTERVIEWS

Since the group that I worked in at the time was responsible for evaluating new products, our primary interest in this analysis was to determine the representativeness and projectability of convention interviews in the evaluation of new products. We decided to concentrate our analysis on a comparison of the conjoint data that was obtained from the two data collection techniques since conjoint was our primary tool in evaluating new products and developing forecasts for them. Since the individual utility scores to a large extent drive the simulation model results, our analysis consisted of a comparison of the utility scores obtained from the central facility interviews with those obtained from convention interviews.

Table I presents the average utility scores for each level of each attribute for both GP/FPs and IMs for both data collection techniques. There were insufficient neurologist convention interviews to conduct a meaningful comparison. A quick examination of the two columns of utility scores for GP/FPs shows that there is a remarkable consistency in the utility scores obtained from the two different data collection techniques. In fact, for each specialty a univariate t-test finds that there is only one attribute/level with a difference between the two techniques that is large enough to be considered statistically significant by a univariate t-test.

With 29 utility scores for each respondent, the traditional univariate t-test is not really the appropriate test to determine the comparability of the two different techniques. With the univariate t-test you would compare the utility scores for the first level of the first attribute to determine if there was a statistically significant difference between the utility scores obtained from the two different data collection techniques. If you use a 95% confidence level for this t-test, there will be a 5% chance that you will conclude that the observed difference is significant when in fact it was due to chance (type I error). When you proceed to do the t-test for the second utility score you again have a 95% confidence level and a 5% chance for a type I error on this test. However, while you only have a 5% chance of making a type I error in each, combined you will have about a 10% chance of incorrectly concluding that there is a significant difference when in fact that difference is solely due to chance. By the time you have done a univariate t-test for each of the 29 utility scores in this study you will have about a 77% probability of saying that there is at least one significant difference when in fact the difference was solely due to chance. Another way of looking at this is that a 5% chance of a type I error means that, on average, in one out of every twenty t-tests that you perform, you will say that there is a difference when in fact the difference was due to chance.

Hotelling’s T-squared test is the multivariate extension of the t-test, and provides a means of simultaneously testing all 29 utilities to determine whether or not the two populations are in fact different. When Hotelling’s T-squared test was performed on the two specialty populations in this data, the results indicated that there was no statistically significant difference between the utility scores obtained from central facility interviews and those obtained from convention interviews (at the 95% confidence level). The results were the same for both GP/FPs and IMs. Since the results of this study show that there is no statistically significant difference between the utility scores obtained from convention interviews and central facility interviews, we can conclude that the two populations are in fact the same in terms of the values that they place on different product attributes when evaluating potential products to prescribe. This then leads to the conclusion that convention interviews provide an efficient, economical means of interviewing physicians and obtaining reliable information about the most likely use of potential new products.

In addition to the ACA portion of the interview used for this study, we also asked respondents a series of background questions about their practice. The responses to these questions are summarized in Table II. An examination of these background questions reveals that there are a few statistically significant differences between the two groups of respondents. Among the GP/FPs, there is a statistically significant difference between the two groups of physicians in terms of the location of their practice – convention respondents are more likely to be from a rural location rather than urban or suburban. Additionally, convention GP/FPs overall reported a larger percentage of their time was spent in the hospital or university setting, as well as a larger percentage of their patient load in the 60-84 age group. Finally, convention GP/FPs were more likely to report that they referred their Alzheimer's patients to a specialist to confirm their diagnosis, and correspondingly less likely to do so to satisfy family members.

Turning our attention now to the background questions and practice characteristics of the IMs, we find a somewhat similar picture. Geographically, the convention IMs were more likely to be rural than their central facility counterparts. As with the GP/FPs, convention IMs spent a greater percent of their time in the hospital (or government/university) setting, and have a larger percent of their patient load among those aged 60+.

CONVENTION INTERVIEWING

While the data presented above indicates that convention interviewing is a viable alternative to central facility interviewing within the pharmaceutical industry, it is clearly not appropriate in all instances. I will now briefly discuss some of the key criteria that must be addressed when considering the appropriateness of a convention interview for a project.

1. **Timing:** Since most medical societies only have annual meetings, there is a relatively narrow window of opportunity for convention interviewing. At least with new pharmaceutical products clinical trials are meticulously planned so that you know well in advance when the product will be coming out of the clinic, allowing you to plan your marketing research activities accordingly, and to take advantage of conventions if appropriate. In order to take advantage of the efficiency offered by convention interviewing it is sometimes necessary to conduct a study a few months earlier than you otherwise would. Obviously before you can do this you need to be sure that you have enough information to define the relevant issues, and that there is a low probability that new data will become available in the interim to invalidate the research that was done at the convention.

2. **Interview Length:** The length of the interview is critical. Doctors have come to the convention to learn, and will not sit still for a lengthy, boring interview. In general, I have found that an ACA interview of approximately 20 minutes or less works fine at medical conventions. Clearly, a lot will depend on the subject matter of the interview. A questionnaire presenting clinical data on new products to treat Alzheimer's or MS, or any other disease for which there are currently very few satisfactory therapeutic options will

generate much more interest than a survey on a new cough/cold product. Along with this, we always tell the physicians before they sit down approximately how long the interview will take. This prevents them from expecting a 5-minute interview and getting extremely impatient after 15 minutes.

3. **Gifts:** It clearly helps to have a nice give-away as a thank-you for the time that the doctor has spent answering the survey. If the convention is being held in a nice area (and most medical conventions are) perhaps a disposable camera would be well received for the doctor to use with his/her family after the convention. For many years we gave away ties that we had made with the caduceus embroidered on it. While these were quite popular at the time, I'm not sure how well it would work in today's business casual environment. Good quality pens have also worked well. The key has been to find something with relatively high perceived value, or something that is appropriate for the venue of the convention, and that is readily portable. For example, disposable cameras have worked very well in places like Orlando where the doctor is likely to have brought his/her family along.

4. **Pre-Test:** While pre-testing is always important, it is especially critical for convention interviewing since there is such a short period of time to collect the data and little chance to correct mistakes in the questionnaire on the convention floor. These conventions only happen once a year, the exhibit floor is typically open for only 2 ½ days, and the heaviest traffic is on the first day. This makes it imperative that every effort be made to have the questionnaire exactly the way you want it at the beginning of the convention.

5. **Convention Suitability:** You need to examine the convention prospectus carefully prior to registering for booth space. In my case with the pharmaceutical industry, some meetings are inappropriate because they are too academic in their approach and content. While academic meetings are clearly appropriate for some projects, most of the time we want to interview physicians who are actually treating patients. You need to ensure that the convention goers are the appropriate respondents for you to draw inferences from and project to the population for your decision making process.

6. **Booth Set-up:** Booth space at a medical convention is typically 10' x 10'. We have found that in this space we can comfortably set up 6 computer interviewing stations. This allows us to make maximum use of the space and time available without making the respondents uncomfortable.

7. **Over-recruit:** One thing that I did not mention earlier was that all analyses comparing convention doctors to central facility doctors were done with "usable respondents". I usually identify usable respondents with the correlation coefficient between the actual and predicted responses to the calibration concepts provided by ACA. With all of the noise and distractions on the convention floor, it is reasonable to expect the discard rate to be higher and to allow for it. The discard rate (using a minimum R of 0.7) may be as high as 15-20% for convention interviews, compared to 5% or so for central facility interviews.

CONCLUSIONS

Overall, these results indicate that convention interviewing is an efficient and economic alternative to central facility interviewing when using conjoint analysis to interview doctors to evaluate potential new products. These results indicate that there is no statistically significant difference between the utility scores developed with either data collection technique. The analysis of the background data indicates that convention interviewing may provide broader representativeness in terms of physician practice type and location. On the other hand, it appears that the patient population treated by convention doctors is somewhat older than that treated by central facility respondents.

While the convention setting offers a broader variety of respondents and has the potential to provide a somewhat more representative sample than central facility interviewing in metropolitan areas, care must be taken to exercise appropriate quality control measures in the interview and screening process.

TABLE I

	GP/FP			IM		
	UTILITY SCORE			UTILITY SCORE		
ATTRIBUTE/LEVEL	Central Facil.	Conv.	t-score	Central Facil.	Conv.	t-score
Side Effect 1:						
Level 1	0.177	0.180	-0.076	0.213	0.189	0.519
Level 2	-0.084	-0.094	0.181	-0.049	-0.091	0.887
Level 3	0.007	0.078	-1.463	0.033	0.078	-0.932
Level 4	-0.286	-0.291	0.094	-0.334	-0.278	-1.162
Side Effect 2:						
Level 1	0.340	0.316	0.375	0.377	0.336	0.741
Level 2	0.072	0.155	-2.149*	0.104	0.107	-0.081
Level 3	-0.134	-0.175	0.867	-0.112	-0.108	-0.061
Level 4	-0.464	-0.423	-0.617	-0.506	-0.436	-1.030
Efficacy 1:						
Level 1	0.472	0.474	-0.024	0.522	0.483	0.574
Level 2	0.013	0.017	-0.080	-0.011	-0.003	-0.227
Level 3	-0.625	-0.585	-0.457	-0.613	-0.556	-0.921
Efficacy 2:						
Level 1	-0.594	-0.671	0.939	-0.599	-0.519	-1.213
Level 2	0.033	0.088	-1.244	0.033	0.050	-0.386
Level 3	0.421	0.489	-1.024	0.464	0.394	1.300
Efficacy 3:						
Level 1	0.584	0.619	-0.382	0.683	0.578	1.173
Level 2	0.220	0.233	-0.206	0.209	0.136	1.264
Level 3	-0.327	-0.294	-0.500	-0.333	-0.246	-1.483
Level 4	-0.663	-0.685	0.227	-0.696	-0.569	-1.590
Efficacy 4:						
Level 1	0.246	0.313	-1.602	0.283	0.236	1.205
Level 2	-0.339	-0.376	0.708	-0.351	-0.286	-1.816
Cost/Day:						
Level 1	0.315	0.315	0.059	0.376	0.306	1.466
Level 2	-0.028	0.010	-0.911	-0.023	-0.033	0.324
Level 3	-0.427	-0.416	-0.174	-0.455	-0.349	-2.190*
Dose Frequency:						
Level 1	0.284	0.284	-0.007	0.322	0.309	0.301
Level 2	-0.018	0.021	-1.068	-0.021	-0.035	0.467
Level 3	-0.405	-0.400	-0.096	-0.404	-0.350	-1.153
Side Effect 3:						
Level 1	0.351	0.410	-1.157	0.384	0.308	1.502
Level 2	-0.037	-0.019	-0.507	-0.003	-0.004	0.014
Level 3	-0.454	-0.486	-0.486	-0.484	-0.381	-1.968

* Significant difference with independent samples t-test @ p = 0.05

TABLE II

BACKGROUND VARIABLE COMPARISON

VARIABLE	GP/FP		IM	
	Central Facil.	Conv.	Central Facil.	Conv.
Percent Urban	43%	35%	36%	58%
Percent Suburban	57%	45%	64%	33%
Percent Rural	0%	16%	0%	8%
Percent Office	96%	65%	99%	50%
Percent Hospital	4%	19%	1%	17%
Percent University/Government	0%	16%	0%	33%
Percent Patients under 40	34%	29%	20%	11%
Percent Patients age 40-59	34%	26%	30%	22%
Percent Patients age 60-84	24%	36%	40%	48%
Percent Patients age 85+	7%	9%	9%	18%
Avg. Number Demented Patients	18	23	22	28
Avg. Number Alzheimer Patients	11	13	14	19
Percent Referring Alzheimer’s Patients	78%	65%	66%	63%
Percent Referring for Confirmation	57%	80%	62%	60%
Percent Referring for Treatment Recommendation	17%	10%	9%	20%
Percent Referring to Satisfy Family	24%	10%	29%	20%
Percent Referring for Other Reasons	2%	0%	0%	0%
Percent Referring to Neurologists	93%	85%	93%	80%
Percent Referring to IMs	0%	0%	2%	7%
Percent Referring to Psychiatrist	5%	10%	3%	7%
Percent Referring to Social Workers	0%	5%	0%	0%
Percent Referring to Others	2%	0%	2%	7%
Percent Alzheimer’s Patients w/Mild AD	37%	33%	29%	28%
Percent Alzheimer’s Patients w/Moderate AD	37%	33%	42%	49%
Percent Alzheimer’s Patients w/Severe AD	26%	33%	29%	23%

DISK-BASED MAIL SURVEYS: A LONGITUDINAL STUDY OF PRACTICES AND RESULTS

Arthur Saltzman, Ph.D.
Para Technologies
William H. MacElroy
Socratic Technologies

BACKGROUND

This paper compares the results of two surveys conducted among market research professionals who use the Disk-by-Mail (DBM) data collection method. The two surveys were conducted four years apart, the first in December of 1994 and the second in November of 1998. Although both surveys contained many identical questions, the 1998 survey contained a few additional questions about the future of DBM vs. online research methods.

The purpose of the original 1994 survey was to determine, among market research professionals, what practices were most effective when conducting DBM surveys. However, the 1998 survey additionally sought to determine the current level of usage of the DBM method relative to online survey methods and also to determine the predicted levels of usage for the different methods three years from now.

This paper will outline the changes in DBM practices over the years and discuss the future of DBM vs. online research techniques.

RESEARCH METHOD

Data for the 1994 survey were collected via a Disk-by-Mail survey programmed by Populus, Inc. using Sawtooth's Ci3 software. The survey was sent to 26 market researchers in November of 1994. These were all known by the authors to have conducted DBM research projects. All responses were received by December 29, 1994.

Data for the 1998 survey were also collected via a Disk-by-Mail survey. The survey was programmed using Socratic Software's Visual Q program. The survey was sent to 40 market researchers on November 5, 1998. All responses were received by December 8, 1998.

The 1994 survey contains responses from 10 research organizations with data from 21 studies, while the 1998 survey contains responses from 13 research organizations with data from 20 studies.

It should be noted that there were a few researchers who participated in the 1994 survey who did not participate in the 1998 survey because they indicated that they no longer used the DBM data collection method.

DETAILED FINDINGS

Changes in DBM practices

The most noticeable change in DBM practices from 1994 to 1998 is the movement toward “personalization.” Researchers today are putting more of an effort into making their surveys appear individualized or personal. They are also making more of an effort to “target” their sample. Specifically, a higher proportion of researchers in 1998 compared to 1994:

- Sent notifications prior to mailing disks to respondents:

Notifications Sent Prior To Mailing DBM

	1994	1998
Sample Size	(n = 21)	(n = 20)
Yes	38%	65%
No	62%	35%

- Pre-screened their respondents:

Pre-Screened Respondents

	1994	1998
Sample Size	(n = 21)	(n = 20)
Yes	24%	50%
No	71%	50%
Don’t Know	5%	--

- Sent personalized outbound envelopes:

Sent Personalized Outbound Envelopes

	1994	1998
Sample Size	(n = 21)	(n = 20)
Yes	81%	90%
No	15%	5%
Don’t Know	5%	5%

- Sent personalized cover letters:

Sent Personalized Cover Letters

	1994	1998
Sample Size	(n = 21)	(n = 20)
Yes	33%	72%
No	62%	22%
Don’t Know	5%	6%

- Provided a telephone number for assistance:

Provided A Telephone Number For Assistance

	1994	**1998**
Sample Size	(n = 21)	(n = 20)
Yes	86%	100%
No	9%	--
Don't Know	5%	--

Not surprisingly, the 1998 data also indicate an increase in the functionality of the surveys compared to four years ago:

- A higher proportion of researchers in 1998 indicated that their surveys included a restart option compared to four years ago:

Restart Option Provided

	1994	**1998**
Sample Size	(n = 21)	(n = 20)
Yes	33%	75%
No	62%	25%
Don't Know	5%	--

Technological advances can also be seen in terms of the physical disks themselves. Only one out of the 20 surveys reported on in the 1998 study was programmed on a 5.25 inch floppy disk. The other 19 were all programmed on a 3.5 inch floppy disk. In the 1994 study, 71% of the surveys were programmed on a 5.25 inch floppy disk.

Response Rates

In spite of this increase in personalization, targeting and functionality, the reported response rates of DBM surveys included in the 1998 data are lower than those reported in the 1994 data. Additionally, the expected response rates were notably lower in the 1998 results (see table below).

DBM Response Rates

	1994	**1998**
Sample Size	(n = 21)	(n = 20)
Expected Response Rate	46%	41%
Actual Response Rate	52%	33%

A possible reason for this finding might have to do with the fact that the DBM was a relatively novel technique in 1994. It could be that the novelty has worn off somewhat and that the degree of personalization is now necessary to maintain viable response rates. Other explanations (such as lower incentive values in 1998) no doubt also play a role.

General Findings

The use of DBM is, of course, best used with those anticipated to have a high degree of access to the equipment needed to run the survey disks. In both studies, the mean percent of potential respondents with access to a PC was 93%. There does appear to be an increase in the expected percentage of respondents with universal access.

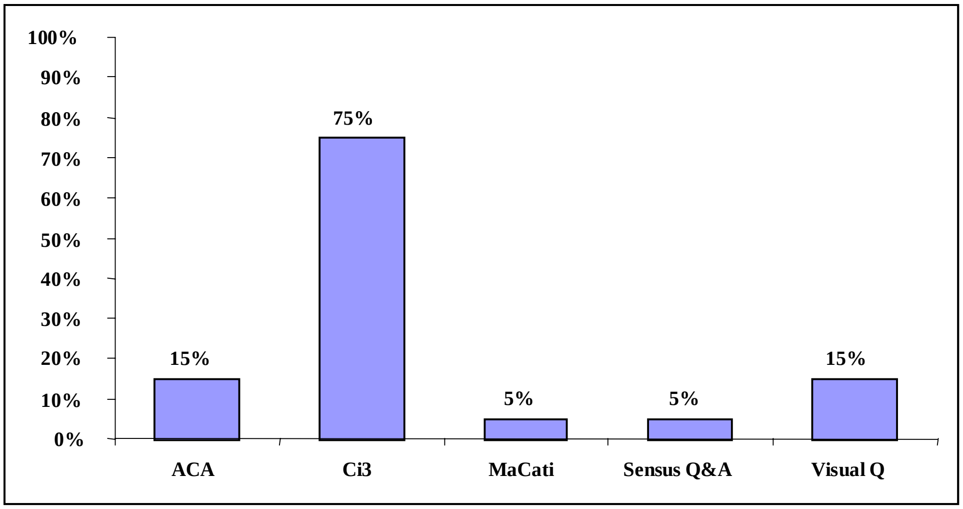
Anticipated Respondent Access to PCs

	1994	1998
Sample Size	(n = 21)	(n = 20)
Percent expected to have nearly 100% access	47%	70%
Mean % with access	93%	93%

In both surveys, the vast majority of agencies believed that almost all potential respondents would be comfortable using the PC to answer the survey questions. In 1994, agencies reported that approximately 86% of users would be comfortable; in 1998 agencies said about 92% of users would be comfortable.

The leading software being used by the agencies polled is Sawtooth's Ci3. It was used in 15 of the 20 surveys studied.

Types of Software Used
(*Note*: Multiple Types Per Project)



The subjects/industries being studied through the use of DBM technique have not changed significantly from the proportions found in the 1994 study.

Subject of DBM Surveys

	1994	1998
Sample Size	(n = 21)	(n = 20)
Computer-related	67%	65%
Telecommunications	5%	10%
Services	14%	10%
Other Topics	14%	15%

Respondent Interest and Affinity

"Interest in the topics to be studied" and "the degree to which respondents would benefit from the outcome of the survey," and "affinity for the sponsor," are three variables that have been hypothesized as having an impact on response rates and cooperation. In 1998, agencies reported much higher anticipated interest, benefit, and *sponsor* affinity from the DBM studies catalogued previously.

Anticipated affinity for the *research organizations* conducting DBM studies has dropped dramatically since 1994. At that time, affinity for the research organization was somewhat ambivalent (45 on a 100-pt. scale, where 100 = High Affinity). In 1998, that rating has fallen to 17.

Respondent Interest, Derived Benefit, and Affinity with Sponsor and Research Organization

	1994	1998
Sample Size	(n = 21)	(n = 20)
Mean Interest in Topic (100 pt scale; 100 = Extremely Interested)	41	69
Mean Benefit from Participation (100 pt scale; 100 = Benefited a Great Deal)	35	51
Mean Expected Affinity for Sponsor (100 pt scale; 100 = High Affinity)	46	81
Mean Expected Affinity for Research Organization (100 pt scale; 100 = High Affinity)	45	17

Mechanics and Logistics

The success of all mail-based surveys has been shown to vary according to the delivery methods, degree of personalization, and other tactical considerations. The mechanics associated with performing DBM studies, does not appear to have changed significantly over the past four years. There is some indication that agencies may be using more expedited delivery services.

Outbound Service Used

	1994	**1998**
Sample Size	(n = 21)	(n = 20)
US Postal Service First Class	72%	55%
US Postal Service Priority Class	0%	10%
Overnight Courier (e.g. FedEx, UPS)	10%	10%
US Postal Bulk	10%	5%
Don’t Know	5%	10%
Some Other Method	0%	10%

Cover letters explaining the process and purpose of the studies were nearly universally used in both 1994 (95%) and in 1998 (90%). As mentioned earlier, one major difference was that the letters in 1998 were more often personalized with the respondents' names (72%) than was the case in 1994 (33%).

The guarantee of anonymity was universal in 1998 (whereas only 85% of the studies reviewed in 1994 carried the same promise). At the same time, the identification of the sponsor/client's name grew from 57% in 1994 to 70% in 1998.

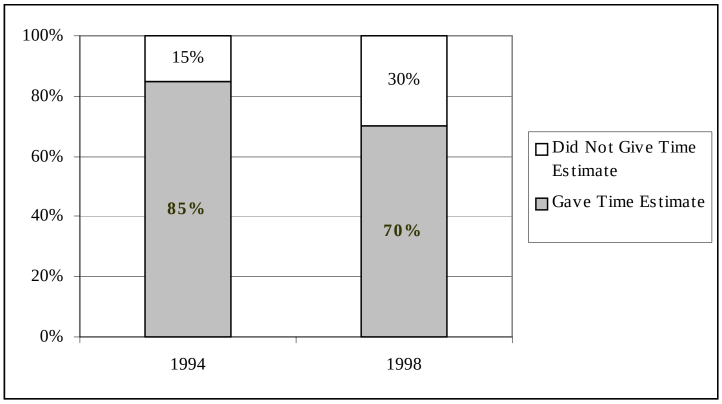
Almost all return postage was pre-paid both in 1994 (91%) and in 1998 (90%). An almost even split between DBM studies using US Postal Service First Class (pre-stamped) and Business Reply Mail was seen in both waves of this research. Few agencies in either wave used overnight couriers for returning completed survey disks.

The reported use of reminders to increase participation was up somewhat in 1998. In addition, the use of telephone calls to encourage returns increased significantly (from 23% in 1994 to 67% in 1998).

As discussed earlier, response rates have dropped about 20 percentage points since the first wave of this study. Accordingly, it appears that agencies are sending out more disks (the mean number sent in 1998 was 1,642 per study versus 1,331 per study in 1994). The use of the more robust 3.5 inch disks probably explains the finding that number of *incomplete* or *unusable* disks returned has dropped sharply from a mean of 61 disks per study in 1994 down to only 16 per study in 1998. Disks returned as *undeliverable* remained constant at approximately 21 disks per study in both waves of research.

Fewer agencies in the recent study gave respondents an estimate of the time required to complete the survey. The average *estimate* of time given was similar in both waves of this research (1994, 24 minutes; 1998, 26 minutes). In both waves, the *actual* amount of time that respondents took was either the same as or less than the estimate given (1994, 24 minutes; 1998 19 minutes).

Time Estimates Given?



In 1994, virtually all agencies reported doing DBM studies only in English and only in the U.S. In 1998, about 1/3 of the studies reported the inclusion of other countries (primarily Canada). In addition, at least one study was programmed in German.

Incentives

In 1994, 86% of the studies recorded offered some form of incentive, in 1998 this number has fallen to 75%. In addition, the proportion of agencies paying incentives only upon receipt of a usable, complete survey has increased dramatically.

How Respondents Qualify for Incentives

	1994	1998
Sample Size	(n = 18)	(n = 15)
Included in Outbound Package	22%	20%
Sent After Receipt of Completed Survey	17%	47%
Sent After Receipt of Complete or Incomplete	11%	--
Drawing at Later Date	39%	33%
Other Incentive	11%	--

Lotteries in general were not used as frequently in the studies described in 1998. Cash and prizes on a quid pro quo basis were more common.

Incentives Offered

	1994	1998
Sample Size	(n = 18)	(n = 15)
Monetary incentive for each respondent	30%	33%
Charitable contributions	12%	--
Lottery for cash incentive	12%	--
Lottery for prize	29%	27%
Copy of results	6%	7%
Free gift for every respondent	--	33%
Other incentive?	12%	--

The total value of incentives for studies reported in 1998 were about half as rich as those offered in 1994. This could help explain the lower response rates covered in the more recent poll.

Total Value of Incentives Offered

	1994	1998
Sample Size	(n = 18)	(n = 15)
Mean Value	$1,117	$550

DBM vs. Online Research Methods

The majority of researchers interviewed expressed a preference for online research methods when compared to DBM.

Preference For DBM or Online Methods

Sample Size	(n = 20)
Online	55%
DBM	25%
Both Equally	15%
Neither	5%

Some of the reasons respondents cited for preferring online methods were:

- “Ease of implementation, quicker response time, lower cost.”
- “Faster, cheaper. Most clients who want to use technology-based surveys don’t want to wait for the disks to be returned.”

- "It is easier to coordinate the implementation. With DBM there are a lot of details concerning the mailing, and respondents are more likely to have trouble with a DBM due to system configurations, etc."

One of the biggest concerns with online research appears to be the limited functionality of the technology currently available. Another major concern is the lack of Internet access among potential respondents.

Here are some comments from researchers who indicated that they prefer the DBM method:

- "Online software is still in its infancy. The population willing to use online is still developing. You can send materials with DBM."
- "Simplicity—they have the physical disk and not much can go wrong. Avoids problems caused solely by the technology for example, e-mail and Internet problems, protocols, incompatibility, network down, and lack of knowledge and skill on the part of respondents."

The Future of DBM

According to those surveyed in 1998, the DBM data collection method does not appear headed for extinction any time soon. Researchers indicate that they plan on using DBM with about the same frequency three years from now as they have during the past 12 months. However, three years from now, they also expect to be doing considerably more research online than they have in the past 12 months (see tables below).

Mean % of Research Projects
Using DBM

Past 12 Months	Three Years From Now
(n = 19)	(n = 19)
12%	13%

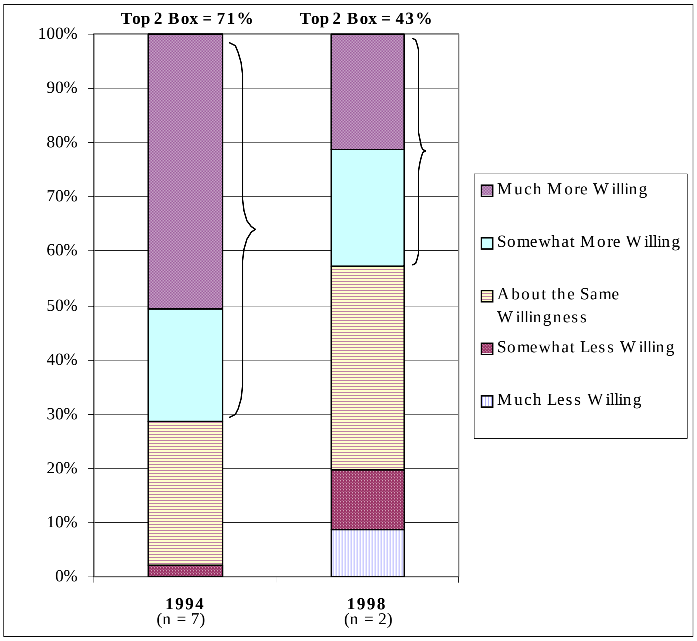
Mean % of Research Projects
Using Online Methods

Past 12 Months	Three Years From Now
(n = 19)	(n = 19)
32%	46%

Several agencies in both waves reported asking whether the respondents for their studies had ever taken part in DBM studies in the past. In 1994, about 1/3 of the respondents in the subject studies had participated in a DBM previously; in 1998, 47% had participated.

Although not many agencies ask "future participation willingness" questions, the few that do show that the willingness to participate is still high overall. The level of strong willingness (Top 2 Box = % Much More or Somewhat More Willing), however, has dropped from 71% in 1994 to 43% in 1998.

Willingness to Participate in Future DBM Studies



CONCLUSIONS

Due to the rapid increase of Internet usage among the general population, online research is clearly the wave of the future according to the agencies polled in this study. However, it is apparent that DBM still has its place, at least in the short run. This will be true as long as there remains a gap between PC users and online users. However, this gap appears to be narrowing, as virtually every new computer purchased today comes Internet-ready. Thus, until PC usage and Internet usage are close to the same level, there will be a need for the DBM survey technique allowing for the broadest possible sample pool.

It is important to note that DBM and online research are not mutually exclusive techniques. They can be used in conjunction with one another and often are.

As several respondents indicated, the online survey technology is still in its early stages of development. Until there is a widely available, fully functional software package, online research will remain limited to those research operations that have the technical capacity to create online surveys that incorporate advanced logic and control functions including:

- screening and quota control
- enforced terminates
- true skip patterns
- complex graphic stimuli

- full randomization of lists and attributes
- real-time error checking
- constant sum and allocation verification
- resumption of mid-term or paused surveys, and
- detection/rejection of multiple survey submissions.

It should also be noted that although DBM will be with us over the next few years, the existence of the floppy disk appears in jeopardy over the long run. For example, some computers are now being built (e.g. Apple’s iMac) without floppy drives. The need for floppy disks has been severely diminished by new technologies such as servers, e-mail attachments, and of course, the Internet. Floppy disks were once the primary way to install software. However, software may now be installed much more quickly via CD ROM or simply by downloading from a manufacturer’s Web site.

Researcher Ian Smith, Socratic Technologies, contributed to the administration of this study and the preparation of this paper.

WHAT WILL WORK OVER THERE? COMPUTER-BASED QUESTIONNAIRES IN FOREIGN LANGUAGES.

Brent Soo Hoo and Lori Heckmann
Griggs-Anderson Research

This paper will discuss the real-world experiences of researchers at Griggs-Anderson Research in international locations using computer-based foreign language interviewing systems for disk-by-mail, computer-aided personal interviewing and central location test applications.[1] The intent is to communicate procedures that have worked and communicate those that have not worked in our international experience.

CONJOINT SYSTEMS

The conjoint systems that will work in foreign countries are entirely dependent upon which language will be used. In the case of Anglo/European-based alphabets, this allows almost any conjoint program to be used. This is because the characters are from the same base ASCII set of codes and character descriptions. The accented characters are available by finding out which character set or code page is being used and then typing the appropriate ALT + character code in MS-DOS (see MS-DOS User's guide).[2] Since most conjoint systems are DOS-based, this procedure will work for many countries. However, it will not work in countries where the character set is double-byte or "multistroke." Here, there is a lot of work yet to be done to develop tools which will work with these languages. The biggest hurdle is the testing and development of the actual programs to be used. If programs are going to be developed for a certain country, then the software must be tested on the same systems that will be used in-country. This will allow the inevitable problems relating to display of the characters to be fixed.

At Griggs-Anderson Research, we've used Sawtooth's ACA (versions 3.1 and 4.0), Sawtooth's CVA (version 2.0, in conjunction with Sawtooth's Ci3 interviewing system), and custom programs in foreign countries. Other Sawtooth programs like CBC (version 1.2) have been adjusted to work with double-byte characters but have not yet been used in the field.

Custom programs require two levels of technical capability. The first level is to create a program that works and collects the necessary data to do the final analysis. The second level is to be able to make this work in the foreign operating system and with the specific character set necessary for the application. Because of this bi-level technical capability, there are few companies that can do both. In an article on software localization in *Computing Japan* magazine, a localization expert from JD Edwards (a large U.S. financial software company) said, "Any time you move into double-byte languages you have to address a whole new set of issues. We have had to be very careful, even with our internal tools, to keep the capabilities from forking and ending up with two different tools."[3] This may change with the added capabilities of Windows 98, Internet

[1] Disk-By-Mail (DBM), computer-aided personal interviewing (CAPI), central location test (CLT).

[2] The MS-DOS manual gives an entire chapter to this information. See the *User's Guide, Microsoft MS-DOS 6.22*. Microsoft Press, 1994. Document MS58574-0594. Chapter 9, pp. 205–220.

[3] Yeaw, Coleman. "Meeting the Challenges of Software Localization," *Computing Japan*, June 1996, p.22.

browser-based tools and more open systems. There are many developments in this area and it will take much timely research to keep up with the technology.

PRACTICES: DBM VERSUS CAPI VERSUS CLT

Since the majority of disk-by-mail, conjoint and computer-aided personal interviewing software is DOS-based, the lowest common denominator needs to be addressed when preparing for international research. This means that a return to DOS-based systems may be necessary. The latest version of standalone MS-DOS is version 6.22.[4] This is also the last and final version of a standalone MS-DOS, as MS-DOS functions have been built into the Windows 95 and Windows 98 operating systems. Because of the technology adoption curve, there will be instances in foreign countries where MS-DOS is still regularly used because of operating system needs or computer configuration issues. In some countries, the technology adoption curve is slowed by the availability of affordable, high-performance equipment (processors, CPUs and RAM). Without the necessary high-end processing power, the users of a particular country may have no other option than lower-end operating systems. Griggs-Anderson Research is still programming some foreign language data collection instruments using MS-DOS (and Windows 3.X) until those parts of the world move to Windows 95 or Windows 98 and use interviewing tools that can utilize all the functions in these operating systems. Using a Windows-based tool demands the country has the computer configuration to allow the use of a higher-end tool. DBM as a procedure is problematic in certain foreign countries because of the technology adoption curve for computer operating systems.

The lack of standards worldwide is a problem which makes any solution a creative one. Here are some examples of situations we have experienced. In China, there is no edition of MS-DOS 6.22 in a traditional Chinese format. It is necessary to find archaic DOS kernel utilities in order to have MS-DOS capabilities in traditional Chinese. This is still problematic, as these DOS kernel utilities are somewhat unstable and hard to work with. Also, the *type* of Chinese used is entirely dependent on *where* the research will take place, as there are two forms of the Chinese language on Intel processor-based computers. One form is traditional Chinese which is used in the Western influenced countries such as Taiwan and Hong Kong. The other is simplified Chinese which is used in the inland Communist-developed areas. The right one must be chosen for the location.

Any language that has never been integrated into a data collection instrument needs to go through an information collection process. The following are specific examples of what information is crucial to know before advancing to the next steps. In our exploration of Hebrew for a potential project, we found out that there is no MS-DOS 6.22 available for Hebrew. The programmers we talked to at the company that sells the Hebrew operating systems also indicate that the Hebrew systems are based on a Fat 16 bit and will not work on a 32 bit system. This calls for a separate partition on the testing PCs, another good use for the System Commander software made by V Communications, Inc. When doing the background research on India, we discovered

[4] See the *User's Guide, Microsoft MS-DOS 6.22*. Microsoft Press, 1994. Document MS58574-0594. Recent discussion with the Microsoft Press noted that there are only two books currently in print and available on MS-DOS. *Step by Step MS-DOS* (ISBN #1-55615-635-9) and *Running MS-DOS* (ISBN#1-55615-633-2).

that there are 15 to 20 official languages.[5] India is a large and populous place! Further consultation with data collection partners in India provided the information that doing a high-end consumer goods computer interview in India would involve programming in seven languages to get national coverage. Our next step in the India background research process led us to a software developer who informed us that there are no ASCII standards in Hindi and that each individual Indian language has its own special fonts. The same software developer told us that 42 different language standards exist in Vietnam!

DOUBLE-BYTE CHARACTER SYSTEMS

There are no better books available on East Asian computing than those written by Ken Lunde.[6] These books give a technical and historical tour de force on the development of computing in the Far East. As far as we know, there are no similar resources for the rest of the world. Lunde gives the reader a basis to understand how computer systems handle characters and text. If applied to other countries, the knowledge has some transfer. The reason for this is that the computer has to handle the character and text somehow. While the Cyrillic and Polish alphabet may not technically be double-byte, they handle additional characters is as if they were. They will not work on the standard U.S. operating system.

At the core of it all is the ASCII system that comes from the Western computer standards. This is the standard that ensures that Western (Anglo/European) computers can talk to each other. The ASCII character set contains 94 printable characters, which is 42 characters more than the uppercase and lowercase letters (26 + 26=52 combined) in the English alphabet. In Japan, there are 1,945 characters in the non-electronic character set. When all the other printable characters are added to make the electronic character set, it totals 6,879 and is called JIS X 208-1990. This is the most updated version of this standard in Japan. Also, there are an additional 6,067 characters in an extended character set called JIS X 0212-1990.[7] The way that a character is stored as a piece of computer information directly relates to the series of cells that holds the encoding information. In the English language, there are limited cells in the encoding of data because the standard ASCII character set is only 94 characters. Even the extended ASCII character set only contains the encoding for 255 total characters. This is a miniscule amount when compared to the sheer bulk of information which is needed to encode the double-byte characters of Japan, China, Korea and other countries. We cannot do this topic justice in this short paper and recommend that any interested parties acquire the books by Lunde and refer to the information resources noted at the end of this paper.

[5] Here are a few examples of how many people speak various Indian languages. Hindi, the official Indian language is spoken by over 275 million people. Bengali, the language of the Bengal region is spoken by over 190 million people in India and Bangladesh. Tamil, a major language in southern India is spoken by approximately 60 million people. Telugu, a major language in southeastern India is spoken by approximately 70 million people. Marathi, a major language in western India is spoken by approximately 65 million people. Another fact to note is that only five languages in the world can claim as many as 190 million speakers. Information provided by an Indian language section at the World Language Resources Web site http://www.worldlanguage.com.

[6] Lunde, Ken (1993), *Understanding Japanese Information Processing*. O'Reilly & Associates, Inc., 1993. ISBN #1-56592-043-0 (out of print).
Lunde, Ken. *CJKV Information Processing (Chinese, Japanese, Korean & Vietnamese Computing)*, O'Reilly & Associates, Inc., 1998. ISBN # 1-56592-224-7. (Published 12/98, includes updated information from the first book).

[7] See Chapter One of Lunde's *Understanding Japanese Information Processing* for a complete description of the Japanese system, pp.1–17.

There is a great amount of market pressure to expand into large world markets such as Japan. Yeaw writes that, "Japan is the second largest market, after the U.S., for many software firms."[8] The sheer amounts of people in countries such as China, India and Brazil have fueled much interest by large corporations trying to sell more products. In order to sell products, research must be done first. In order to do research, programs to collect data are needed. Programming that will work in the English language has difficulty with the encoding present in double-byte applications. It simply is not built to handle that type of information. That is why specialized localization programming firms exist. It is one thing to be a programmer; it is another thing to have the language expertise, and yet another to have the development tools available. Why does this present a problem for data collection/interviewing computer programs? These tools, with a few exceptions, were developed in the Western ASCII standard. They have a hard time with the extra information inherent in double-byte systems. Until tools are developed in these countries or by native speakers, there will be a lack of sophisticated data collection/interviewing programs. There are some tools that do exist already and they seem to originate from Pacific Rim software developers and research organizations who have an eye on the "Tigers of Asia" as an important future market.

TRANSLATION ISSUES AND BEST PRACTICES

Translation usually originates from the data collection partner organization. Our company sends a final version of the questionnaire and terminology in English to the data collection partner. In a few days, a foreign language version of the questionnaire is finished. We usually ask for this in both a word processing program (such as Microsoft Word) and in text-only form (if possible). We ask for these files to be both e-mailed and physically sent on floppy disk. At the same time that the final questionnaire is finished, it is sent into programming in English. An English language version is the basis for all subsequent programming efforts. When the foreign language files come back, we use the foreign language PCs and appropriate foreign language word processing and line and/or text editing software to finalize the language portion of the survey instrument. It generally takes a week to translate English to another language. When it is done, the beta version of the program is sent via e-mail and also on disk to the country where it will be used. The data collection organization then tests the program for accuracy of translation. At the same time, a foreign language interviewer at our offices will also test the instrument. Feedback from both the data collection partner and the in-house foreign language interviewer is collected and the programs are updated. Final versions are completed and tested. The final version is tested one final time and the data are used as test data to see if the instrument is behaving properly in the data files.

Triple translation generally occurs upon client request. In which case, a final version of the survey instrument is sent on disk to the client. They test it with their own foreign language staff for accuracy of translation and usability and then provide final feedback to us. There has been much more written on the topic of world markets, translation and localization. When viewing the amount of international work being done at Griggs-Anderson Research, which has grown to over 30% currently, no one can argue that the future is in the international marketplace.

[8] Yeaw Coleman, "Meeting the Challenges of Software Localization," *Computing Japan*, June 1996, p.19.

TESTING

Testing is one of the biggest issues which ensures success. It can almost be said that no amount of testing is enough. With the increasing complexity of the survey instruments, this becomes critical for checking proper branching and skip patterns. We do a lot of testing on the English language instrument and make sure that the questionnaire is final before moving into the integration of the foreign languages.

The way we do testing is to set up a separate PC (or hard disk/hard disk partition) for each country from which we are going to collect data. This is where the luxury of swappable hard drives and the multi-boot capability of V Communications' System Commander come into play.[9] With the help of the IS department we set up the computers to run the exact operating systems which will be encountered in the countries being researched. Loading the word processing program and line editor that work in the foreign language is an added advantage to development. This way, the testing and development can be done on the same machine. It takes a while to get the foreign operating systems set up and it definitely helps to have a native speaker available to read the install messages and possible errors. Once this is done, great care must be taken to properly back up the foreign configured machines. Another hot tip is to connect a laser printer locally so that the machine will have the ability to print. Getting these foreign machines on the network is an option, but it involves more complexity. Generally, we leave them as standalone machines connected locally to laser printers. The modern laser printers have network ports as well as local ports and the network port usually carries the traffic in the larger companies.

Test disks are created once the English programming is finalized. This testing is done in-house by numerous people, other than the programming team. If the programmers become too familiar with the program, they are not able to test as effectively. This is not to say that the programming team does not test their work, it simply means that having another perspective and different sets of eyes to test the procedure improves its overall accuracy. All errors are corrected, all edits are made and the final English version is created.

At this point the foreign translation is introduced. The translation is integrated into the program and over the course of about a week's time, a final foreign version is available for testing. It is sent to the foreign countries for testing and is concurrently tested by our native-speaking, in-house personnel. All comments, fixes and changes are compiled and the program is updated one last time. In a perfect world, this procedure would only happen once, but iterations take more updating and tweaking than one might think. Another potential problem is that foreign languages can be more verbose and take many more words than English does—or as with German, the words themselves are longer. When this occurs, the space a sentence takes may be longer, paragraphs grow and the programming syntax and structure is impaired, possibly to the point of program failure or bugs. The other thing to consider is that double-byte characters take up *twice* the amount of space. Where there are 25 lines by 80 columns in English, there are more like 25 lines by 40 columns in a double-byte language. This causes more questions and screens to be in a program. The only way to overcome this is to allow plenty of extra space when programming the English version. Experience says cramming as much as possible onto one screen is not as effective as making three screens to allow for more lines of double-byte text.

[9] System Commander was given the *Computing Japan* Editors' Seal of Approval in the June 1996 issue of *Computing Japan*. Editors. "The Help Desk; Installing Win95E and Win95J: A Better Solution," *Computing Japan*, June 1996, p.45.

The testing personnel are then told what case IDs to use. This helps to identify the test cases and also helps with debugging. If certain case IDs are assigned a testing structure, the different branches and skip patterns of a large instrument can be properly tested. The more complex the instruments are, the more standardized testing procedures become necessary. The instrument must be as close to perfect as possible as the data depends on it. All calculations and arithmetic are then checked for accuracy. When the instrument is sent overseas we assure that the data collection partners have been assigned test case IDs. Invariably there are occasions when the program or computer did something unpredictable and the foreign interviewers want to test the program to see if they can identify or replicate the problem. In this case, some test data may come back in the final data. If these cases cannot be identified in these cases, the result is a corrupted sample. Additionally, we always collect a name field in the instrument. If the interview is a test interview, it can be noted as such in the name field.

Potential problems are sometimes noticed in the matching of qualification (recruit) screener data and the survey instrument data. These two instruments are often administered at separate times. The same case ID must be assigned to both the screener and the survey to match up the two data records. This is where a data collection partner that is very detail-oriented is a big plus. If any cases are not matched, they cannot be used. Good record keeping on the part of the field staff is necessary to ensure that all cases match the subsequent survey data. We generally negotiate with the data collection organization to agree that payment will be given only for the correct amount of *matching* cases. This type of record keeping also pays other dividends in the possible need for callbacks or clarification. If further questions are necessary, the data collection partner is then able to quickly find out who a particular respondent is according to what a respondent's case ID is. This usually comes up in the data cleaning phase and it is important to resolve these data problems so that the final data set can be cut.

Another item to consider is the issue of oversampling. It is preferable to oversample by five to ten cases in each major quota group to ensure enough clean cases after the data is initially cleaned and matched. The oversample must be balanced by the cost of each interview and also by how much it would cost to redo an interview after the data collection phase is finished. Having an oversample saves time and money when nonqualifying cases have to be discarded.

DATA TRANSMISSION AND BACKUP

Protection and backup procedures save a lot of time and help to avoid corrupt data files. If a procedure is in place, the data are always properly backed up when being transferred from one location to another. The file names of the data files are always identified for the data collection personnel. Additionally, they are also cautioned to not try to open or edit these files to avoid the problem of data file corruption. The data collection partners are told that they need to rely on the home office to crack the data and provide completion reports. We use the same procedures for both domestic and foreign jobs. These procedures have become the company standard when doing jobs of this type.

We have written a procedure where each computer that is being used for data collection has its data files backed up on a nightly basis. A separate set of backup disks is used for each day in the field. In addition, each night the same day's worth of backup files are uploaded to a BBS system at our main office. An alternative is to "zip" or compress the files and e-mail them back to the main office. Neither transmission method is infallible. There is sometimes a need to ask for retransmission in order to get a clean file. At the main office, a data analyst checks the uploaded data files every morning. A report is generated on completes and incompletes and a listing of case IDs (and sometimes the name field as well) is generated. This is checked against the recruited amounts. This becomes more complicated when working with a foreign data collection partner. Still, the amount of completes and incompletes can be checked to verify numbers.

The actual backup disks are either sent to the home office at the end of the data collection or sent on a daily or weekly basis. The schedule of shipping the disks is dependent on how long the project is in the field. There may be the need to refer to these backup disks in the event of a data problem.

A test of this procedure is done to get the test data from the foreign data collection partner to the home office. This serves a twofold purpose. First, it tests the actual backup procedure and familiarizes the data collection staff with the process. Second, it provides a set of test data to start the procedures in the data processing department. With this set of test data, the initial data export procedures and specifications are then worked on.

COMPUTER CONFIGURATIONS

The knowledge of what types of systems and configurations are used in a particular country will lead to the success of foreign projects. This section will present historical context using Japan as an example of a country with unique needs to illustrate the reasons why the computer configuration is important. In Japan, there used to be an issue concerning a different operating system which was incompatible with Western “Wintel” operating systems.[10] This is/was the NEC computer and operating system. According to data that was compiled in 1998 by our Japanese research partner Nikkei Research, it appears that this platform and operating system have been supplanted by the increasing dominance of the “Wintel” systems. According to data supplied by Nikkei Research the following is the breakdown of home and workplace operating systems currently used in Japan.

Operating System	Home Percentage (circa 5/98)	Work Percentage (circa 4/98)
MS DOS	4.2%	4.0%
MS Win 3.X	2.6%	12.9%
MS Win95	86.6%	75.0%
MS WinNT4.0	1.6%	6.6%
Apple Macintosh	4.0%	0.5%
Other OS	0.4%	1.0%
NA	0.6%	0.0%

According to Nikkei Research, a percentage of 17%–20% of Japanese households currently have computers. The majority of those computers are running MS Win95 operating system. It

[10] Wintel is defined as the Microsoft Windows operating system running on an Intel or Intel-compatible processor.

would seem that Microsoft has made great strides to capture the powerful Japanese market since the old days of IBM DOS/V.[11]

Because of the popularity of disk-by-mail procedures, it is important to know if households can run the programs. Before deciding on a methodology, knowing the breakdown of operating systems and home/work percentages is important.

Seventy-five percent of Japanese home computers are desktop types with the remaining 25% being laptops. The Japanese workplace is more evenly split; 52% are desktops and 48% are laptops. There is an amount of computer sharing in the Japanese workplace, but there are no hard figures on this practice. Nikkei reports, “We cannot tell you an exact percentage, but generally companies don’t have a 1:1 ratio of computers to employees in Japan.”[12]

These percentages are shown as an example of what occurs in foreign countries. It will be different in every country. Effort should be made to get data of this type when planning a foreign, computerized data collection project. This drives what platforms are needed to target for the maximum coverage. This also demonstrates how a disk-by-mail study can go wrong if the programmed disk is incompatible with the predominant computer platform or operating system in the foreign country.

TIMELINES

In our experience, foreign integration projects take a great deal of time. In terms of steps, nothing can move forward until the English language version is finalized. By this we mean that the English version of the survey instrument be completely tested and fully debugged on all branches and skip patterns. Concurrently, the translators translate the final English language version of the instrument into the foreign language(s). These translations are back-checked for accuracy and understandability by our in-house, native-speaking staff. Once a final, translated version is agreed upon, then the process of foreign language integration begins. This process takes a week or two. In this period of time, the version is finalized to the point of testing. It is tested by both the in-house, native-speaking staff and also sent to the country where it will be used. Then the data collection partner checks the foreign language instrument for accuracy and usability. When all parties agree that the instrument is as final as it can be, it is ready to collect data in the field. There can be no compromises in the quality of the instrument, as it is the crucial part in the collection of top-quality and accurate data. Careful thought goes into forming schedules that can be kept. One thing to note is that on projects of this type, we accrue approximately one-and-a-half to two times the budgeted amount of hours for this process. At times, we arrive at a budget and timeline figure and double it to be on the conservative side. The complexity also increases when dealing with multiple countries. We try to approach them sequentially, finishing one instrument completely before moving on to the next. An alternative here is to split the programming team so that multiple countries can be programmed at the same time. This requires good technical understanding and capability from the programming staff. They should know exactly what they are doing and communicate any deviations to the rest of the programming department as this impacts the data output.

[11] IBM DOS/V was the initial system from IBM that allowed Japanese computers to work with the Japanese character set. It used IBM Japan’s own front end processor IBMMKK.

[12] Quote from Nikkei Research when asked the question “Do Japanese workplaces have a higher percentage of 1 computer per user, or do individuals share computers?” 12/98.

Activity	Timeframe
Client approves questionnaire (English version)	Approximately three to four weeks after project begins
Questionnaire translated	One week
Translation checked	One week
Programming start to completion	Two to three weeks
Program testing (in-house)	One week (each language)
Cut and paste translation	One week (each language
Program testing (in-country)	One week
Final changes	One week
Data collection	Two to three weeks
Data extraction and processing	Two to three weeks

RESPONDENT CONFIDENTIALITY

There are laws or standards that need to be examined on a per country basis. In Germany and Japan, we have run into strict laws and/or market research industry standards that add a level of complexity to the extraction of the data and the administration of the possible callback procedures. In which case, the data collection partners are asked about the rules on respondent confidentiality. For Germany and Japan, the data collection partners replace the respondent name and/or company name with a code number to protect respondents' identity. A list of these names and numbers is kept in their offices. When the need arises to recontact a respondent to clarify an answer or ask additional questions, the code number is used to identify which respondent to recontact.

IN-COUNTRY EXPERIENCES

One thing to be avoided at all costs is data collection agencies opening files that they should not be in. To prevent this, we explicitly tell them what files to leave alone. Since they are focused on getting completed interviews, this is a big temptation. We ensure that they follow the procedure of collecting and forwarding the files so that the files can be opened with the proper tools. We had one experience where a field facility opened a binary data file with a text editor and destroyed the format. The file was saved by much data massaging and the necessity of writing a custom application to properly extract the data from the file. But what happened is a good example of how important this step is.

Last-minute translation challenges are one of the hardest things to overcome. There is no substitute for proper scheduling to allow for enough time to tune and fine-tune the translated survey instrument. Because of the multiple procedures involved in updating the questionnaire and changing the translations, the necessary time must be budgeted. The programmer is brought on-site to complete and fine-tune the instrument with the foreign staff. If enough time isn't allowed, it is necessary to have the proper editing tools available to work in the languages and countries in which the data is being collected. These tools include a good word processing program for that language and a line editor which works in that language. Another method of dealing with a limited timeframe, is to have a team of programmers available, one "in-country" and others at the home office. This enables work to get done (depending on the time zone difference) while each part of the team sleeps. In order to do this, strict rules on versions and who has the most updated copy are paramount. It is necessary for the in-country programmer to do all that is possible before retiring for the night. When finished, the programmer e-mails or downloads a version of the program code to the team at the home office. While the in-country programmer sleeps, the home office programmers fix and refine the program code. As they make sections final, they compile the program to be sure that things work properly. At the end of their work day, they e-mail or upload a new, final version to the in-country programmer. By morning local time, the "in-country" programmer has a new version of the program code and the application is compiled and finished. There is a big trust factor involved in this type of code or file sharing. It is important when doing this type of programming to make liberal use of the "comment" function inside most programming environments. Without these comments, either side of the programming team can be uncertain as to what functions certain parts of the code are to perform.

When the translated survey instrument works as it should and respondents understand what they are being asked, it is a beautiful thing. As a standard procedure, we have the interviewers and staff administering the instrument actually go through the whole questionnaire just as a respondent would. That way, they know what the error messages are, they know what kind of input is necessary for the questions being asked, and they become familiar with the timing and sections or breaks. They will know when to show the new products or when to introduce new concepts. When doing the briefing, the researcher (and/or programmer) need(s) to show the staff how to handle questions and possible problems where data is being collected. A crucial step in the data collection process is an in-person, comprehensive briefing with the interviewers and data collection staff in the country where data is being collected. Cultural and communication differences make this type of information exchange hard to do any other way. The briefing also allows the quality of the interviewer and the quality of the data collection partner's organization and organizational skills to be gauged.

DATA PROCESSING

A discussion of the data processing, extraction and analysis is important. Vital decisions are made at the time of programming and at the time of extraction—for example, how to treat the different countries and the dissimilarities in questionnaires. We decide if the instrument can be housed and programmed in one big instrument that only changes in certain parts for the foreign applications, or if each country has to have a separate instrument that should be programmed independently. Once made, the decision is communicated to the data analysts who determine if the data is to be merged or if a cross tabulation is to be done on a grand total data set. If the data sets from individual countries will never be merged, then the approach of programming separate instruments will work. However, if a huge, international, merged data file is the objective, then the approach of programming a master instrument deserves some serious thought. In a merged international data file, the data map has separate locations for each specific country and allows the use of a single data file. Creating separate data locations for each country streamlines the process for the data processing team and allows tables to be built which can be broken down on a per-country basis while still showing the overall grand total of responses. Any time that the sheer number of necessary data files is decreased, it is easier to work on the data set.

Naming conventions for questions and variables is important. There are times when we ask the same question in each country, but depending on the format (master instrument versus individual country instrument) care is taken to ensure the names of questions and variables are consistently written in the data files. Thought is required on how to handle differing response lists for questions based on each country when the data is collected.

OPEN-ENDED RESPONSES

Open-ended responses present an interesting issue in the field of foreign data collection programming. The potential problem is based on the entry systems used in a particular country. Because of the multiple strokes necessary to type or "draw" a character on-screen, there are multiple standards involved in foreign countries. In Japan, there are two basic types of input methodologies. The first is a direct method which uses the input of the actual encoding value of the target character. The second is an indirect method which usually involves typing out the pronunciation of the characters on the keyboard. The user selects the character which matches the intended usage of the character in the word or sentence. This method gives the user choices on which characters are the closest matches. This is the same situation experienced with the entry systems of China.

That particular method used impacts the entry of open-ended fields. If a computer is set up to allow text entry in one method, but the respondent cannot type in that method, it is a problem. Open-ends and "other, specify" types of questions should be used and tested with great care. An additional issue is that with multiple strokes, input can actually take longer to do than it can in English language situations. A solution that we use a lot is to set up an answer booklet or a form that has blanks for the respondent to write in their "other" or open-ended responses. These responses are then translated and upon translation, the data collection partner provides the files and responses in electronic form. These files are merged into the final data set. The benefit for the respondent is that they are able to write their answers much more quickly. The benefit for the data collection partner is that it removes the step of having to compile an open-ended file. This quickens the procedure as no file has to originate from the electronic data to be sent to the foreign location. Instead, the paper responses are compiled in-country and an electronic file is sent to the home offices from the foreign country.

INCENTIVES AND COOPERATION FEES

The incentives that are paid in the United States may not be sufficient to cover similar projects overseas. In-country data collection partners can provide good ideas on what amounts or incentives are appropriate. The culture of some locales dictates different interviewing procedures which vary from what works in the U.S. and in Europe. The base interview can take longer in a foreign language as some concepts are harder to convey. If interview length increases, then the incentive needs to follow.

CLIENT/SPONSOR CONFIDENTIALITY

The reason so much effort goes into localization of the data collection instrument is to increase the appearance that it is a native product rather than one from outside the country. In most projects, client sponsorship is kept confidential to decrease any demand bias.

DATA EXTRACTION PROBLEMS

In one research project in Japan, China and Korea, we noticed a problem with the Ci3 LIST function which had something to do with the double-byte characters. We received files back from the data collection partners and found that the files were corrupt. We thought that there was some kind of error in transmission, so we looked more closely at the disks that had been sent back (this is why there are two steps in the back up and data transmission standards). Once again, we started to analyze the data. It was still corrupt. Our programming staff frantically tried to sort out the data problem, but could not recreate the error. Finally, the programmers looked at the raw binary data files and compared them to a set of binary test data that we generated at our office. The files were very similar. The programming staff arrived at the hypothesis that the double-byte characters had reacted in an unforeseen manner with the binary data structure. This corrupted some key characters in the files which made them virtually unreadable with the analysis tools provided. The fix involved a painstaking process of viewing and editing the affected files and trying to make sense of what could be read in the uncorrupted part of the file. This was a time- consuming process, but finally the data was successfully extracted. Needless to say, we now avoid the use of the Ci3 LIST function and try to use other methods to capture the data.

Another problem we encountered involved the import of CVA data from a double-byte binary Ci3 data file. For some reason the data extraction tool for the CVA program had difficulty converting out the data. It was believed that the tool had difficulty with the double-byte characters in the data file. After some testing, a CVA data export was done on just the CVA data in plain ASCII format. Then, this ASCII data was read into the CVA data import tool. This worked and is currently part of our procedures when using double-byte binary Ci3 data files with CVA data contained.

As noted earlier, it is imperative to use the proper tools to open, use, edit, extract the data. Problems occur when a binary data file is opened in Microsoft Word (or other commercial word processing programs). Changes to the file may not be seen, but the file becomes corrupted. It costs time and money to save and extract the data files. Suppliers must be told which files they can view, and which ones they shouldn’t. If there are any doubts, it is better to err on the side of not opening the files with any tools.

REFERENCES AND RESOURCES:

Whenever we are asked to find out about programming a new language we have not had experience with, our first stop is the Internet. Doing keyword searches on the particular language in a variety of terms can lead to very useful and informative information sources and companies. From there, the hardware, software, programming help and other information can be found that will make projects successful. No two foreign language integration projects are exactly alike. It is important to find out as much as possible about the foreign language that needs to be integrated. Both historical and programming information is needed on the operating systems, the fonts and character display, word processing programs, line editors and programming tools and utilities. The Internet has a wealth of knowledge on these topics for all languages.

User's Guide, Microsoft MS-DOS 6.22. Microsoft Press, 1994. Document MS58574-0594. Chapter 9 is the chapter on International usage.
Microsoft Order Center (800) 249-8314. Ask for information on foreign operating systems. (Do a keyword search on DOS.)
http://www.microsoft.com/.
http://www.msdn.com/.

Griggs-Anderson Research Web pages on international research
http://www.gar.com.
http://www.gar.com/intl.htm.
http://www.gar.com/primer/intlstrt.htm.

Nikkei Research, Japan. http://www.nikkei-r.co.jp/.

Assembling and Creating Foreign Language Web Materials (Web page)
http://www-japan.mit.edu/presentations/nflrc96/chars.html.

Walnut Creek CD ROM. (800) 786-9907. http://www.cdrom.com
"*East Asian Text Processing CD ROM*." Copyright June 1994.
"*Japanese Text Processing CD ROM*." Copyright January 1996.

Computing Japan. ISSN 1340-7228. http://www.cjmag.co.jp/.

Qualitas Trading Company. (510) 848-8080. http://www.holonet.net/qualitas/.

Pacific Rim Connections. (650) 697-0911. http://www.pacrim.net/Welcome.html.

Sawtooth Software. (360) 681-2300. http://www.sawtoothsoftware.com.

Chase Computers. (949 626?) 308-5888. http://www.chasecom.com/.

ComStar Company. (408) 257-9480. http://www.gy.com.

V Communications, Inc. (408) 965-4000. http://www.v-com.com.

System Commander software (multi-boot systems)

Pacific HiTech Inc. (801) 261-1024. http://www.pht.com.
"FontAsia TextPro CD ROM"

World Language Resources. (310) 996-2300. http://www.worldlanguage.com.

Lunde, Ken. *Understanding Japanese Information Processing.* O'Reilly & Associates, Inc., 1993. ISBN #1-56592-043-0.

Lunde, Ken. *CJKV Information Processing (Chinese, Japanese, Korean & Vietnamese Computing).* O'Reilly & Associates, Inc., 1998. ISBN # 1-56592-224-7. http://www.oreilly.com/catalog/cjkvinfo/.

ACKNOWLEDGEMENTS

Aaron Tam for all his programming help and for speaking and writing Chinese.

Client Services Department at Griggs-Anderson Research for all the support for the foreign programming.

Kaz Terauchi and Lisa Kim at Nikkei Research, Japan, for all the experience and data on Japanese computing.

Chris King at Sawtooth Software for all the support on program testing/improvements.

Lane Webster for believing in the procedure.

COMMENT ON SOO HOO AND HECKMANN

Karlan J. Witt
IntelliQuest, Inc.

First, I'd like to agree with the authors on their recommendations. I actually commend them for documenting at a detailed level many of the problems that can occur, along with proposed solutions. The problems the authors describe are very real. At an overall level, I'd also like to thank them for raising the awareness of some very real issues when conducting market research studies internationally.

So armed with their recommendations, are we all ready to launch into new markets?

Well, some good news is, some of the barriers they cited are slowly diminishing.

Hardware and software adoption in emerging markets is increasing, and with the lowered price points of the newer technology, we are seeing some "leap frogging" activity, where buyers in emerging markets are not buying up from a 386 to a 486 processor, but to a Pentium, for instance. However, even when the technology is available, cultural attitudes towards electronic surveys may be slower to change.

One of the greatest barriers they cited was the lack of availability of survey software with multi-language capabilities. The good news is this is changing too, minimizing the type of intensive manual labor described in the paper undertaken only by the most committed researchers.

However, even with these advances, it is still very expensive, and the ROI may not be there for every market. Additionally, where the conventional wisdom counsels a researcher to create computer-aided interviews with the lowest common-denominator assumption of technology available, there is a recent complication to applying that advice. With Internet usage growing on a global basis, would-be respondents are faced with the most stimulating images that companies can create, and users' expectations of online applications is greater. A very plain vanilla interface intended to prevent exclusion of certain target audiences may now encourage non-response among more sophisticated audiences. A clear trade-off.

Finally, more on the good news front in the form of Unicode. This will aid in addressing the single byte/double byte language problem that the authors cite. Unicode technology will now allow Sequel Server, Oracle and Informix to manipulate and store data in multiple formats in one file by using a translation table referencing back to a universal code page. Commercially-available data collection packages utilizing this technology are lagging, although two or three proprietary packages are now on the market.

EFFICIENT FEE STRUCTURES FOR MUTUAL FUNDS

Ronald T. Wilcox*
Carnegie Mellon University

ABSTRACT

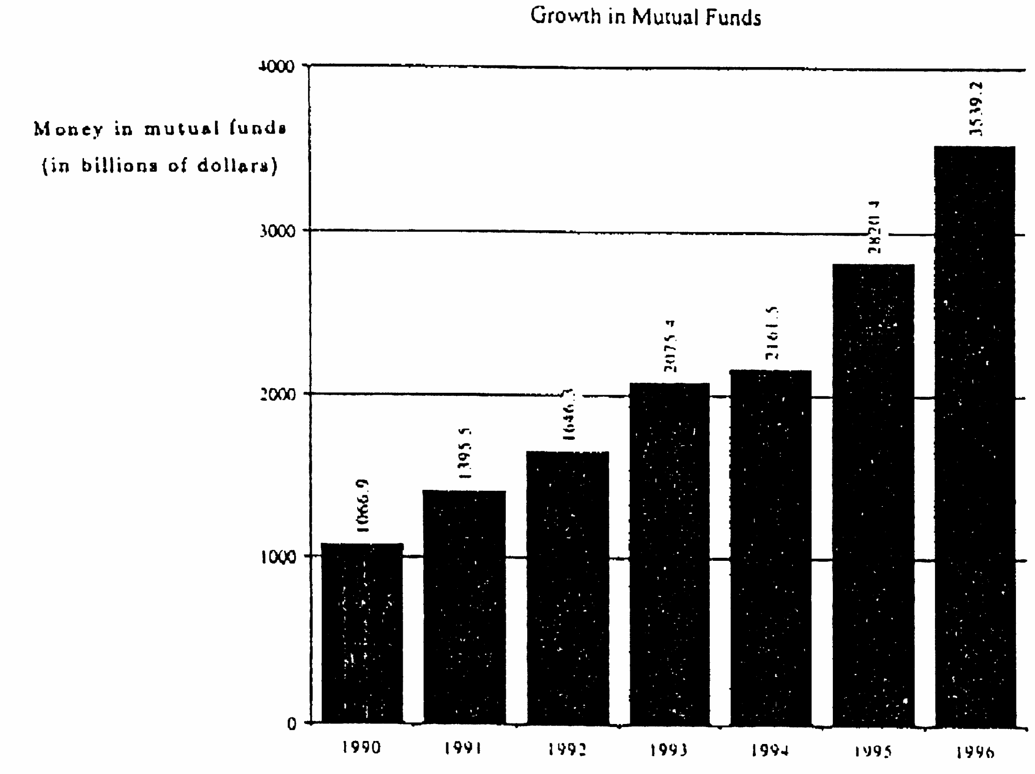
This research presents a management decision model for setting mutual fund fee structures. The model pairs information obtained from consumer conjoint experiments, designed to uncover consumers' preferences for different fee structures, with information on the expected revenue generated from various fee structures to suggest a set of "efficient" structures to mutual fund managers. Mutual fund companies can use this information to refine the fees they impose for their various funds. We also examine the data using latent-class choice-based conjoint analysis to uncover different groups with similar choice heuristics for the funds. We then use information gathered from pre-conjoint questionnaires to characterize the classes which arise in our model. Taken together, this information allows us to examine opportunities for price discrimination in this marketplace. The model is demonstrated using data collected from a cross-section of mutual fund investors.

1. INTRODUCTION

The 1990's have witnessed explosive growth in mutual fund assets under management as well as the number of publicly traded mutual funds. Against this backdrop of tremendous growth, mutual fund companies have spent ever increasing amounts of money marketing their funds. It is difficult to look at an issue of any popular business publication, listen to the radio, or watch television without seeing advertisements boasting the benefits a particular company's mutual funds. The competition in this industry has evolved from spirited to fierce over the last few years. Given the increased competition, marketing has evolved as a core functional area of these organizations. Traditional marketing activities such as channel selection, advertising, and pricing have been moved to the forefront of these organizations in their struggle for a piece of the seemingly ever-growing asset pie. As one industry executive put it

> It used to be that when a fund had good performance, the money would naturally flow into it, . . . Now even funds with unspectacular numbers are seeing substantial asset increases, which is a clear testimony to the power of marketing.
>
> -Marshall Front, Executive Vice-President, Stein Roe (Kihn 1996)

[*] Assistant Professor of Marketing, Graduate School of Industrial Administration, Carnegie Mellon University. We wish to thank Don Lehmann, Chakravarthi Narasimhan, and Christine Parlour for providing feedback on this manuscript. We also wish to thank the following organizations for providing support for this project; The Faculty Development Fund at Carnegie Mellon University, The Marketing Science Institute, and Sawtooth Software, Inc.



Source: 1997 Mutual Fund Factbook

Figure 1

Mutual fund companies make money not by holding positions in the financial instruments in which they trade, but rather by charging fees, which are tantamount to retail prices, to investors who choose to invest in their funds. By pooling together the money of many investors, mutual fund companies offer investors increased diversification and professional money management in return for these fees. By law, mutual fund companies must divulge the fees they charge to potential investors. These charges then form a basis of comparison for investors who are shopping for mutual funds. Beyond traditional measures of performance, investors can examine the increasingly complex fee structures imposed by these funds either directly from their prospectuses, through a number of well-read business publications[1], or increasingly through a variety of on-line services.[2]

Although the U.S. Security and Exchange Commission (SEC) is vested with the power to regulate the fee structures imposed by mutual funds, it in general does not.[3] Mutual fund managers, and the boards charged with their oversight, are left largely unfettered in determining what they will charge investors. These managers must carefully weigh the impact of pricing decisions

[1] The *Wall Street Journal* currently prints load information for mutual funds every day and expense ratio information once a week. Many other publications, including the well cited *Morningstar* mutual fund rating service provides consumers with a wealth of information about mutual fund fee structures.

[2] Some companies make a substantial amount of money simply by providing comparative information about mutual funds. Following what *Business Week* called the "Schwab Revolution" in 1994, the number of online services providing this information has increased dramatically.

[3] The one practical exception to this is the 1% cap the SEC imposes on distribution (12b-1) fees.

on both their ability to generate revenue from current and potential future customers as well as the impact such decisions will have on the fund selection process of these same investors.

Further complicating this issue is the fact that mutual fund managers have a series of retail prices, rather than a single price, at their disposal. Unlike traditional retail markets, the retail price of a mutual fund often involves a number of sub-prices or fees which the manager must set and potential investors must evaluate. A few of the more common types of fees are listed here:

- Front-End Loads – A commission or sales charge, typically expressed as a percent of the amount invested, paid at the time of purchase. Loads generally range from 0%, a "No-Load Fund", to as high as 8 or 9%.
- Back-End Loads – Similar to a front-end load except that the sales charge is deferred until the time the investor redeems fund shares.
- Operating Fees – A fee charged the fund to cover administrative costs as well as the fees paid to the investment manager. Usually expressed as an annual percentage of the fund's net asset value and called an "expense ratio", these fees are generally in the .25% - 2% range.
- 12b-1 Fees – Referring to rule 12b-1 of the U.S. Securities and Exchange Commission (SEC), these fees are a method for charging marketing and distribution related expenses directly to the fund's assets. Approximately half of all publicly available funds currently charge 12b-1 fees, and the SEC mandates that these fees must be no more than 1%.
- Account Maintenance Fees – An annual fee generally charged to customers who do not maintain a minimum balance in a given fund.

This list is hardly exhaustive, but gives the reader a notion of the often complex fee structures that are pervasive in the marketplace for mutual funds. Fee structures are by no means uniform across funds managed by the same company, or even across similar types of funds managed by different companies. For example, two "international equity" funds managed by different mutual fund companies may have very different fee structures.[4]

A clear understanding of the trade-offs that consumers make when evaluating the multiple components of the fee structure is critical for choosing fees that will maximize the profitability of the fund. The goal of this research is to lay out a framework for determining a set of "efficient" fee structures, fee structures which take into account both the revenue potential of a given set of fees as well as the utility consumers derive from such fees. Subsequent to this determination, we will explore ways to develop multiple efficient fee structures targeted towards different groups of consumers.

The rest of this paper is organized as follows. In the next section we review the previous literature on mutual funds paying particular attention to those researchers who have examined mutual fund fee structures. Section 2 briefly reviews the extant literature on mutual funds. Section 3 defines efficient fee structures and details our marketing decision model.
To demonstrate this model, Section 4 reports the results of a field study we conducted with

[4] As a specific example consider the "emerging markets" funds of Scudder and Vanguard. As of this writing, Scudder charged no load and operating expense of 2.00% on their fund while Vanguard charged a 1.5% front-end load, a 1% back-end load, and an operating expense fee of .60%.

current mutual fund investors. We derive a single set of efficient fee structures for the pooled data, as well as two sets of fee structures for the two latent classes of investors which arise in our model. Finally, Section 5 discusses the limitations of our modeling approach, summarizes the work, and points to new research directions in this area.

2. PREVIOUS RESEARCH ON MUTUAL FUNDS

Research on mutual funds has been confined almost entirely to the finance literature. The chief concern of this literature has been whether mutual fund performance is predictable (Lehmann and Modest 1987; Grinblatt and Titman 1989; Blake, Elton and Gruber 1993), persistent (Grinblatt and Titman 1992; Hendricks, Patel, and Zeckhauser 1993; Brown and Goetzmann 1995; Malkiel 1995; Elton, Gruber, and Blake 1996) as well as which metrics accurately capture fund performance (Murthi, Choi, and Desai 1997; Grinblatt and Titman 1993). While there are conflicting findings in the research cited, the overall flavor of this literature is that, if past performance does provide information about future performance, such information is weak at best.

The only extant research on mutual fund fee structures are the recent works of Chordia (1997), Christoffersen (1997), and Kihn (1996), as well as less recently Chance and Ferris (1987). Chordia uses a game theoretic model to argue that mutual funds charge loads in order to discourage redemption. Thus, funds whose investment strategy entails holding less cash, typically more aggressive funds, are more likely to charge loads than their more conservative counterparts. Christoffersen (1997) examines the conditions under which mutual fund managers waive some of the fees they had initially informed investors they would impose. Kihn (1996) studies 2,496 mutual funds to try to explain why some funds choose to charge front-end loads while others do not. One particularly interesting finding of both Christoffersen (1997) and Kihn (1997) is that mutual fund companies use different fees within the overall fee structure as substitutes. For instance, a greater load generally implies a lower expense ratio, which suggests that the fee structure is in large part a marketing decision, where management balances the revenue of different fee structures with the preferences of their target audience.

We take a different approach with this research. Unlike the previous research in this area, our goal is not to describe current market phenomena. Rather, we seek to improve the retail pricing decisions of mutual fund companies through constructing a management decision model. The decision model we propose takes into account the preferences consumers have for different mutual fund fee structures as well as the revenues derived from such fees. The output of this model is a set of “efficient” fee structures, fee structures which have the greatest potential for maximizing the profit of the managing firm.

3. A FRAMEWORK FOR DETERMINING FEE STRUCTURES

3.1. Defining Efficient Fee Structures

Many of the fee combinations that a mutual fund company can choose may generate equivalent or nearly equivalent expected revenue streams from any given consumer. It is important for the company to understand how each of the possible fee structures it might impose not only affects expected revenue from its current assets-under-management, but also how these different fee structures affect consumers’ utility for the fund itself. If one fee structure generates the same expected revenue as another, but offers consumers greater utility, it is clear that the company should select the one consumers prefer. Incorporating consumers' utility for various fee structures into the decision process enhances the prospects for increasing assets-under-management and hence long run profitability.

In this discussion we will make use of the following notation

p = the initial investment amount

r = the expected annual return on the investment p

i = the firm's internal rate of return

l = the front-end load charged by the firm, given as a percentage of the initial investment

f = the expense fee (ratio) charged by the firm, given as a percentage of the initial investment

F = the entire set of fees charged by the mutual fund

R = the expected revenue the firm generates, on a per customer basis, from imposing a given fee structure

U = a function which maps fee structures to consumer utility.

We define an efficient fee structure $\left(F^* \middle| R, p, r, i\right)$ as a fee structure such that there exists no other fee structure $\left(F' \middle| R', p, r, i\right)$ which jointly satisfies the properties $R' > R$ and $U(F') \geq U(F^*)$. In other words, an efficient fee structure implies that the firm cannot make more money from its current customers with another fee structure while maintaining these customers' utility at a level at least as high as that generated by the efficient structure.

In order to operationalize our definition we need to measure the expected revenue generated from any given set of fees, the utility for those same fees, and then combine those two pieces of information to form the efficient frontier. The next three subsections lay out a framework for accomplishing this.

3.2. Generating Revenue from Loads and Expense Ratios

While the set of possible fees, F, is very large in this marketplace, we model as if the fund makes a decision only on the load (front-end) and expense ratio to charge its investors. These two fees are the most common found, and are widely publicized in the popular press. The framework we develop here can easily be extended to include other types of fees under consideration by management.

The expected net present value of the revenue stream generated from managing some initial investment p for n years with a load and expense ratio structure (l, f) can be given by

$$E[R] = pl + \sum_{t=1}^{n} \frac{(p - pl)(1+r)^t (1-f)^{t-1} f}{(1+i)^t} \qquad (3.1)$$

As is evident from this expression, there are many potential fee structures which will lead to the same expected revenue per investor. For example, if the expected holding time of the mutual fund for any given investor is 5 years (*n* = 5), the expected market return 10% (*r* = .1), the internal rate of return 12% (*i* = .12), the expected initial investment $1000 (*p* = 1000), and the desired revenue $75 (*R* = 75) then Figure 2 depicts all load and expense ratios which will produce the desired revenue amount.[5] To develop efficient fees, this information must now be paired with consumer preferences. The methodology for measuring those preferences is now detailed.

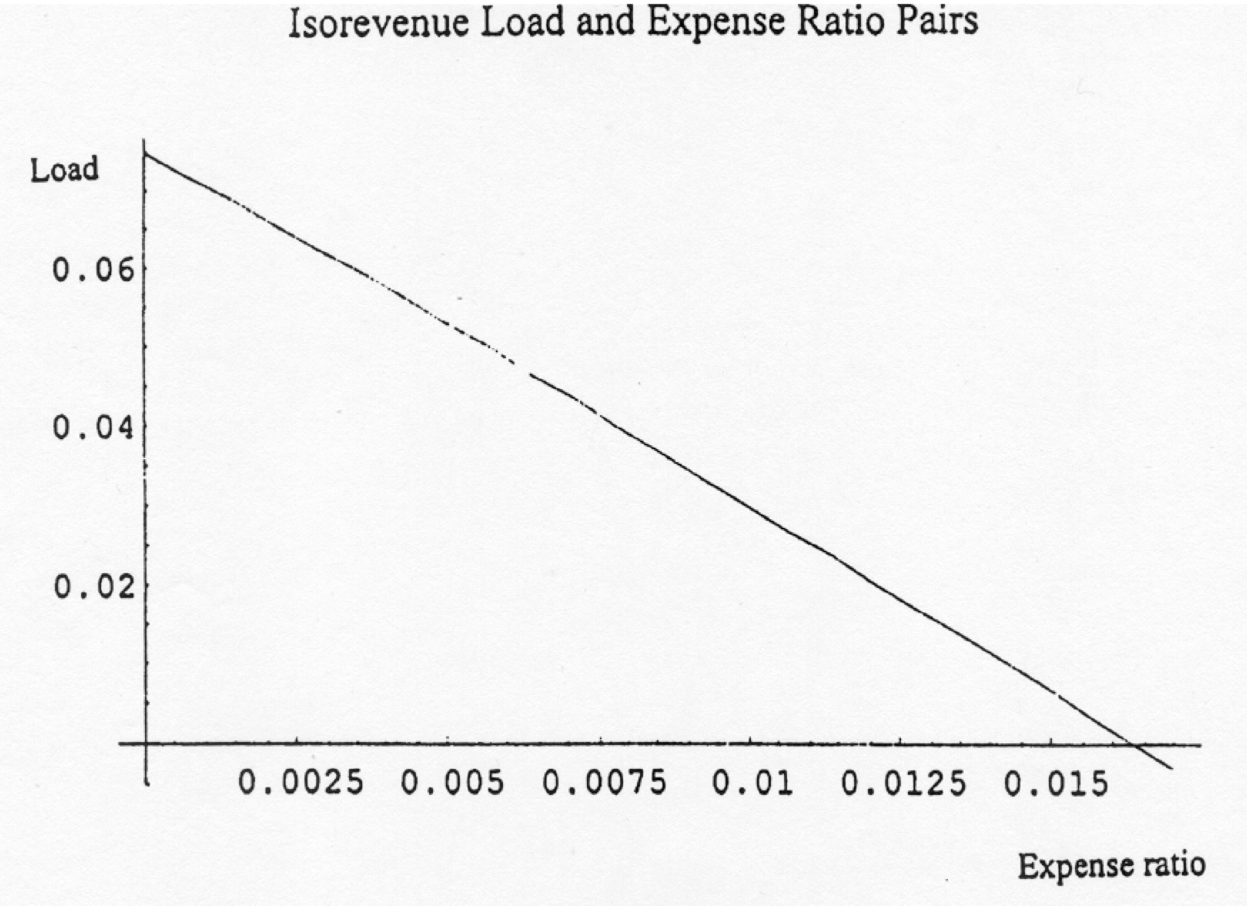


Figure 2

3.3. Measuring Consumers' Preferences for Different Fee Structures

Consumers derive different amounts of utility from different fee structures. While it would seem obvious that lower prices are preferred to higher prices, it is less obvious how consumers compare fee structures when one fee structure does not uniformly dominate another. For instance, would consumers rather pay a load of 1% and an expense ratio of .3% or no load and an expense ratio of .8%. While it is easy to solve for the fee structure that would minimize expected overall cost, given a planning horizon and expected market return, this information is of limited

[5] We have specified the revenue equation such that the firm considers the expected initial investment amount (*p*) and the target revenue amount (*R*). This equation can easily be reformulated to reflect the target revenue as a percent of the investment amount instead of an absolute dollar figure.

value to mutual fund managers. We find that many consumers' preferences for fee structures are inconsistent with strict expected cost minimization. Fund managers need to understand the actual choice processes of consumers to make reasonable pricing decisions.

We propose measuring the utility derived from different fee structures via choice-based conjoint analysis, where consumers are presented profiles of different stock mutual funds and asked to select their most preferred fund. Through observing consumers' choices we can estimate the utility generated from each fee structure. The particular type of conjoint analysis we propose employs a latent-class multinomial logit model to uncover utilities for each of the attribute levels.[6] We then can use simple linear interpolation to generate utilities for any possible load and expense ratio combination within the bounds of the most extreme loads and expenses tested.
For example, consider the following conjoint design intended to examine the utility of different load and expense ratio combinations. For simplicity, assume that there are only three attribute levels for each attribute. The general form of the conjoint output can be succinctly described by Table 1.

Table 1: Estimating Utility from Conjoint Output

	Load	Load Part-Worths	Expense Fee	Expense Fee Part-Worths
Levels	l_1	u_{l1}	f_1	u_{f1}
	l_2	u_{l2}	f_2	u_{f2}
	l_3	u_{l3}	f_3	u_{f3}

where

l_i = the value of the i^{th} tested load level

f_i = the value of the i^{th} tested expense ratio level

u_{li} = the utility, or part-worth, of the i^{th} tested load level

u_{fi} = the utility of the i^{th} tested expense ratio level

and the levels are ordered such that $l_1 < l_2 < l_3$ and $f_1 < f_2 < f_3$. Values for u_{li} and u_{fi} are estimated directly from the conjoint analysis. To generate the utility for any fee pair which lies within the bounds of the most extreme values tested we can compute.

$$U(l, f) = \underline{u_l} + \frac{l - \underline{l}}{\overline{l} - \underline{l}}(\overline{u_l} - \underline{u_l}) + \underline{u_f} + \frac{f - \underline{f}}{\overline{f} - \underline{f}}(\overline{u_f} - \underline{u_f}) \tag{3.2}$$

[6] We use Sawtooth Software's Latent-Class Choice Based Conjoint package to facilitate participant interviews and estimate part-worths. For a description of the algorithm used by the software to generate part-worths as well as latent-class memberships see DeSarbo, Ramaswamy and Cohen (1995).

where the upper and lower bars represent the attribute levels and estimated utility levels which lie immediately above and below the proposed fee pair respectively. Equation (3.2) is simply a linear interpolation between the relevant logit utilities. The precision of any utility estimate generated from this interpolation will necessarily be a decreasing function of the distance between each of the attribute levels tested.

We propose latent-class choice-based conjoint instead of aggregate choice-based conjoint for two reasons. First, recent research in conjoint methodology has indicated that latent-class models recapture the known data structure better than either aggregate models or other methods of segmentation (Vriens, Wedel, and Wilms 1996; Moore, Gray-Lee, and Louviere 1996; DeSarbo, Ramaswamy, Cohen 1995). This, combined with the research in incorporating heterogeneity in panel data models (Gönül and Srinivasan 1993; Chintagunta, Jain, and Vilcassim 1991) suggests that unbiased multinomial logit parameter estimates are obtained only when differences in consumers' choice processes are captured by the model. Second, uncovering the different latent classes in the data allows us to examine the antecedents of the different choice heuristics for mutual funds. Through demographic and other information collected from respondents, detailed in the next section, we can characterize the different classes of respondents and subsequently uncover opportunities for fee discrimination in these markets.

Following the findings of Moore, Myhta, and Pavia (1994) and Johnson and Pinnell (1995), we impose no *a priori* utility constraints on the conjoint estimation. Rather, we allow consumers' choices to completely determine the parameters of the model, even if these choices lead to ordered attribute preference reversals.[7]

3.4 Constructing Efficient Fee Structures

In order to generate a map of efficient load and fee structures, the preference information derived from the conjoint analysis must be paired with information on expected revenue. Essentially we are looking for the envelope of the tangency points between the iso-revenue functions tested and the utility functions derived from the conjoint. From these tangency points we can produce a map of the efficient load and expense ratio combinations for various desired revenue levels.

[7] We also estimated a model where the load and expense ratio utilities were constrained to follow the expected ordering. This model failed to uncover the data structure as well as the unconstrained model.

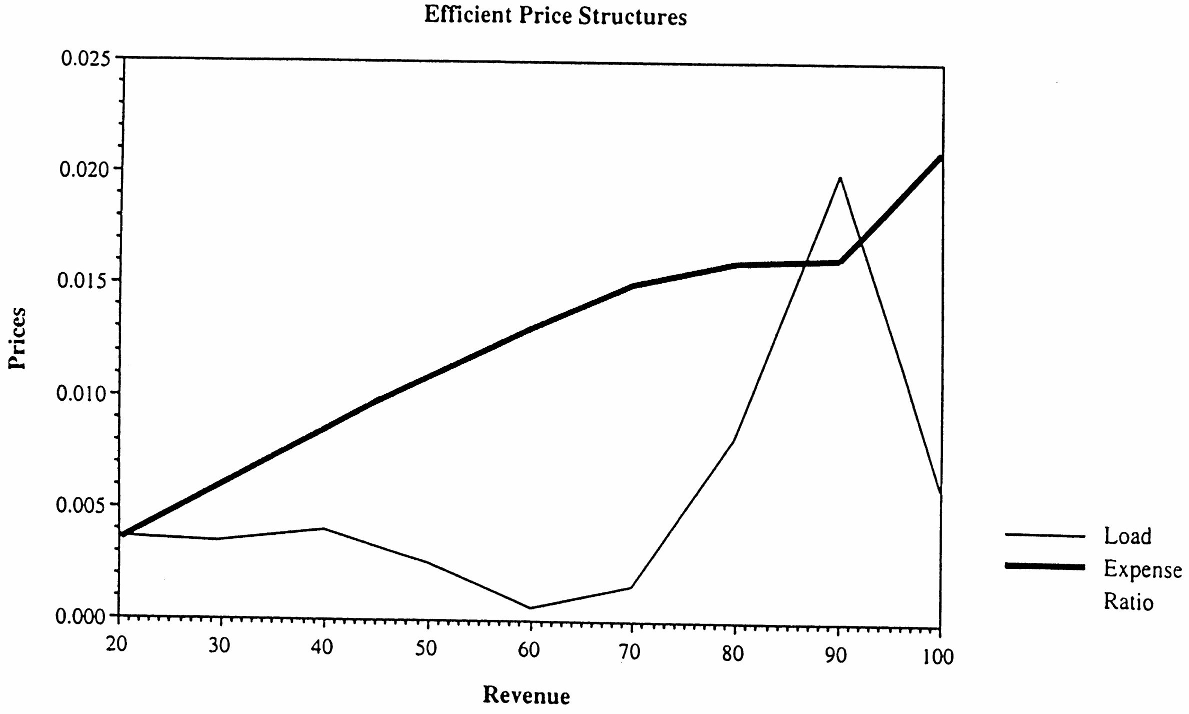


Figure 3

Figure 3 depicts a hypothetical solution to a problem of this type. It is a map which displays efficient load and expense ratio combinations for a wide range of possible revenue levels.

One increasingly common way to solve for these types of efficient frontiers is to use a technique called Data Envelopment Analysis (DEA). The problem at hand can be formulated as a single-input/single-output DEA problem. In the nomenclature of DEA, each load and expense ratio combination examined would constitute a Decision Making Unit (DMU), the revenue generated from each DMU the output, and the total utility cost (measured relative to the utility of $l = 0$ and $f = 0$) of each DMU, the input. We could then solve a set of appropriately specified linear programs to form the envelopment surface, or efficient frontier, and map the efficient frontier back to the load and expense ratio combinations that generated it. While this is not an unreasonable way to approach the problem, there are more direct ways to solve for efficient fee structures in this context. We present the most direct way, and hence what we believe is the most straightforward way, here.

Define $U^*(R)$ as function which maps each candidate revenue value R to the load and expense ratio pair which maximizes utility. To find $U^*(R)$ we solve the following maximization for successive value of $R_0 \in \left[0, \overline{R}\right]$, where $\overline{R}$ denotes that maximum amount of revenue that can be generated from a fee structure within the tested range.

$$U^*(R) \;=\; \max_{l,\, f} U(l, f) = \underline{u_l} + \frac{l - \underline{l}}{\overline{l} - \underline{l}}(\overline{u_l} - \underline{u_l}) + \underline{u_f} + \frac{f - \underline{f}}{\overline{f} - \underline{f}}(\overline{u_f} - \underline{u_f}) \tag{3.3}$$

$$s.t. \qquad pl + \sum_{t=1}^{n} \frac{(p - pl)(1+r)^{t}(1-f)^{t-1} f}{(1+i)^{t}} \geq R_{0} \tag{3.4}$$

$$0 \leq l \leq l_{\max}, \;\; 0 \leq f \leq f_{\max}$$

Since we have placed no *a priori* restrictions on the relationships between the part-worths, and hence the shape of the utility function, this maximization problem has no analytical solution. The solution strategy we employ involves generating fee structures which satisfy the constraints of this optimization, at each level of $R_{0,}$ and then testing each one of these candidate fee structures, via the objective function, to determine which one maximizes utility. We generate load and expense ratio combinations that satisfy the constraints by writing equation (3.4) in terms of *l* and then repeatedly solving this equation for different values of *f*, over the interval tested in the conjoint study.[8]

$$l(R_{0}, f, i, r, n) = \frac{R_{0}(i-r) + f(1+r)(R_{0} + (p(-\frac{(f-1)(1+r)}{1+i})^{n} - 1))}{p(i - r + f(1+r)(-\frac{(f-1)(1+r)}{(1+i)})^{n})} \tag{3.5}$$

For some values of $R_{0,}$ particularly the most extreme values examined, some values of *f* will generate a required load which not in the interval examined in the conjoint analysis. These load and expense ratio pairs must be discarded since accurately computing the utility of such a pair is impossible.[9] The optimization is computationally intensive, but straightforward, and provides the best method for determining optimal fee structure without imposing additional structure on the model.

One weakness of these types of extreme point envelopment methods is that the efficient frontier generated from such methods can be very sensitive to measurement error in the underlying constructs. We were concerned that the estimated efficient fee structures might be very sensitive to small changes in the estimated conjoint parameters. To mitigate this concern we numerically integrated over the range of each of the estimated part-worths via Monte Carlo integration and performed the numerical optimization for each draw from the candidate distributions. This method invokes the asymptotic normality of the parameter distributions. Since, in the application we will detail in the next section, the parameter estimates are derived from 1000 choice experiments, and industrial applications are likely to have substantially larger sample sizes than this, we feel comfortable in sampling from normal distributions. The reported efficient fee structures are the expected values of the optimized load and expense ratio combinations generated from

[8] In the application detailed in the next section we allow *f* to vary over the interval $[0, .025]$. We increment *f* by 1 basis point (.0001) and solve equation (3.5) at each step.

[9] The GAUSS code used for this optimization is available from the authors upon request.

1000 Monte Carlo draws. Further, we repeated this procedure for different numbers of draws to insure stability of the final solution. Unless the variances of the estimated conjoint parameters are very small, we recommend taking this precaution in industrial applications.

Now that we have laid out the methodology for generating efficient fee structures, we will demonstrate this methodology via a field study. We begin by estimating an aggregate version of the above model, solving for a single efficient frontier for the entire sample, and then move to latent-class analysis where we explore whether it is reasonable to develop multiple fee structures for the same fund, targeting these different fee structures to the relevant consumers.

4. A FIELD STUDY

4.1. Data Collection

We collected experimental choice task data from 50 current mutual fund investors.[10] Participants were paid $5 for what was on average a 20-30 minute task. Each participant was asked 20 choice tasks.[11] In addition, prior to answering the choice tasks, each participant provided demographic information and answered a short quiz designed to test their knowledge of financial markets and investing.[12] The demographic characteristics of these investors were quite diverse.[13] In addition to information on load and expense ratios, the choice task also included the attributes

1. Mutual fund company name
2. Fund performance over the previous year
3. Average annual performance over the previous 10 years
4. The fund’s beta rating.[14]

These attributes were selected to correspond to the information provided to consumers in a fund’s prospectus, as well as information commonly provided in a fund’s annual report. While there certainly are other attributes that may influence fund choice, these attributes are commonly highlighted in the literature distributed by mutual fund companies. Respondents were told that these were attributes of hypothetical U.S. common stock mutual funds.

We presented fund fee information in a format consistent with that found in a fund’s prospectus.[15] Mutual fund companies are required by law to calculate the actual dollar cost of the fee structure they currently charge for various holding times, assuming a 5% market return and a $1,000 initial investment. They are also required to then inform potential investors that this information cannot be used to predict future fees the fund may charge. In the conjoint task we present investors, there are 36 possible load and expense ratio combinations. For each of these combinations we developed a schedule of actual costs, virtually identical to that which would appear in a prospectus. These 36 schedules were organized into a short manual which was given

[10] We suggest a considerably larger sample size for industrial applications.

[11] The number of choice tasks is within the range suggested by Johnson and Orme (1996).

[12] Please see Appendix A for the questionnaire on demographics and Appendix B for the finance quiz. The finance quiz was adopted from the quiz that Vanguard Inc. uses to determine the investment knowledge of potential clients.

[13] A breakdown of the demographics of the sample is available from the author upon request.

[14] A beta rating measures the volatility of a fund in relation to a relevant index (usually the S&P 500 for domestic equity funds). For example, a beta rating of 1.2 indicates that the fund is approximately 20% more volatile than the index.

[15] See Appendix C for an example.

to participants and could be referenced by them at any point during the conjoint task. Thus, for any profile that could appear in the choice task, participants had access to the same type of cost information that they would have had from a prospectus.

The choice task presented full product profiles in triplets and asked consumers which fund they preferred. We employed a random profile design. The levels chosen for each of the attributes are given in Table 2.

Table 2: Experimental Choice Design

	Attributes					
	Company	Load	Expense Ratio	1-Year Return	10-Year Return	Beta Rating
Levels	Fidelity	No Load	0%	10%	5%	0.7
	Vanguard	1%	.5%	20%	10%	0.9
	T. Rowe fee	2%	1.0%	30%	15%	1.1
	Dreyfus	3%	1.5%	40%	20%	1.3
	Pecunia	4%	2.0%	50%	25%	1.5
		5%	2.5%			

The levels of the attributes were chosen to reflect market conditions at the time of this study. The company name "Pecunia" is fictitious and is included as a benchmark for the utility of company names. To generate expected revenue for any given load and fee structure we must specify values for the expected holding time for each investment (*n*), the expected market rate of return (*r*) as well as the internal rate of return (*i*). We use 5 years ($n = 5$) as the expected holding time for the given mutual fund.[16] In practice, mutual fund companies will almost certainly have access to data which would provide a more precise measure of *n* for the particular fund they are interested in examining. Since we are demonstrating this procedure for a general class of funds, and not any particular fund, we used the holding time which seemed most appropriate for market conditions as of this writing. We use the long-run average annual return of 10.7% on U.S. equities as a measure of the expected market rate of return. The internal rate of return (*i*) is set to 12%, which is a conservative figure for this industry.

4.2 Estimated Preferences and Efficient Fee Structures

Table 3 provides the aggregate and two-class solution to conjoint estimation,[17] while Figure 4 depicts efficient fee structures derived from applying the optimization detailed in section 4.4 to the aggregate solution. While it is possible to generate fee structures for any revenue level from $0 to $157 in this application, we depict only fee structures corresponding to revenue levels in

[16] This measure is consistent with conversations we have had with practitioners who have indicated that during times of strong market conditions the average holding time of U.S. equity mutual funds is about 5 years, while during times of weak market conditions it can move up to about 7-8 years.

[17] The Consistent AIC measure suggested a two-class solution to this problem. Reported standard errors of the parameter estimates have been inflated by 15% to account for within respondent serial choice dependencies.

the $20-$100 range for clarity of presentation. We will discuss the aggregate solution here and discuss the implications of the two-class solution in the following subsection.

Table 3: Conjoint Results

		Aggregate Model	Latent-Class 1	Latent-Class 2
Company				
	Fidelity	0.09 (0.09)	0.20 (0.14)	-0.09 (0.14)
	Vanguard	0.01 (0.09)	0.05 (0.15)	-0.03 (0.15)
	T. Rowe Price	0.08 (0.09)	0.16 (0.14)	-0.03 (0.14)
	Dreyfus	0.00 (0.09)	-0.01 (0.14)	0.05 (0.14)
	Pecunia	-0.19 (0.09)	-0.39 (0.14)	0.10 (0.14)
Load				
	No Load	0.88 (0.10)	0.92 (0.18)	1.13 (0.18)
	1%	0.48 (0.10)	0.55 (0.18)	0.52 (0.18)
	2%	0.05 (0.11)	0.14 (0.18)	-0.10 (0.18)
	3%	-0.25 (0.11)	-0.15 (0.19)	-0.47 (0.19)
	4%	-0.49 (0.13)	-0.64 (0.18)	-0.37 (0.18)
	5%	-0.67 (0.13)	-0.82 (0.21)	-0.70 (0.21)
Expense Ratio				
	0%	0.99 (0.10)	1.20 (0.18)	0.96 (0.18)
	.5%	0.50 (0.10)	0.62 (0.18)	0.46 (0.18)
	1.0%	0.05 (0.11)	0.10 (0.18)	0.08 (0.18)
	1.5%	-0.07 (0.11)	-0.02 (0.18)	-0.17 (0.18)
	2.0%	-0.72 (0.13)	-1.07 (0.19)	-0.52 (0.19)
	2.5%	-0.74 (0.13)	-0.83 (0.21)	-0.80 (0.21)
1-Year Return				
	10%	-0.27 (0.10)	-0.49 (0.17)	-0.08 (0.17)
	20%	-0.14 (0.10)	-0.38 (0.16)	0.21 (0.16)
	30%	0.06 (0.10)	0.10 (0.16)	0.00 (0.16)
	40%	0.09 (0.10)	0.19 (0.16)	0.02 (0.16)
	50%	0.27 (0.10)	0.58 (0.17)	-0.15 (0.17)
10-Year Return				
	5%	-1.23 (0.14)	-0.68 (0.91)	-3.81 (0.91)
	10%	-0.70 (0.11)	-0.30 (0.36)	-1.25 (0.36)
	15%	-0.06 (0.10)	-0.02 (0.28)	0.59 (0.28)
	20%	0.76 (0.09)	0.39 (0.26)	1.82 (0.26)
	25%	1.23 (0.09)	0.61 (0.28)	2.65 (0.28)
Beta Rating				
	.7	0.05 (0.10)	0.12 (0.17)	-0.11 (0.17)
	.9	0.06 (0.10)	0.00 (0.16)	0.14 (0.16)
	1.1	0.13 (0.10)	0.09 (0.16)	0.16 (0.16)
	1.3	-0.10 (0.10)	-0.17 (0.17)	-0.07 (0.17)
	1.5	-0.14 (0.10)	-0.05 (0.16)	-0.12 (0.16)

*standard deviations of parameter estimates in parentheses

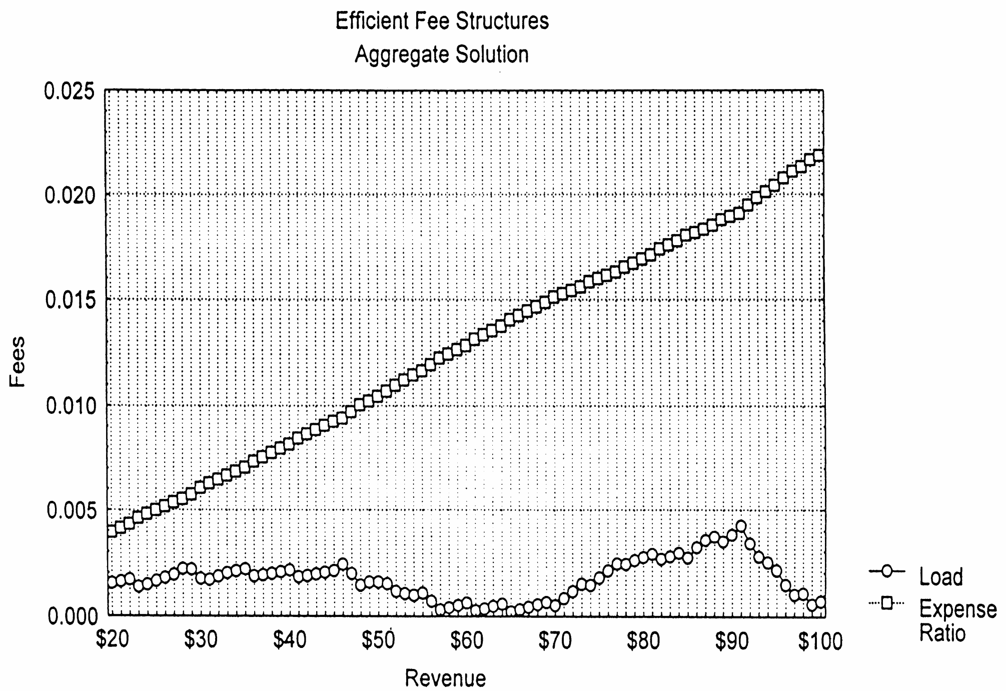


Figure 4

The efficient fee structure suggested by the aggregate model clearly points to maintaining very low loads. As the desired revenue level increases the efficient load rarely rises above 50 basis points. This is not an entirely surprising result given what we observe in the marketplace. Many domestic equity funds maintain expense ratios over 100 basis points while not charging a sales load. These fund managers believe that the revenue forgone from not charging a load is more than offset by the increased probability that a consumer shopping for a fund will choose their fund because of its no-load status. Expense ratios, on the other hand, generate revenue more efficiently relative to their utility cost. Hence, our model suggests that managers should be less wary about letting expense ratios rise to generate additional revenue for their funds.

Notice that the model we developed does not suggest a single best load and expense ratio combination for a given mutual fund. The model rules out many fee structures as being inefficient, but is silent on which of the resultant efficient structures the manager should choose. The fund manager must then use his/her knowledge of market competition as well as the overall strategic role the fund plays in the company's portfolio of funds to determine which of the candidate fee pairs is most appropriate for the fund.

4.3. Some Opportunities for Price Discrimination

If systematic differences exist in how consumers use fee structure information to make mutual funds choices, fund managers can use this information to design and target different fee structures to different consumers and more fully capture the profit potential in the marketplace. Some basic price discrimination practices already exist in this marketplace. Some mutual fund companies waive, or lower, the sales load for their funds if an investor invests an amount above a set lower bound. We provide a simple framework for uncovering fee discrimination opportunities in the marketplace for mutual funds.

A common way of examining the importance of each attribute in the choice process is to compute attribute importance indices. In the application at hand these indices can be computed by taking the ratio of the difference between the most extreme utility values for a given attribute to the sum of the differences in extreme utility values across all attributes within a given class and multiplying this ratio by 100.[18] The attribute importances for each of the consumer classes can be found in Table 4.

Table 4: Attribute Importances

	Class 1	**Class 2**
Company	8	2
Load	24	17
Expense Ratio	31	16
1-Year Return	15	3
10-Year Return	18	59
Beta Rating	4	3

The data suggests that the first of the two classes identified by the conjoint analysis can be reasonably described as fee sensitive, their two highest importance weights correspond to the attributes Load and Expense Ratio, while the second class primarily focuses on 10-Year Return in making their fund selection. We will call these groups the "fee-oriented group" and the "past-returns-oriented group" respectively. The choice processes of these two groups imply different efficient fee structures. Figures 5A and 5B depict the efficient fee structures generated from these different choice processes.

[18] For example, the importance weight for Load for Class 1 can be computed as

$$\left(\frac{.92+.82}{(.20+.39)+(.92+.82)+(1.20+.83)+(.49+.58)+(.68+.61)+(.12+.05)}\right)*100\approx 24$$

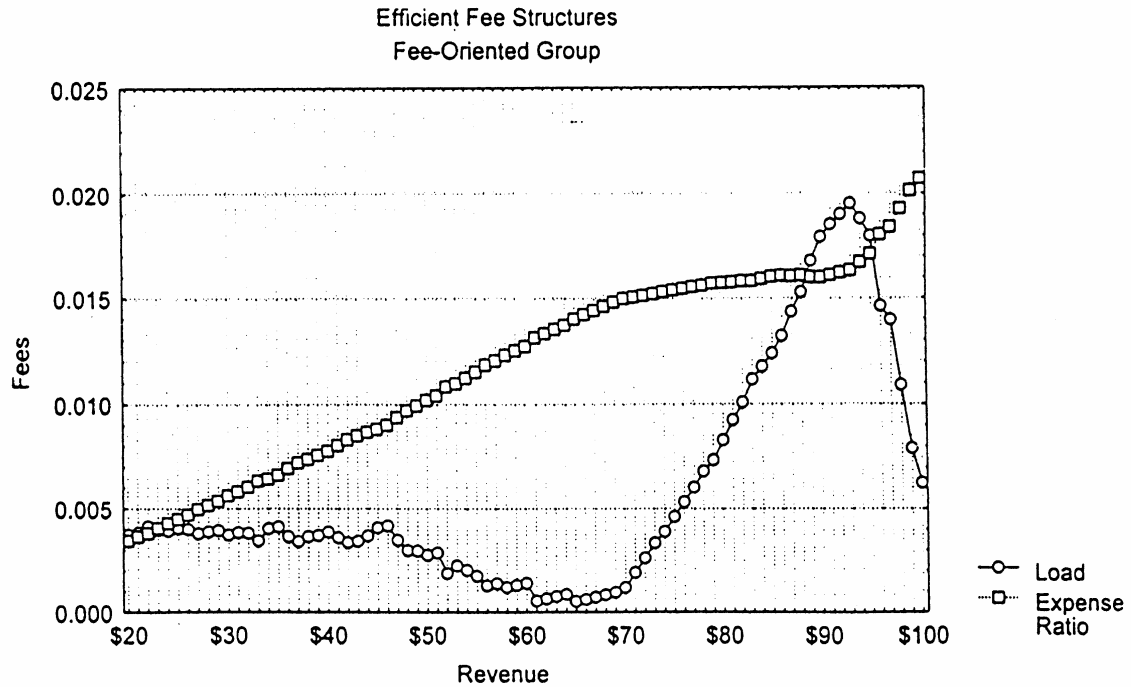


Figure 5A

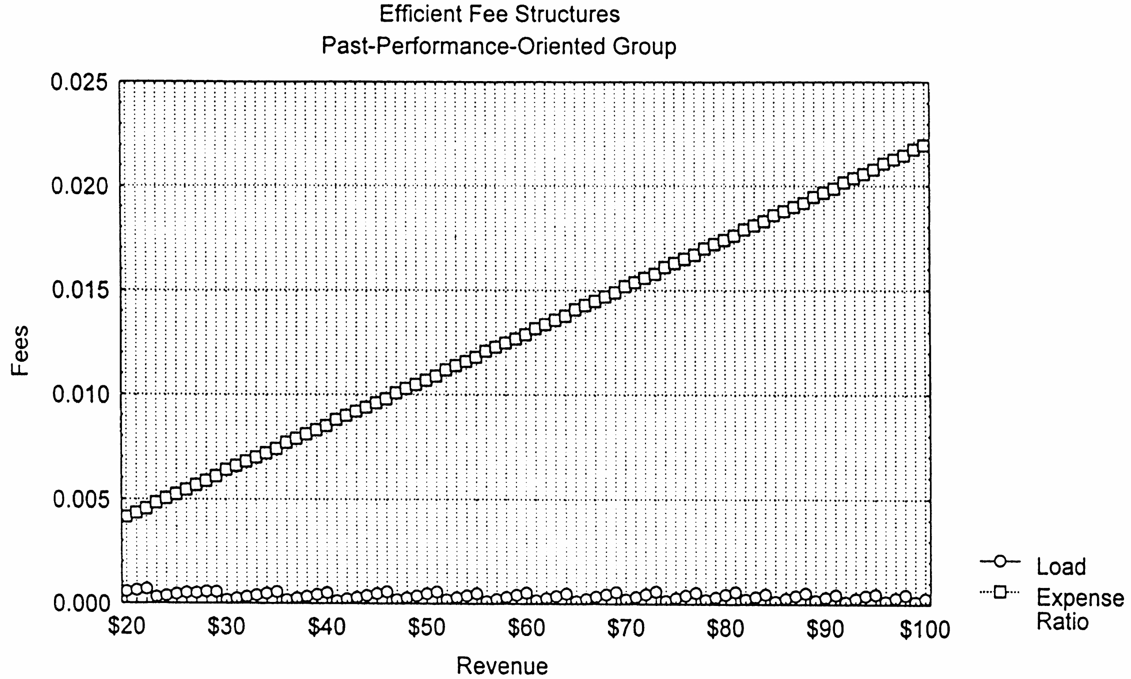


Figure 5B

The efficient fee structures generated for the past-performance-oriented group differs little from those generated from the aggregate solution. With this group, fee increases should generally involve around increasing the expense ratio and not the load. The fee-oriented group, in addition to their greater overall focus on prices, places greater emphasis on the expense ratio relative to the load. This translates into efficient fee structures which have higher loads and lower expense ratios than those of the past-performance-oriented group. In particular, notice that efficient loads rise sharply between revenue levels of about $70-$90 while efficient expense ratios remain relatively flat. If we examine the conjoint parameter estimates for the Latent-Class 1 group (fee-oriented group) it is relatively easy to see why these flat expense ratios and sharp rise in loads arise over this range. Estimated utilities for expense ratios drop sharply from the 1.5% level to the 2% level (-0.02 to -1.07). Thus, the algorithm we are using to generate these fee structures imposes a heavy utility penalty as the expense ratios rise above 1.5%. This creates resistance for efficient expense ratios near the 1.5% level and forces the loads upward to satisfy the revenue requirements.

What of course would be interesting to fund managers is if they could identify these different groups of investors and target their fees accordingly. To investigate this issue we used the information collected in the demographic questionnaires and pre-conjoint finance quiz. In particular we believed that a consumer's income, educational level, planning horizon, and knowledge of and experience in investing might provide insights into how they evaluate mutual funds. To this end, we performed a logistic regression where a dummy variable indicating latent-class membership served as the dependent variable and the above mentioned constructs the independent variables. There are no clear, theoretically sound, hypotheses to test here. We are simply interested in examining whether some variables which are readily available to fund companies have the ability to discriminate between different efficient fee structures. The variables in this regression are operationalized as follows:

- Class – a dummy variable taking the value 1 if the respondent was assigned to the first latent class (fee-oriented group) and the value 0 if he/she was assigned to the second latent class (past-performance-oriented group)
- Income – a categorical variable denoting the income range indicated on question 5 of the questionnaire. The variable is coded such that increasing values represent increasing income.
- Education – a categorical variable denoting the educational level indicated on question 7 of the questionnaire. Again, increasing values represent increasing educational levels.
- Portfolio Complexity – a variable indicating the number of different kinds of mutual funds in which the respondent indicated that they were currently invested. This variable is a simple summation of the categories marked in question 4.
- Planning Horizon – The number of years, indicated in question 9, that the respondent planned on keeping a "substantial amount of money" in the mutual funds in which they were currently invested.
- Fee Question – A dummy variable taking the value 1 if the respondent answered question 10 of the finance quiz correctly (the correct answer is true) and the value 0 otherwise.
- Basic Financial Knowledge – The total number of correct answers on the finance quiz net of the fee question.

The results from the logistic regression of the independent variables on the dependent variable, Class, are given in Table 5.

Table 5: Latent-Class Predictors

Variable	
Income	-0.44 (0.38)
Education	-0.71 (0.44)
Portfolio Complexity	0.42 (0.28)
Planning Horizon	0.02 (0.03)
Fee Question	1.76 (0.79)*
Basic Financial Knowledge	-0.57 (0.25)*

*p<.05

The results of this analysis are quite interesting. While it is difficult to argue that there exists a single optimal decision rule for mutual fund selection, a significant body of literature in finance suggests that past performance is a very poor indicator of future performance and hence would tend to advocate fee minimization as a mutual fund selection strategy. In our study, consumers who demonstrated a greater knowledge of basic finance were more likely to be classified in the past-performance-oriented group. However, those mutual fund investors who answered the fee question correctly were more likely to focus on fees in their selection process. What this result seems to imply is that knowledge of basic finance does not necessarily translate into the ability to reasonably interpret the impact of mutual fund prices on overall fund performance. Somewhat surprisingly, the demographic variables Income and Education did not play a significant role in determining to which group respondents were assigned. Likewise, Planning Horizon and Portfolio Complexity were not significant. In particular, we were surprised that Planning Horizon did not have a significant impact. What little evidence exists that past performance provides clues to future performance is restricted to the short run. This suggests that consumers with longer planning horizons should be particularly concerned with fees as opposed to past performance. Our results do not bear this out.

Our attempts at characterizing the consumers who comprise the latent-classes that the conjoint uncovered met with only limited success. While the efficient fee structures our model suggests for the two groups are quite different, we have discovered only a very limited way to profile these two groups.

5. Discussion, Managerial Implications and Limitations

We developed a decision model for aiding mutual fund managers in determining the fee structures of their funds. This model is both easy to understand and easy to use. It rules out a large number of possible fee structures as being inefficient and suggests a much smaller set of structures for further consideration by management. It points to opportunities for price discrimination in this market.

While working towards a model that could be implemented by mutual fund managers, we made several trade-offs in favor of parsimony and ease of implementation at the expense of fully characterizing the richness of this marketplace. We modeled mutual fund returns as independent of the fees set by management. Clearly this is not the case in practice. Higher loads and expense ratios reduce the overall return measurements for mutual funds. While the finance literature suggests that potential investors should not put much weight on past performance figures, and should instead minimize costs, there is little doubt that consumers do in fact put considerable weight on these performance figures. Many companies advertise their past performance and popular mutual fund ratings services rate funds based on past performance. While a moderate increase or decrease in fees will have only a marginal impact on the performance figures of a given fund, even small changes in performance measurements may have a considerable impact on consumer behavior if this change reverses the position of the fund relative to some established benchmark (i.e. The S&P 500 Index). Since, *ex ante*, managers do not know what their position will be with respect to these benchmarks it is difficult to build these kinds of contingencies into a pricing decision model. Yet, some understanding of how pricing decisions may affect performance, particularly with respect to competitors and common benchmark measures, is necessary to make effective pricing decisions.

The latent-class conjoint methodology we utilized to capture consumers' preferences for different fee structures assumes that consumers can be divided into homogenous groups. This is almost certainly not true. While the methodology utilized here is better than the standard aggregate models common in practice it still suffers from the possibility of obscuring potentially important differences across respondents. Research on conjoint methodology continues on how to best obtain individual-level parameter estimates from conjoint tasks (Zwerina and Huber 1996; Lenk, DeSarbo, Green, and Young 1996). Since contributions to conjoint methodology is not the goal of this research, we used what we believe is the best available method, backed by published research, at this time. As conjoint methodology evolves it will become possible for practitioners to estimate the model at an individual-level and hence use this information, combined with the revenue data, to tailor fee structures for individual investors.

Finally, our model takes the profits generated from any given mutual fund as independent of the prices charged by other funds in the company's fund portfolio. Given that many mutual fund companies have, at a minimum, several funds in their portfolio, it is reasonable to believe that pricing decisions made for one fund may affect capital inflows and outflows in other funds in the portfolio. These sorts of cross-funds issues are not dealt with by our model.

Financial markets provide a rich environment to study consumer behavior. Consumers make decisions that can potentially make large differences in their comfort at retirement, ability to afford a college education for their children, and many other contingencies about which people care a great deal. As the marketplace for financial products and services grows increasingly complex, it is important that we understand how consumers make these types of decisions. It is important both for good business practice, as well as a better understanding of how this marketplace should be regulated. Our research has uncovered the consumer choice process for mutual funds in a practical, yet incomplete, way. We hope that other researchers will more closely examine the consumer choice process in these marketplaces to uncover the richness of the phenomena.

APPENDIX A

Respondent Information

We would like you to provide us some demographic information. Please be assured that we are not interested in this information for any other reason than to conduct research on mutual fund selection. Your personal demographic information will be revealed to no one. While it is important to our research to have this information, if you feel uncomfortable answering any of these questions please tell the research assistant that you do not wish to answer the questionnaire.

1. Please indicate your sex.
   1. Male
   2. Female

2. Please indicate your age ___________

3. Your employment status could best be described as
   1. Full-time employed outside the home
   2. Part-time employed outside the home
   3. Homemaker
   4. Retired

4. In which kinds of mutual funds are you currently invested (circle all that apply)
   1. Domestic Stock Funds
   2. Government Bond Funds (either federal, state, or municipal)
   3. Corporate Bond Funds
   4. Money Market Funds
   5. International Stock Funds (non-emerging market)
   6. Emerging Market Funds
   7. Other (please specify) ____

5. Please indicate the range of your gross (pre-tax) yearly income (include all sources of income)
   1. Less than $15,000
   2. $15,000 - $25,000
   3. $25,000 - $40,000
   4. $40,000 - $65,000
   5. $65,000 - $100,000
   6. over $100,000

6. Do you currently use a professional financial planner to help you with your investment choices.
   1. Yes
   2. No

7. Your education level can best be described as
    1. Did not finish high school
    2. High school diploma or GED
    3. Took some college courses
    4. Associates Degree
    5. Bachelors Degree
    6. Masters Degree
    7. Doctorate Degree

8. Please briefly describe your main sources of information about mutual funds (i.e. Wall Street Journal, a friend who you believe is good at investing, a financial planner, etc.)

9. What is your current anticipated time horizon on money that you have invested in stock mutual funds. In other words, how long do you think that it will be before you spend a substantial amount of the money you have invested in these funds?

APPENDIX B

Preview Quiz for Mutual Fund Selection

To help us determine your knowledge of mutual fund investing, we have devised the following short quiz. Please answer the following ten questions by circling the number corresponding to the answer you believe is correct. You may be assured that your performance on this quiz will be used solely for research purposes and no information about your performance will be provided to anyone not directly involved in this research.

1. If interest rates rise, the price of a bond fund will
   1. Rise
   1. Fall
   2. Remain the same

2. A stock fund’s beta rating can best be described as
   1. A measure of the relative performance of the fund vs. the S&P 500 index.
   2. A measure of the relative volatility of the fund vs. the S&P 500 index.
   3. A measure of the relative growth of the fund vs. the S&P 500 index.
   4. A measure of the relative capital outflow of the fund vs. the S& P 500 index.

3. The long run historical average return for U.S. common stocks is about
   1. 10%
   2. 15%
   3. 20%
   4. 25%

4. A mutual fund’s capital gains distribution can best be described as
   1. The interest and dividends earned by a fund’s securities.
   2. The gain you earn when you sell fund shares for a profit.
   3. The cost of operating a mutual fund, expressed as a percentage of net assets.
   4. The payments to shareholders of profits from the sale of securities in the fund’s portfolio.

5. A 12b-1 fee is
   1. A charge incurred at the time you initially purchase fund shares.
   2. A fee for reinvesting fund dividends.
   3. A fee charged against the fund’s assets for distribution and marketing related expenses.
   4. A charge incurred to cover the management and administrative costs of the fund.

6. A money market mutual fund is guaranteed by the U.S. government against principal loss.
   1. True
   2. False

7. If you invest in a bond mutual fund with an average maturity of five years, this means that you cannot withdraw your money from the fund within a five year period without incurring a penalty.
   1. True
   2. False

8. The income provided by a stock mutual fund is free from Federal income tax.
   1. True
   2. False

9. Market risk is the potential for a decline in the value of an investment.
   1. True
   2. False

10. All other things being equal, the lower the fees charged by a mutual fund, the higher its return.
    1. True
    2. False

APPENDIX C

Fee Structure 2-1

Expenses and Fees

Sales Commission to Purchase Shares (Load)	2%
Investment Management Fee	1%

Example

Assuming a 5% annual return and redemption at the end of each period, the total expenses related to a $1,000 investment would be

1 Year	3 Years	5 Years	10 Years
$30	$52	$76	$143

This example assumes reinvestment of all dividends and distributions and that the total fund operating expenses listed above remain the same each year. This example should not be considered a past or future example of expenses. Actual expenses vary from year to year and may be higher or lower than those shown.

REFERENCES

Blake, C. R., E. J. Elton and M. J. Gruber (1993), “The Performance of Bond Mutual Funds,” *Journal of Business* 66 (July): 371-403.

Brown, S. J., and W. N. Goetzmann (1995), “Performance Persistence,” *Journal of Finance* 50 (June): 679-98.

Chance, Don M. and Stephen Ferris (1987), “The Effect of 12b-1 Plans on Mutual Fund Expense Ratios: A Note,” *Journal of Finance* 42(4), 1077-1082.

Chintagunta, P., D. Jain, and N. Vilcassim (1991), “Investigating Heterogeneity in Brand Preference in Logit Models for Panel Data,” *Journal of Marketing Research*, November.

Chordia, T. (1996), “The Structure of Mutual Fund Charges,” *Journal of Financial Economics* 41: 3-39.

Christoffersen, Susan Kerr (1997), “Fee Waivers in Money Market Mutual Funds,” unpublished doctoral thesis.

DeSarbo, Wayne, Venkatram Ramaswamy, and Steven H. Cohen (1995), “Market Segmentation with Choice Based Conjoint Analysis,” *Marketing Letters*, 6, 137-48.

Elton, E. J., M. J. Gruber, and C. R. Blake (1996), “The Persistence of Risk-Adjusted Mutual Fund Performance,” *Journal of Business*, 69 (April): 133-57.

Elton E. J., M. J. Gruber, S. Das, and M. Hlavka (1993), “Efficiency with Costly Information: A Reinterpretation of Evidence for Managed Portfolios,” *Review of Financial Studies* 6 (Spring): 1-22.

Gönül, Fusun, and Kannan Srinivasan (1993), “Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models: Methodological and Managerial Issues,” *Marketing Science* 12(3): 213-229.

Grinblatt, M., and S. Titman (1989), “Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings,” *Journal of Business* 62, (July): 393-416.

Grinblatt, M., and S. Titman (1992) “Performance Persistence in Mutual Funds,” *Journal of Finance* 47 (December): 1977-84.

Grinblatt M., and S. Titman (1993) “Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns,” *Journal of Business* 66, 47-68.

Hendricks D., J. Patel, and R. Zeckhauser (1993), “Hot Hands in Mutual Funds: Short-run Persistence of Performance, 1974-88,” *Journal of Finance* 48 (March): 93-130.

Johnson, Richard M. and Bryan K. Orme (1996), “How Many Questions Should You Ask in a Choice-Based Conjoint Study,” *Sawtooth Software Technical Paper.*

Johnson, Richard M. and Jonathan Pinnell (1995), “Comment on Incorporating Prior Knowledge into the Analysis of Conjoint Studies,” *Sawtooth Software Technical Paper.*

Kihn, John (1996), “To Load or Not to Load? A Study of the Marketing and Distribution Charges of Mutual Funds,” *Financial Analysts Journal*, May/June: 28-36.

Lehmann, B. N., and D. Modest (1987), “Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons,” *Journal of Finance* 42 (June): 233-65.

Lenk, P. J., W. S. DeSarbo, P. E. Green, and M. R. Young (1996) “Hierarchical Bayes Conjoint Analysis: Recovery of Part-Worth Heterogeneity from Reduced Experimental Designs,” *Marketing Science*, 15(2), 173-191.

Malkiel, B. G. (1995), “Returns from Investing in Equity Funds,” *Journal of Finance* 50 (June): 549-72.

Moore, W. L., J. Gray-Lee and J. Louviere (1997), “A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation,” *working paper*.

Moore, W. L., Raj B. Myhta and Teresa M. Pavia (1994), “A Simplified Method for Constrained Parameter Estimation in Conjoint Analysis,” *Marketing Letters* 5(2), 173-181.

Murthi, B. P. S., Yoon K. Choi, and Preyas Desai, “Efficiency of Mutual Funds and Portfolio Performance Measurement: A Non-Parametric Approach,” *European Journal of Operations Research* 98: 408-18.

The 1997 Mutual Fund Factbook, The Investment Company Institute.

Vriens, M., M. Wedel and T. Wilms (1996), “Metric Conjoint Segmentation Methods: A Monte Carlo Comparison,” *Journal of Marketing Research*, 33, 73-85.

Zwerina, K. and J. Huber (1996), “Deriving Individual Preference Structures from Practical Choice Experiments,” working paper, Duke University.

COMMENT ON WILCOX

Michael G. Mulhern
Mulhern Consulting

I believe Ron's paper offers research practitioners several major insights for improving the quality of our research.

First, this paper suggests a way to provide managers what they really want from a conjoint study; a linkage between preference/choice shares and revenue or profit measures. By using the microeconomic concept of iso-preference curves, Ron generates iso-revenue curves which allow managers to investigate various optimal expense/fee combinations (that is, the "efficient frontier").

Secondly, Ron illustrates the value of disaggregating choice data via latent class modeling. The aggregate model indicates 10-year performance is the most important attribute. However, latent class found two segments, approximately equal in size. One was a performance driven group while the other was a fee driven segment. Using only the aggregate results would have fostered a strategy that addressed the needs of only half the market!

A final strength of this paper is the realistic presentation of the fee and expense structure. By presenting this information in a format very familiar to most investors (i.e. the prospectus format), reliability was likely enhanced.

Suggestions for improvement are

- Suspected design interactions need to be reviewed. Several attribute combinations may be correlated. They include brand and fee structure, 1 and 10 year performance, as well as beta and performance.
- Consider a single performance attribute with levels of 1, 3, 5, and 10 years. The current design does not incorporate 3 and 5-year performance indicators.
- Evaluate the impact of restricted or prohibited combinations on design efficiency. Not all mutual funds companies offer funds at the load specified in the design. In fact many of the combinations would need to be prohibited (assuming current fee structures are maintained), dramatically reducing design efficiency. Perhaps combining brand/mutual fund company and load into a single attribute would resolve this problem. However, recognize that doing so would eliminate the ability to investigate their relative contribution to choice.
- The term "beta" may be foreign to many respondents; particularly the less sophisticated investor. I would suggest either defining the term in the introductory material or renaming the attribute "fund volatility."
- No matter how the design is modified, model interactions in an attempt to improve the model's forecasting ability.

- Consider consumer factors unique to this product/market that may impact your results. Two come to mind. First, there is a wide range of investor knowledge. At the extreme, consumers can be very naïve or very sophisticated. In addition, mutual fund buyers may have different objectives; some may be traders while others may be investors using a buy and hold strategy.

In conclusion, the strengths of Ron's paper far outweighed the weaknesses. Particularly insightful were the use of iso-revenue curves and latent class modeling.

KNOWING WHEN TO FACTOR: SIMULATING THE TANDEM APPROACH TO CLUSTER ANALYSIS

Andrew Elder
IntelliQuest, Inc.

BACKGROUND

One of the many decisions an analyst has to make in the course of cluster analysis segmentation concerns pre-treatment of input variables. Should they be standardized, centered, double-centered, mean substituted, weighted, factor analyzed, or just left alone? One alternative is to use the "tandem" approach to cluster analysis, which involves pre-treatment of input variables by principal components factor analysis (Punj and Stewart 1983). Factor analysis of intercorrelated input variables produces orthogonal composite "factor scores" that are linear combinations of the input variables. The respondent-by-factor score matrix then becomes the input to the cluster analysis.

Some analysts justify factor analysis of input variables primarily as a data reduction technique (Wilcox 1991). Another rationale for tandem cluster analysis is that if unequal numbers of raw variables load on the various factors, then clustering on the raw variables themselves would cause "essentially an implicit weighting of these variables" (Aldenderfer and Blashfield 1984). This weighting could bias the cluster solution in favor of over-represented factors and against poorly represented factors. For example, if ten variables measure brand quality and only one measures price, a cluster analysis using raw items rather than factors will be influenced ten times as much by quality differences as by price differences.

On the other hand, some arguments have been advanced against the tandem approach. According to Dunteman (1989), there is "no advantage in transforming the original observations to principle component scores prior to clustering." Rholf (1970) notes that clustering on factor scores rather than on raw items tends to diminish differences between groups that are not widely separated and this makes such groups more difficult to differentiate and hence to detect. A related complaint invokes the central limit theorem to suggest that "when variables are grouped into factors, a lot of 'smoothing' will occur. Cluster analysis can take advantage of the 'lumpiness' of data and will be impeded by any smoothing that takes place" (Sawtooth Software 1995). Fiedler and McDonald (1991) found in an empirical study that, in three of four data sets tested, clusters produced from raw variables produced larger differences on external profiling variables than did cluster analysis on factor scores.

Chang (1983) and Dillon, Mulani and Frederick (1989) note that the factors with the largest eigenvalues are not necessarily the factors on which groups of objects are most widely separated. Analyzing real and simulated respondents, these studies show that if only the factors with the largest eigenvalues are retained for subsequent use, important distance information may be lost and known groups may be insufficiently distinguished. In their recent review of cluster analysis, Arabie and Hubert (1994) cite these authors and attack tandem clustering. They describe tandem clustering as one of several "misunderstandings," and they conclude that tandem clustering is an "outmoded and statistically insupportable practice." Green and Kreiger (1995) investigate this issue empirically, but the data they employ, individual-level conjoint utilities, is atypical at best: factor analysis of such data is neither routinely performed nor obviously meaningful. Conjoint utilities are not the data for testing the appropriateness of tandem cluster analysis.

In addition to the two extremes of ignoring the inter-relationships between raw input variables and transforming raw variables into orthogonal composites, two other pre-treatment alternatives lie somewhere in between:

1. Using principal components analysis (PCA) to select one or more raw variables per factor to represent each underlying dimension and entering equivalent numbers of these exemplars into the cluster analysis; and
2. Using variable weighting to redress representational imbalance among factors.
    a. The approach automated in the CCA software package used below simply multiplies variables by positive integers. Thus if Factor 1 has two variables (v_1 and v_2) to represent it and Factor 2 has six ($v_3 – v_8$) , the variables representing Factor 1 might have their values multiplied by 3.
    b. Replication weighting is an alternative approach. In this case, the Factor 1 variables might all be copied twice. Then these six variables (v_1, v_2, and two copies of each) could enter the cluster analysis with $v_3 – v_8$, giving the two factors equitable representation.

This paper investigates the attack on tandem cluster analysis put forth most vigorously by Arabie and Hubert (1995). Using data sets from simulated respondents, I compare the validity of clustering on raw variables, factor scores, one item per factor, and two types of weighted variables. I predict that results will be sensitive to the level of representational imbalance (the number of raw items per factor entered into the cluster analysis), so the five pre-treatment options will be tested on the four levels of representational imbalance, for a total of 20 design cells. Each design cell contains 20 cluster analyses on 10 replicated data sets. Each data set is generated using a different initial seed and subjected to two k-means clustering algorithms.

STUDY DESIGN

Structure of Simulated Data

Each data set contains 1,000 simulated respondents in four segments of 250 respondents each. The data possess the characteristics assumed by k-means cluster analysis: segments are hyper-spheres of respondent-points randomly dispersed around a centroid. Three factors underlie 35 raw variable scores for all respondents. Twenty variables load on Factor 1 (variables v_1 to v_{20}), ten on Factor 2 (v_{21} to v_{30}) and five on Factor 3 (v_{31} to v_{35}).

The variable generation process begins with a random, normally distributed variable (v_a) with a mean of 0.0 and a standard deviation of 1.0. For each respondent, v_1 through v_{20} are equal to v_a plus another random normal variate with a mean of 0.0 and a standard deviation of 1.0. The use of a consistent base (v_a) keeps each respondent's values for v_1 through v_{20} close to one another. Subsequent variables are similarly constructed, using a v_b to generate v_{21} to v_{30} and a v_c as the basis for v_{31} to v_{35}.

Additive constants adjust these initial raw scores to create the desired between-cluster differences among all 35 variables loading on the three factors. The pattern of directional differences between segments is shown below. A by-product of this pattern is that it generates the desired factorial structure. A more collinear pattern of differences would have required further steps to effect the proper correlational pattern (e.g., Grisaffe 1993). Finally, I round the scores to integer values, truncate the distribution so that values range from one to five, and create a left skew. SPSS code for generating this artificial data set appears in the Appendix.

	Factor 1 Variables	Factor 2 Variables	Factor 3 Variables
Segment 1	–	+	–
Segment 2	+	+	–
Segment 3	+	–	+
Segment 4	–	+	+

The resulting raw scores resemble typical marketing research data: left- skewed ratings on a five-point scale, with a mean of about 3.5 and a standard deviation of about 1.0. Raw variables load on their respective factors at $r \approx 0.8$ and on the other factors at $r \approx 0.0$. I systematically employ subsets of the 35 raw variables for different analyses. Representational imbalance varies from perfectly balanced (five variables per factor), less balanced (nine variables on Factor 1, six on Factor 2 and three on Factor 3), moderate imbalance (10, 5, and 1), to a more extreme case of imbalance (20, 10, and 1).

Cluster Analysis

I perform the cluster analyses using two popular k-means clustering algorithms. Sawtooth Software's Convergent Cluster Analysis (CCA) program offers a robust algorithm for k-means cluster analysis that searches for convergence by replicating solutions for a given number of clusters up to ten times using different cluster seeds and seeding strategies. Quick Cluster – the SPSS procedure for k-means cluster analysis – lacks CCA's search for convergence but is a convenient alternative given its integration into the SPSS package.

I assess the success of various cluster solutions by cross-tabulating derived cluster membership by known segment membership. Exactly four segments exist, so only four-cluster solutions are sought from both packages. Respondents correctly assigned to segments by the cluster analysis are "hits."

Hypotheses / Hypothesis Testing

I expect that tandem cluster analysis based on composite factor scores will identify known segment membership better than will cluster analysis based on the raw items as the degree of representational imbalance increases. I also expect the other two input variable pre-treatment options (one raw variable per factor and weighting of raw input variables to redress imbalance in factorial representation) to fare better than the use of raw input variables under similar circumstances.

Based on the results of Neal (1989), I expect CCA and Quick Cluster to perform comparably in terms of hits. Since this indeed turns out to be the case, I combine data from the two algorithms in the analyses below.

One additional element I will examine is the significance of changes in hits by varying levels of imbalance and pre-treatment options. A two-way analysis of variance (ANOVA) will measure both main effects and the interaction between them.

Results

The results of the ANOVA are shown in Table 1. Both main effects and their interaction are highly significant. The interaction does not interfere with the one-at-a-time interpretation of the main effects and is not by itself interesting, so the discussion focuses on the main effects.

The mean numbers of hits by pre-treatment option and representational imbalance appear in Table 2. When clustering by raw items, the number of hits falls as representational imbalance increases. For tandem clustering, however, hits do not decrease with increasing imbalance. This is as hypothesized, and demonstrates that certain data conditions will require tandem cluster analysis. I also hypothesized that using pre-treatment options of one item per factor and variable weighting would avoid the problem of representational imbalance. Of these options, one item per factor clustering and replication variable weighting avoid the problem, while multiplicative variable weighting shows the most severe degradation in validity as representational imbalance increases.

Although hits using one item per factor do not fall with increasing representational imbalance, its validity is dominated by the other pre-treatment alternatives. In the presence of representational imbalance, this pre-treatment option gets fewer hits than do either tandem cluster analysis or replicated weighting. In the absence of representational imbalance, single item clustering performs less well than does clustering on raw variables.

Discussion

The ability to replicate known cluster membership can be skewed by representational imbalance, and tandem cluster analysis will sometimes be preferable to clustering on raw variables. Given the level of between-groups separation simulated in this study, raw and tandem clustering produce similar numbers of hits in situations where the factor with the most items has three times (3x) as many as the factor with the fewest. Tandem clustering is superior beyond this level of imbalance, but raw variable clustering produces more hits in the absence of such disproportionate representation.

The use of one variable per factor is dominated by raw and tandem cluster analysis. In the absence of imbalance, using all raw variables produces more hits than does a cluster analysis using one item per factor. Single item clustering is also inferior to the tandem approach when representational imbalance is present. Multiplicative weighting is a mounting disaster as representational imbalance increases, but replication weighting is at least as good as raw and tandem cluster analyses at their best.

In this simulation, tandem cluster analysis provides segmentation results superior to raw variable cluster analysis if representational imbalance is worse than 3x. If representation imbalance is less than 3x, however, tandem cluster analysis is clearly inferior. While these results are specific to the separation levels incorporated into the data sets, they suggest that the decision to use raw versus tandem cluster analysis should be data-driven. Let the results of a preliminary factor analysis serve as a guide.

Likely one's initial impression of tandem clustering depends on one's psychometric orientation. Thinking in factor analysis terms, raw variables contain unique and common variance. Some portion of the unique variance is real information, and some of it is just measurement error. Analysts who view the measurement error component as large may be more inclined to cluster on composite factor scores than on error-laden raw variables. On the other hand, analysts who believe valuable information resides in the unique variance component may prefer clustering on raw items rather than on overly parsimonious factor scores.

For purposes of this experiment, the pattern of segment differences deliberately contains crucial information in the often under-represented dimension. This artifice specifically illustrates a problem with using raw variables in situations of representational imbalance. While it is true that not all real world data sets will suffer from important differentiation along an under-represented dimension, it is also true that analysts who fail to look for such a problem are unlikely to find it.

It may be idealistic to hope that all sets of variables used in cluster analysis are constructed with psychometric sense. At the very least, however, the analyst should factor analyze the set of input variables and check for representational imbalance. If it exists to a relatively severe degree, the analyst may want to try separate cluster analyses on raw variables and on composite factor scores.

The question of whether tandem cluster analysis is a good idea or a bad one thus depends on how equitably the input variables cover the measurement domain. If dimensions are relatively evenly covered, raw variables will do a better job than factor scores. If not, however, using raw variables may exacerbate the problem identified by Chang (1983) and Dillon, Mulani and Frederick (1992): artificially (albeit implicitly) down-weighting dimensions containing important between-segment difference information can bias cluster analysis results.

Thus Arabie and Hubert (1994) draw exactly the wrong conclusions from the work of Chang (1983) and Dillon, Mulani and Frederick (1992). It is not the pre-treatment of factoring, *per se,* that these works call into question, but rather *under*factoring. This may be seen as another reason to err on the high side when choosing the number of factors to retain (Tabachnick and Fidell 1983, Stevens 1996).

REFERENCES

Aldenderfer, Mark S. and Roger K. Blashfield (1984) "Cluster Analysis," Sage University Paper series on Quantitative Applications in the Social Sciences, 07-044, Beverly Hills: Sage.

Arabie, Phipps and Lawrence Hubert (1994) "Cluster Analysis in Marketing Research," *in Advanced Methods of Marketing Research*, Richard P. Bagozzi, ed. Oxford: Blackwell.

Chang, Wei-Chien (1983) "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics*, 32: 267-75.

Dillon, William R., Narenda Mulani and Donald G. Fredrick (1989) "On the Use of Component Scores in the Presence of Group Structure," *Journal of Consumer Research*, 16: 106-12.

Dunteman, George H., (1989) "Principle Components Analysis," Sage University Paper series on Quantitative Applications in the Social Sciences, 07-069, Beverly Hills: Sage.

Fiedler, John A. and John D. McDonald (1991) "Market Figmentation: Clustering on Factor Scores vs Individual Variables," paper presented at American Marketing Association' 1991 A/R/T Forum.

Green, Paul and Abba Krieger (1995) "Alternative Approaches to Cluster-Based Market Segmentation," *Journal of the Marketing Research Society*, 37: 221-39.

Grisaffe, Doug (1993) "Generating Correlational Data for Multivariate Simulations: A Two-Stage Principal Components Procedure for PC Users," paper presented at the TIMS Marketing Science Conference, Tucson.

Neal, William D. (1989) "A Comparison of 18 Clustering Algorithms Generally Available to the Marketing Research Professional," *Proceedings of the Sawtooth Software Conference*, 299-324.

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

Rholf, F. James (1970) "Adaptive Hierarchical Clustering Schemes," *Systematic Zoology*, 19: 58-82.

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

SPSS. Inc. (1998), SPSS 8.0. Chicago: SPSS, Inc.

Stevens, James. (1996) *Applied Multivariate Statistics for the Social Sciences*, Mahwah, NJ: Lawrence Erlbaum.

Tabachnick Barbara G. and Linda S. Fidell (1983) *Using Multivariate Statistics*, New York: Harper and Row.

Wilcox, Marsha (1991) "A Comparison of Factor Analysis and Correspondence Analysis as Data Reduction Techniques with Clustering in Market Segmentation," *1991 Sawtooth Software Conference Proceedings,* Ketchum: Sawtooth Software.

Table 1.
ANOVA of Hits by Representational Imbalance and Pre-Treatment

Source of Variation	Sum of Squares	df	MSE	F	p
Imbalance	350,966	3	116,989	60.86	<.001
Pre-Treatment	844,458	4	211,115	109.83	<.001
I x P	610,959	12	50,913	26.49	<.001
Within Cells	730,423	380	1,922		

Table 2.
Average Number of Hits By Representational Imbalance and Pre-Treatment

	Pre-Treatment				
Representational Imbalance	Raw Variables	Tandem	One Item Per Factor	Multiplicative Weighting	Replication Weighting
5, 5, 5	596	547	496	596	596
9, 6, 3	528	530	454	457	562
10, 5, 1	473	549	494	369	576
20, 10, 1	476	567	504	364	577

APPENDIX

SPSS Code for Generating Artificial Data Set
(with between group differences of 1.0)

```
data list free /seg.
begin data.
1 2 3 4
.
.
.
1 2 3 4
end data.
compute a=.6667.
compute b=1.
compute c=.77.
compute d=1.
compute Va=normal(b).
compute Vb=normal(b).
compute Vc=normal(b).
if (seg=1) Va=Va-c.
.
.
.
if (seg=4) Vc=Vc+c.
compute V1=Va+normal(d).
.
.
.
compute V35=Vc+normal(d).
compute V1=rnd((V1*a)+3.5).
.
.
.
compute V35=rnd((V35*a)+3.5).
```

THE NUMBER OF LEVELS EFFECT: A PROPOSED SOLUTION[1]

Dick McCullough
MACRO Consulting, Inc.

ABSTRACT

The existence of the number of levels effect (NOL) in conjoint models has been widely reported since 1981 (Currim et al.). Currim et al. demonstrated that the effect is, for rank-order data, at least partially mathematical or algorithmic. Green and Srinivasan (1990) have argued that another source of this bias may be behavioral.

In this paper, we first offer an analytic approach that confirms the existence of the algorithmic component to the number of levels effect for rank-order data. We then describe a second study which demonstrates a practical solution to the number of levels effect, regardless of the source of the effect.

INTRODUCTION

The existence of the number of levels effect in conjoint models has been widely reported since 1981 (Currim et al.). The effect occurs when one attribute has more or fewer levels than other attributes. For example, if price were included in a study and defined to have five levels, price would appear more important than if price were defined to have two levels. This effect is independent of attribute range, which also can dramatically affect attribute relative importance.

NOL was originally observed for rank-order preferences but has since been shown to occur with virtually all types of conjoint data (Wittink et al. 1989). Currim et al. demonstrated, for rank-order data, that the effect is at least partially mathematical or algorithmic. Green and Srinivasan (1990) have argued that a source of this bias may also be behavioral. That is, attributes with higher numbers of levels may be given more attention by respondents than attributes with fewer levels. If true, this might cause respondents to rate attributes with a greater number of levels higher than attributes with fewer levels. Steenkamp and Wittink (1994) have argued that the effect is, at least partially, due to non-metric quality responses, which computationally causes ratings data to behave similarly to rank-order data.

The number of levels effect has been widely reported. It is generally agreed that the effect is a serious problem that can and often does significantly distort attribute relative importance scores, utility estimates and market simulation results. And largely due to the fact that the only known practical method for removing this effect has been to hold the number of levels constant across attributes, it has often been ignored in commercial studies.

[1] The author wishes to thank Rich Johnson, Dick Wittink, Jayme Plunkett and Jamin Brazil for their invaluable assistance with this paper.

In this paper, we confirm the existence of an algorithmic component to the number of levels effect for rank-order data and offer a solution to remove the levels effects bias from full-profile conjoint models, both rank-order and ratings. The solution is demonstrated using ratings data.

Two separate studies are reviewed in this paper.

METHODOLOGY-FIRST STUDY

In the first study, we examine a rank-order conjoint data set. The data come from a trade-off study of a new high-technology product. The trade-off data are derived from a rank-order card sort data collection exercise. The study design consists of 21 cards and three attributes. Attribute Price has 3 levels and is a vector attribute. Attribute Brand has 5 levels and is a partworth attribute. Attribute V has 2 levels and can be considered either vector or partworth. The combinations have been selected so the experimental design is reasonably orthogonal and balanced.

The data set is prepared two ways: 1) the data set is kept in its original format and 2) the data set is altered, i.e., degraded, so that one attribute at a time is redefined to have fewer levels. This degradation is achieved by simply removing any cards from the rank- order which include the omitted attribute levels. The degraded form of the data set for attribute Price has 14 cards. The degraded form of the data set for attribute Brand has six or eight cards, depending on which levels were exterior for a given respondent. Attribute V only has two levels so no degradation of the data set is necessary.

Conjoint utilities are estimated for each version of the data set. We estimate the existence and magnitude of the levels effect for each version of the data set using a slight variation of the regression model approach used by Steenkamp and Wittink (1994). In that approach, the relative importance scores for each respondent for a fixed attribute are regressed against the number of levels for that attribute. In the Steenkamp and Wittink study, the sample was split so that half of the sample saw attributes with one set of levels and half saw attributes with another set of levels. In our study, the two data sets, i.e., the complete data set and the degraded data set for a given attribute, are merged to provide variance in the number of levels. Because respondents are exposed to exactly the same stimuli in all versions of the data set (original and degraded), no behavioral component of NOL is possible. Any NOL effect detected will necessarily be due entirely to an algorithmic component.

ANALYSIS-FIRST STUDY

The regression models results (Table 1) show a levels effect for two different attributes. For the well-ordered, i.e., vector, attribute Price, the magnitude of the effect is approximately the same as cited by Steenkamp and Wittink.

For the non-well ordered, i.e., partworth, attribute Brand, the effect, although significant, is substantially less in magnitude.

Table 1

Regression of Attribute Importance on Number of Levels

Attribute	Beta[a]	R^2	ANOVA
Price[b]	.077 *	.03	.001 * (F=10.32)
Brand[c]	.026*	.02	.004 * (F=8.226)

[a] Levels effects for each attribute were estimated by regressing the merged attribute importance scores of the original attribute and the degraded attribute on their respective number of levels.

[b] Price is a well-ordered or vector attribute.

[c] Brand is a non-well-ordered or partworth attribute. However, interior levels were excluded on a per-respondent basis to ensure no exterior levels were excluded in the two-level case.

* Significant at the 95% confidence level

A Proposed Solution to the Number of Levels Effect

The above analysis suggests a possible solution to eliminate entirely the number of levels effect regardless of its source.

Methodology-Second Study

The subject of this study was high-end ice hockey skates. The study was designed to be a full-profile metric conjoint study that had these attributes:

Brand (4 levels)

Price (4 levels)

Visual Design (3 levels)

Weight (3 levels)

Psychological Price Point (2 levels)

Brand and Visual Design were partworth attributes. Price and weight were vector attributes. Psychological price point was a metric attribute with $0 and $1 as its two levels. Respondents were shown a product price that was the sum of the values for the price attribute and the psychological price point attribute. For example, if price were $399 and psychological price point were $1, respondents would see a price of $400. If price were $399 and psychological price point were $0, respondents would see a price of $399.

Respondents participated in a two-stage conjoint exercise. The first conjoint exercise had only two levels for each attribute. The levels used were the exterior levels, i.e., those levels that had maximum and minimum utility for each individual respondent. For the partworth attributes, exterior levels, that is, the most preferred and least preferred levels, had to be identified for each respondent prior to the first conjoint exercise. This was done by direct questioning. For the vector attributes, the numeric maximum and minimum values were assumed to be exterior for all respondents. There were 18 different versions of this two-level design (6 different Brand pairs times three different Visual Design pairs). Respondents in this section of the study rated 12 different hockey skates for purchase interest.

The second conjoint exercise was a full-profile metric conjoint exercise utilizing all levels of all attributes. For this exercise, respondents rated 18 cards.

As was the case in the earlier study, both of these experimental designs were reasonably orthogonal and balanced.

The general concept is to identify attribute relative importance scores from the first stage conjoint exercise (exterior levels only). Utility estimates from this stage should exhibit no number of levels effect since all attributes have the same number of levels. The second stage conjoint exercise should establish the relative preference of levels within attribute.

The full-level utility estimates can then be linearly scaled into the two-level estimates. The resulting utilities will exhibit the correct attribute relative importance and also maintain the relative positions of levels within each attribute.

Data for this second study were collected in late December, 1998, via a Web-based survey, which allowed greater design flexibility and experimental control than paper-and-pencil data collection.

ANALYSIS-SECOND STUDY

Prior to analysis, the data sets were edited so that respondents with individual-level conjoint models that were not significant at least at the 75% confidence level were excluded from further analysis. Approximately one-third of the sample was discarded at this stage.

Additionally, respondents were excluded who did not provide consistent claimed and derived exterior levels, i.e., exterior levels from the direct questioning which were the same as the exterior levels computed from the full-levels conjoint. This second criterion caused a dramatic reduction in sample size. Approximately two-thirds of the remaining sample was discarded at this stage.

The initial sample size was 425. The final sample size was 79.

Upon reviewing possible sources for this high percentage of inconsistent respondents, it was concluded that the wording of the exterior levels direct questions was confusing and misleading. Other sources of this inconsistency might have been model instability, irrational respondents or poor data quality due to the Web-based collection method.

However, the exterior levels questions were redesigned for a subsequent paper-and-pencil study which employed the same study design. Results from that study, while improved, were still disappointing. Approximately half of the sample did not provide consistent claimed exterior levels when compared to derived exterior levels. If either question wording or quality of Web-based data were the primary source of this inconsistency, the paper-and-pencil study should have shown much greater improvement.

Further, we would expect most unstable models and irrational respondents to be excluded by discarding all models which were not significant at least at the 75% confidence level.

Additional possible explanations include respondent indifference to alternative levels, respondent fatigue, confusion- or fatigue-motivated simplification where the respondent would focus on one attribute that was important to him or her and ignore the others. Interaction effects, i.e., respondents may impute certain properties to certain levels that are not inherent in those levels, may also distort the claimed exterior levels identified by direct questioning. For example, a respondent may assume that a specific brand is expensive, heavy and/or traditional looking during the direct questioning (thus coloring his or her responses to those questions) but may change that opinion when shown an alternative that lists that brand with the attribute levels low price, light weight and stylish.

Additional research needs to be conducted to explore possible reasons for the high degree of inconsistency in respondent data between claimed and derived exterior levels.

RESULTS-SECOND STUDY

Table 2 shows the attribute relative importance scores for the full-levels stage and for the two-levels stage. There are differences in relative importance, particularly for psychological price point. It is suspected that differences in attribute relative importance scores might have been more dramatic if the variance in number of levels across attributes had been greater.

Table 2

**Attribute Relative Importance:
Complete and Two-level**

	Mean Relative Importance	
Attribute	Full-levels n=79	Two-levels n=79
Price	13%	14%
Brand	53%	51%
Psychological Price Point	6%	10%
Visual Design	17%	14%
Weight	11%	11%

However, the two-level design does not provide information about all of the attribute levels that may be of interest to management.

If, for each respondent, his/her utility weights for an attribute with three or more levels are linearly scaled into his/her utility weights for the same attribute with two levels, then attribute relative importance is maintained as well as level importance within attribute.

Table 3 shows the utility weights for attribute levels for the second conjoint exercise (full-levels) and attribute levels for the second conjoint exercise rescaled to have the same attribute relative importance scores as the attributes from the two-levels conjoint. The relationship between levels within attribute from the full-levels stage (stage 2) are preserved while the attribute relative importance scores from the two-levels stage (stage 1) are also preserved.

Table 3

Original and Rescaled Attribute Level Utilities

Attribute	Original Full-levels (n=79)	Rescaled Full-levels (n=79)
Price	-.00287	-.00274
Brand		
A	-.461	-.479
B	.657	.724
C	-.221	-.383
D	.024	.186
Psych. Pt.	-.088	-.026
Design		
A	-.009	.036
B	-.06	-.091
C	.069	.125
Weight	-.067	-.002

Alternative Solutions to Respondent Inconsistency

To avoid the problems of respondent inconsistency, there are at least three possible alternatives. For aggregate models, one could conduct a full-levels conjoint exercise, calculate utility weights, identify exterior levels among aggregate, mean utilities, then conduct a subsequent two-level exercise with a fresh sample. It may be the case, however, that the problems of heterogeneity normally associated with aggregate models could affect the accuracy and usefulness of this approach.

For disaggregate models, one could conduct a full-levels conjoint exercise, calculate utility weights during the interview, identify exterior levels for each respondent, create an appropriate questionnaire (in real time), then conduct a subsequent two-level exercise with the same respondent. This approach is necessarily adaptive and would require some form of computer-assisted

interviewing. It also assumes that the psychological component of the number of levels effect is extremely short-term. This assumption would need to be tested before this alternative could be accepted.

Another alternative for disaggregate models would be to rescale the full-levels utilities into the two-levels utilities, regardless of whether or not the two-levels utilities are exterior. It is not clear that the resulting attribute relative importance scores would or would not accurately reflect the true two-levels importance scores, i.e., the attribute relative importance scores that would have been computed had all respondents been shown exterior levels in the two-levels exercise.

SUMMARY

The existence of both psychological and algorithmic components to the number of levels effect has been demonstrated in prior studies.

Here, we have demonstrated a potential solution to eliminate the number of levels effect regardless of its source. Given an appropriate data collection methodology, such as Web-based surveys, and a two trade-off study design, conjoint utilities can be estimated for all attributes in their original specifications as well as for all attributes redefined to the two level case. The original utility weights can be linearly scaled into the two-level utility weights to remove the number of levels effect and more accurately reflect attribute relative importance.

More work must be done, however, to increase the consistency between claimed exterior levels and derived exterior levels or to find an alternative way to identify exterior levels.

REFERENCES

Currim, I.S., C.B. Weinberg, D.R. Wittink (1981), "The Design of Subscription Programs for a Performing Arts Series," *Journal of Consumer Research,* 8 (June), 67-75.

Green, P.E., and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing,* 54 (October), 3-19.

Steenkamp, J.E.M., and D.R. Wittink (1994), "The Metric Quality of Full-Profile Judgments and the Number-of-Levels Effect in Conjoint Analysis," *International Journal of Research in Marketing,* Vol. 11, Num. 3 (June), 275-286.

Wittink, D. R., (1990), "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions," *1990 Advanced Research Techniques Forum Proceedings,* American Marketing Association.

Wittink, D. R., J. C. Huber, J. A. Fiedler, and R. L. Miller (1992), "The Magnitude of and an Explanation for the Number of Levels Effect in Conjoint Analysis," working paper, Cornell University (December).

Wittink, D. R., J. C. Huber, P. Zandan, R. M. Johnson (1992), "The Number of Levels Effect in Conjoint: Where Does It Come From and Can It Be Eliminated?," *1992 Sawtooth Software Conference Proceedings,* 355-364.

Wittink, D.R., L. Krishnamurthi, and D.J. Reibstein (1989), "The Effects of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters,* 1, 113-23.

COMMENT ON MCCULLOUGH

Dick R. Wittink[1]
Yale University

I am very pleased with Dick McCullough's research on the number-of-levels (NOL) effect in conjoint analysis. The biases in part worths due to this effect have been known to exist for about 20 years now. And, although we have made some progress on this issue, reasonable people disagree about the source(s) of the effect and about its resolution.

Almost independent of the source of the problem, the most obvious way to avoid the NOL problem is to use the same number of levels for all attributes. Indeed, it appears that in an increasing number of commercial applications the number of levels is constant across the attributes. This solution is consistent with the hypothesis that the effect has a "psychological" basis: respondents pay more attention to an attribute as its number of levels increases, ceteris paribus.

Before I delve into specific aspects of Dick McCullough's paper, I offer two caveats. One is that the NOL effect can be avoided if we use a fully self-explicated approach. Indeed, Srinivasan and Park (1997) find that in predicting actual MBA job choices, a self-explicated approach outperforms a customized full-profile approach.

The other caveat is that if the NOL effect in conjoint analysis has a "psychological" basis, it is likely that marketplace choice behavior is subject to a similar effect. If so, the solution is not to equalize the number of levels across the attributes, but to create a study design that replicates the marketplace conditions individual decision makers are expected to face, and for which we want to make predictions.

In this comment I address the following issues: (1) is the NOL effect large enough for us to be concerned about; (2) if it is large enough, and if we want to use full-profile conjoint, and if we do not want to use the same number of levels for all attributes, what adjustments can we make in the part worths?

IS THE NOL EFFECT LARGE?

The NOL effect was first documented in a study of choices of paired objects constructed from a tradeoff matrix conjoint design (Currim et al. 1981). These authors showed that for each matrix involving one two-level and one three-level attribute, the minimum derived importances were 0.2 (two-level) and 0.4 (three-level) and the maxima were 0.6 (two-level) and 0.8 (three-level). Based on this systematic difference, Currim et al. adjusted the results. They found that the attribute with the highest derived importance after the part worths were adjusted would have been only fourth out of six attributes based on the unadjusted part worths.

[1] Dick R. Wittink is the General George Rogers Clark Professor of Management and Marketing, Yale School of Management, Box 208200, New Haven, CT 06520-8200, e-mail: dick.wittink@yale.edu

Subsequent studies have shown the NOL effect to occur in all data collection methods (e.g., tradeoff matrix ranks, full-profile ranks and ratings, ACA, CBC) and all data analysis methods (e.g., OLS, MONANOVA, LINMAP). Thus, the NOL effect generalizes across all types of data collection methods (except for self-explicated approaches), all types of measurement scales (rank order, choices, ratings, magnitude estimation), and all estimation procedures. In addition, the magnitude of the effect is very large such that substantive conclusions about the relevance of individual attributes (holding the ranges of variation constant!) are strongly affected.

WHAT ADJUSTMENT CAN WE MAKE?

Assuming that a full-profile study is desired, and assuming that it is undesirable to use an equal number of levels for all attributes, can the NOL effect be eliminated? Dick McCullough's paper discusses approaches for rank-order preferences and ratings. For ranks, his solution is fairly straightforward. To illustrate, I use least squares on hypothetical ranks in a design with two attributes.

Suppose attribute A has 3 levels, denoted A_1, A_2 and A_3 (where $A_1 > A_2 > A_3$), and B has 2 levels, B_1 and B_2 (where $B_1 > B_2$). I show below two rank orders, one most favorable to attribute A, the other most favorable to B:

	Rank Order	
Object	A favored	B favored
A_1 B_1	1	1
A_1 B_2	2	4
A_2 B_1	3	2
A_2 B_2	4	5
A_3 B_1	5	3
A_3 B_2	6	6

A least-squares analysis of these ranks generates the following part worths:

	(Unequal Number of Levels) Part Worth	
Attribute-level	A favored	B favored
A_1	2	1
A_2	0	0
A_3	-2	-1
B_1	0.5	1.5
B_2	-0.5	-1.5

It is easy to see that when A is favored the relative importances are 0.8 and 0.2, whereas when B is favored these are 0.4 and 0.6. These are exactly the limits identified by Currim et al. (1981).

To eliminate the NOL effect, Dick McCullough proposes to delete the objects with A_2, the interior level for A, and to analyze the resulting ranks. Thus, if the 2 x 3 design used above is the desired one, and rank orders are obtained as shown for two different respondents, we would then create:

	Rank Order	
Object	A favored	B favored
A_1 B_1	1	1
A_1 B_2	2	3
A_3 B_1	3	2
A_3 B_2	4	4

A least-squares analysis of these ranks provides:

	(Equal Number of Levels) Part Worth	
Attribute-level	A favored	B favored
A_1	1	0.5
A_3	-1	-0.5
B_1	0.5	1
B_2	-0.5	-1

To obtain adjusted part worths for all attribute levels, including A_2, we could multiply all part worths for a given attribute by the ratio of the attribute's importance in the "equal number of levels" case over the corresponding importance in the "unequal number of levels" case. These ratios are 0.5 for A and 1.0 for B when A is favored, and 0.5 for A and 0.67 for B when B is favored. In this example, this rescaling is unnecessary, since the part worths for A_2 are zero in all cases. The resulting adjusted part worths for the desired design are:

	Adjusted Part Worths	
Attribute-level	A favored	B favored
A_1	1	0.5
A_2	0	0
A_3	-1	-0.5
B_1	0.5	1
B_2	-0.5	-1

Now it is useful to note that none of the solutions are precise. Indeed, there are many alternative sets of part worths that can reproduce the original ranks (as the numbers of attributes and levels increase, the likelihood of a unique solution to exist increases, under the assumption that a compensatory model applies). Certainly, the raw part worths provide perfect predictions of the original ranks. The adjusted part worths when attribute B is favored also perfectly reproduce the original ranks. However, when A is favored the adjusted part worths predict ties for A_1 B_2 and A_2 B_1 as well as for A_2 B_2 and A_3 B_1. Thus, even though the adjustment seems appealing, it is possible that the adjusted part worths predict the original ranks less well than the unadjusted part worths do. This is an issue that needs additional attention.

An appealing feature of the adjustment is that it can be done separately for each respondent. Of course, one needs to make sure that the degraded design still allows for the part worths to be reliably estimated. Thus it is important for both the degraded design (in which objects with interior levels are deleted) and the original design to satisfy the usual criteria we employ for the construction of full-profile conjoint designs. For example, the degraded design will necessarily increase the (statistical) uncertainty of the part worths.

For ratings the adjustment is much more complex. The reason is that we do not know what ratings a respondent would have given for a degraded design. Dick McCullough's solution is to use two experimentally equivalent groups of respondents, one evaluating profiles created from a design with equal numbers of levels for all attributes, the other evaluating profiles created from a design with the desired unequal numbers of levels.

Ideally one would then match two respondents, one from each of the two designs, who have identical preferences except for the difference in design characteristics. Such linkages would allow for exactly the same type of transformation of the part worths resulting from the design with equal numbers of levels. Of course, in practice it will be impossible to find pairs of respondents with identical preferences, a problem that is compounded by the NOL effect which applies only to one design.

One possible approach is to ask all respondents to provide self-explicated data as well. It should then be possible to align respondents, if not in a pairwise fashion then at least at a segment level. Appropriate transformations can be estimated, so that the part worths for the design with unequal numbers of level can be adjusted. Alternatively, it is possible to use respondent characteristics (including purchase behavior) to explain differences in part worths between respondents in the design with equal numbers of levels. These explanations can then be used to predict part worths for all levels in the design with unequal numbers of levels. A caveat associated with the latter approach is that the transformed part worths will have reduced heterogeneity between respondents. The amount of this reduction depends on the lack of explanatory power associated with the respondent characteristics.

I find that Dick McCullough's paper provides very sensible ideas for full-profile conjoint studies. Although his paper does not resolve the uncertainty about the source(s) of the NOL effect, other papers also fail to provide an unambiguous resolution (see, for example, Wittink and Seetharaman, 1999). However, the proposal to rescale the part worths obtained from a desired conjoint study design, based on results for a design with equal numbers of levels, seems attractive, and I believe worth considering if there are good reasons not to impose the constraint of equal numbers of levels on the study design. It remains to be shown that the degraded designs provide sufficient precision in the part worths. In addition, it is important to demonstrate that the adjusted part worths do not reduce the correspondence between predicted and original preferences, and that self-explicated importances can be obtained in such a way that there is acceptable consistency with the inferences from the full-profile study.

REFERENCES

Currim, Imran S., Charles B. Weinberg and Dick R. Wittink (1981), "Design of Subscription Programs for a Performing Arts Series," Journal of Consumer Research, 8 (June), 67-75.

Srinivasan, V. and Chan Su Park (1997), "Surprising Robustness of the Self-Explicated Approach to Customer Preference Structure Measurement," Journal of Marketing Research, 34 (May), 286-291.

Wittink, Dick R. and P. B. Seetharaman (1999), "A Comparison of Alternative Solutions to the Number-of-Levels Effect," Sawtooth Conference Proceedings, this volume.

MATCHING CANDIDATES WITH JOB OPENINGS USING WEB-BASED ADAPTIVE CONJOINT

Man Jit Singh
Futurestep
Sam Kingsley
dataDirect

INTRODUCTION

The widespread availability of Web access has led to re-evaluation of traditional relationships between companies and their customers. Industries which have a heavy information component, such as news and financial services, are experiencing large shifts to online delivery.

The business logic of online job search and recruiting for middle-management and professional positions is that the Internet allows Futurestep to use sophisticated assessment methodologies effectively with a large database of management candidates. Candidates, including those who may not be actively looking for career changes, register with Futurestep because of the personal information and feedback they receive. Recruiting professionals manage the process with employers, particularly the job specification step, preparation of an offer and negotiation.

Futurestep was the first company to use ACQNET, a Web adaptation of Sawtooth Software's ACA interviewing module. Over 70,000 ACQNET interviews have been completed to date. In the Futurestep process, direct interview questions are used to identify likely candidates for potential jobs. Structured questionnaires also assess decision-making and communication styles, and career motivators. The ACA results are used to identify potential candidates in the early stages of a search by eliminating those who would be unlikely to accept offers within the hiring company's set ranges: Is this candidate ready to move? Which job factors are the most important to this candidate? The ACA results are also used in the final stages of a search to provide a client company with information about how to tailor an offer that appeals to a candidate's preferences.

Since the analysis of conjoint results is conducted at an individual level, minor adjustments were necessary in the ACA interviewing algorithm and analysis.

TRADITIONAL RECRUITING

Whether traditional or not, recruiting firms generally follow three stages in conducting a search:

- FINDING means sourcing a large pool of potential candidates. It is usually done by industry researchers. As much as 50% of a firm's resources may be devoted to this stage, which is repeated for each new search conducted.

- ASSESSMENT of candidates means reviewing the pool and understanding the top candidates beyond their objective experience levels. This gives the recruiter a chance to probe and test the way candidates handle themselves in the context of the client's needs.
- PLACEMENT of candidates includes articulating to both the candidate and employer why this is a good match, plus negotiating an offer.

The job search and recruiting business has traditionally relied on a "telephone tree" of Researchers to source potential candidates for job openings. This concentrates over half of the resources of a firm on simply obtaining a list of potentially qualified candidates. In a typical recruiting firm, dozens of Researchers may contact hundreds of candidates by telephone for a given search. Many of the contacts may be wasted, because they surface neither a potential candidate nor a referral to someone else in the telephone tree.

```
                 Top Management
                        /   \
               Ptr1             Ptr2
               /   \
        Assoc.          Assoc.
         /   \
   Rsch           Rsch
```

Researchers identify a large pool of potential candidates with qualifications that match the job specifications. The particular skills and experiences required for the job are assessed at this stage.

Next, in the traditional firm, Senior Associates or Partners review the qualifications and select a small number, perhaps 2-6 candidates, for additional interviewing. This is a time-consuming process that usually involves a personal contact with each candidate.

Finally, one or more candidates are presented to the employer, who will often interview several of the candidates. At this stage, the recruiting firm seeks to solidify the match: both sides should be convinced of the appeal. A specific offer will be made, usually with input from the recruiting firm. The inefficiency at this stage occurs when there are protracted, iterative rounds of interviews with client managers or rounds of offer negotiations and/or a candidate ultimately declines the offer—a "slow-no."

FUTURESTEP RECRUITING: PHASE 1—FINDING CANDIDATES

To reduce the inefficiency of the Finding step, a large database of qualified candidates was developed. Futurestep launched a nation-wide radio and print advertising campaign. The Wall Street Journal is an important partner with Futurestep because of its recognition among potential candidates and clients. Within a month there were over 50,000 registrants. These individuals are generally mid-career managers and higher.

Why do candidates come to Futurestep? Because they receive:

- Complimentary career and market value feedback

- Guaranteed confidentiality of candidate's information
- Credibility of Korn/Ferry International and the Wall Street Journal
- Consideration for exclusive, non-advertised positions from blue chip organizations as well as smaller, high-growth firms.

Some of the feedback from candidates illustrates their level of involvement with the process.

> "This whole process ... provides an opportunity to establish a personal relationship in a high-tech world. The business world rarely sees technology as an enabler and technology rarely sees beyond its owns bits & bytes."
>
> "Thank you! What a great way to get to know your business style. The more I know, the better I can communicate my needs in a job interview."
>
> "This is an outstanding and leading edge tool for career development."

Futurestep is part of a sea of change in the business of searching for jobs (stated from the employee's perspective) and hiring (employer's perspective). It is changing because of the amount of information available to candidates and employers. As we have seen in other markets, such as home mortgages and foreign exchange, when more information is made available in a standardized format to both buyers and sellers, "liquidity" improves due to a greater transparency in pricing. Transaction costs decrease and there is an overall greater level of transactions.

Also, as uncertainty is diminished due to greater transparency of information, the decision process is shortened. Overall, the effects on the employment marketplace may eventually be as large as the introduction of employment advertising!

FUTURESTEP RECRUITING: PHASE 2—ASSESSMENT OF CANDIDATES

In both the traditional and Futurestep approaches there is a requirement to verify the information that has been submitted and check that the candidate in-person matches the description on paper. Futurestep uses a proprietary video-conferencing system to permit recruiters to interview potential candidates in their own homes. To keep the comparison between candidates more consistent, the questions are asked from a common script. In fact, the interviews are recorded and can be used in future searches as well. Once Futurestep has selected one or more candidates to present to the employer, a computer CD is prepared that contains the recorded assessment interviews. In addition to the ACA data collected, Futurestep uses a battery of validated self-assessment instruments to better match personal preferences in decision-making and managerial styles. This data is matched with information provided by hiring managers and potential peers of the candidate who fill out a set of “mirror” instruments prior to the search.

FUTURESTEP RECRUITING: PHASE 3—PLACEMENT

Having identified several potential candidates for a search—usually 2-3 individuals—the next challenge for the recruiting firm is to market the job to the candidate, and vice versa.

- The wealth of standardized information available to the recruiter and employer about each candidate's approach to work environments gives the employer a clear basis for selecting the top candidate.
- Conjoint utilities provide the values that the candidate places on the most important aspects of the job. This helps the recruiter and employer in several ways:
  - Examining the utility curves helps recruiters understand what characteristics of a job might be very attractive, or are potential sticking points.
  - Using calibration concept results, the recruiter gets an idea of the candidate's willingness to move for any job. Is this a "tough nut" or an "easy sell"? While the really tough nuts (those with little likelihood of moving) have already been selected out, there may be some borderline candidates left in the pool because of their clearly superior abilities.
  - Since a given job may represent different levels to different candidates, a tailored "job spec" must be created for each candidate in order to calculate likelihood of accepting the job. At present, this is a manual process that is only used for the small number of finalists in a search. An input screen for Job X for Candidate Y has been designed to permit the recruiter to estimate what the job represents to that candidate, and then to adjust the input values to see what's driving the job's appeal: what ifs.
  - The attributes contain both annual compensation and wealth creation (stock or options). Therefore, it is possible to tailor an offer by shifting the salary/wealth balance to maximize the value to the candidate, within parameters established by the employer.

Having conjoint results also lets the recruiter highlight to the candidate those factors of the job that would be most important to him/her, rather than spending time on unimportant characteristics.

Overall, another feature of having the conjoint data collected in advance is that candidates' answers aren't tailored to their perceptions of a particular job opening. It serves as a natural check on the enthusiasm with which candidates, and sometimes recruiters, approach a search.

When combined with the other structured questionnaires, it gives the candidate and recruiter a richer language with which to discuss a particular job as well as the candidate’s needs and values. Since this type of standardized data collection is not typical in the recruitment industry, training for Futurestep recruiters has been built into the process.

HOW HAVE CANDIDATES REACTED TO THE CONJOINT INTERVIEW?

The Futurestep conjoint design consists of 11 attributes with a total of 40 levels. The interview itself asks for respondent input just as it would in ACA Version 4 for DOS:

Except for one attribute, all preference rankings are set *a priori*.

Each attribute is rated on importance using Sawtooth’s suggested 4-point scale.

15 pairs are presented, 10 with 2 attributes and 5 with 3 attributes each.

4 calibration concepts are used.

Comments from candidates who completed the conjoint portion show that they are very involved in this part of the process, and generally find it an accurate reflection of how they would make choices. There is little drop-off in participation at this stage. The conjoint is the last portion of the registration, so candidates have generally invested over 45 minutes up to this point. Although exact statistics aren't available, over 90% of respondents who get to the end of the prior section complete the conjoint part too.

The comments below highlight the need for more complete explanations of the conjoint process than might be required in a "normal" market research interview.

> "The characteristics described above are the person I am. I did not believe the questions asked could reveal so much. This indeed is very interesting and exciting. Thank you."
>
> "Thanks for the Desired Job Characteristics—a good reflection in my case EXCEPT: major event = too high and making an impact = too low; probably more like 8 and 8 respectively. Thanks again."
>
> "It was surprising and revealing to me that my career questionnaire feedback was in accordance with my career experiences to date. I was very impressed that my job characteristics (as defined by your assessment) is almost a perfect match with the job that I am doing. It was also heartening to read about the job characteristics of positions that I am applying for actually dovetail with my desired job characteristics. As a result, I would say that your evaluations are 'on the money.' Thank you very much."
>
> "I essentially agree with the results and explanations associated with each area except in the section "Desired Job Characteristics." I am disappointed in that my responses indicate that I favor personal wealth twice as much as making an impact. Although achieving wealth is important, I believe that it should follow making a positive impact, especially in a new position. I guess I didn't answer the questions with this in mind. Is it possible to change my responses in order to make this point? I found this to be a very enlightening exercise."
>
> "Very pleased. The analysis "felt" right on target. It also brought up a career option/class that I didn't know existed, which matched really well with the characteristics of my ideal job."
>
> "I really enjoyed the exercise—I think the results are right on and consistent with other profiling techniques I have experienced. In this case, I think the attributes in the "job characteristics" report were well chosen—had to be via focus groups." [Editor: No focus groups.]
>
> "I have some [Job Characteristics] that are different by only 1% or 2%. Are these statistically significant differences?"
>
> "I would like to comment on the "relocation" weighting in the job characteristics section. Because I live within five miles of Chicago, and am willing to work at locations within a 20 mile radius of one of the major employment centers in the country, my reluctance to relocate should not be weighted in the same way as someone who lives in a much less urban area. Does your analysis take geographic location of the respondent into consideration?" [Editor: Yes it does.]

> "I somewhat disagree with results of Desired Job Characteristics. That may be how my answers fell out, so to speak, however if I obtain and successfully perform a job that provides more opportunity and responsibility, I, personally, would only assume that I will impact the performance results and the bottom line, in a positive manner, significantly."
>
> "Absolutely amazing! You have drawn a very accurate picture of my personality, characteristics and personal goals."
>
> "It was a pleasant experience registering to Korn/Ferry's Futurestep site on the Internet. I was specially impressed by the immediate "Desired Job Characteristics" feedback upon conclusion of the registration process."

CHANGES REQUIRED OF ACA DATA COLLECTION

Web Delay

In the DOS version of ACA, calculations can be made after each entry by the respondent. Even on an original IBM PC (8088), there is an insignificant delay as ACA does its computations.

Since ACQNET has avoided client-side execution, the results of each page must be sent to the server for processing, no matter how trivial. This imposes a transmission delay for Web-based ACA of approximately 3 seconds. There are two places in ACA which this affects:

- During the priors section, ACA uses either a preference rating or ranking for levels within attributes. In the DOS version, preference rankings are simplified for the respondent by use of a disappearing list. After a choice is made, that answer is removed from the screen. This makes it possible to ask the respondent to rank order a large number of levels.

  To do this on the Web without client-side execution would require a page to be submitted and returned for each judgment. Instead, ACQNET implements preference rankings by asking the respondent to put the rank order (1 to n) into a text box. Rankings on all levels of all attributes are obtained before the page is submitted. While ACQNET allows the researcher to specify either preference rankings or ratings, ratings are recommended if there are more than 3 levels per attribute.

- The pairs section presents a different kind of problem. In DOS ACA, utilities are re-calculated after the respondent enters a judgment on each pair, which can affect the left-to-right balancing of the next pair. However, doing re-calculations during the pairs section requires a balance between the respondent's patience with delays and the additional accuracy obtained if ACA does frequent re-calculations.

- ACQNET addresses the need to re-calculate utilities during the pairs without stretching the patience of the respondent. It does so by presenting a variable number of pairs on one Web page — currently 3 at a time. When the respondent submits a page of answers to pairs, ACQNET re-calculates the utilities and selects the next set of pairs to present.

Reversal of Preferences From Priors

There is another characteristic of ACA that was not particularly suited to individual level analysis. Preferences for levels within attributes are either set *a priori* (e.g., higher cost is always less appealing than lower cost) or asked in the priors questions (“Which do you prefer?”). Answers to the pairs questions, however, can actually lead to utilities that contradict the ordering of preferences from the priors section. In the DOS version, this is permitted and the final utilities sometimes contain reversals. Our experience with aggregate-level analysis of ACA is that such reversals have no measurable effect on simulation results.

In individual level analysis, however, such reversals can become confusing to respondents and analysts. When the utilities for two levels are reversed in preference, ACA will not properly balance a pair that contains those levels. This means that the respondent could be presented with the following type of choice:

Annual comp: no change	Annual comp: 20% increase
Wealth creation: $0	Wealth creation: $500K stock options

In this situation, some respondents pause and ask of themselves (or an interviewer, if present) “There’s no trade-off here. What should I do?”

A preliminary analysis of two previously collected ACA data sets was performed to determine how often a pair was presented in which either the Left or Right side contained all the preferred levels. In one ACA study there were practically no reversals: 4 out of a total of 2,595 pairs presented to 145 respondents contained no trade-off between left and right. However in another study there were 102 non-trade-off pairs out of 1,918 pairs judged by 100 respondents. In other words, 5% of the pairs involved decisions where there was no trade-off.

Additionally, respondents seem to use the extreme scale values when rating pairs where there is no trade-off. Out of the 106 reversals noted above, 74% of the pairs involved were rated either 1, 2, 8 or 9. This may indicate that respondents were rating their <u>certainty</u> of preference rather than the strength of preference.

Further analysis is being considered to determine the situations that are likely to produce preference reversals.

For the Futurestep application, a decision was made that the preference rank orders from the priors section were to be accepted. Of the 11 Futurestep attributes, 10 were rank ordered *a priori* (e.g., more salary is always preferred to less). There was one attribute that required the respondent to make a judgment: whether relocation to a more desirable location was preferred to no relocation. For this application, it was determined that the respondent’s judgments of preferences during the priors would be more accurate than preferences derived from pairs.

ACQNET contains an option to restrict the judgments in the pairs section from reversing the ordering of preferences for levels from the priors. If a reversal is found, the utilities for the two levels are averaged, then the tie is broken by adding .01 to the utility for the level that was preferred in the priors.

WHAT LIES AHEAD?

- Futurestep isn't the answer for all businesses.

  Small business employers are least likely to have well-developed job specifications and are also quite cost sensitive. Their adoption of this method for staffing would require an increased level of formalization for the business as well as lower cost delivery by the recruiting firm.

The Futurestep model works well where there is a well-defined skill set or product knowledge required. If a job requires a high degree of personal fit, individualized assessment or personal relationships for hiring, the standardized Futurestep approach is unlikely to be adopted. Examples would be employment by a movie studio, creative fields such as design or advertising, or any job that requires an audition.

- Futurestep isn't for all candidates, but since online job searching significantly expands a candidate's opportunities, it is expected to be adopted with increasing frequency. Candidates will still use multiple methods of job searching, including networking, classifieds and Internet job postings.
- Further development of tools and training for recruiters in the "language of conjoint" will be required. Currently under development with Futurestep is a method to specify jobs in terms of the attributes and levels for each potential candidate. This is starting as a subjective process with a long-term goal of systematizing this analytic step.
- There is also an opportunity to better understand employer's trade-offs for a candidate: what would drive a greater dollar offer, or other benefits? What would make a candidate exceptional — such that an employer would willingly pay more? ACA has the ability to make the dialogue between recruiter and employer more systematic and less based on assumptions. As conjoint has forced researchers to be explicit about product attributes, so would employers communicate their trade-offs with greater accuracy.
- What's ahead for ACA on the Web? We see several trends:
  - More access by researchers and non-researchers.
  - Online design and testing of ACA via the Web is already available. Prototyping and testing can be greatly simplified.
  - Humongous data sets can be developed, increasing the opportunities for subgroup and segmentation analyses.
  - When conjoint is used with individual feedback, it places greater demands on researchers to provide clear and meaningful explanations of individual conjoint results

PREDICTING PRODUCT REGISTRATION CARD RESPONSE RATES WITH CONJOINT ANALYSIS—AN EXPERIMENT IN DATA COLLECTION

Paul Wollerman
The Polk Company

ABSTRACT

Many manufacturers of consumer durables pack "warranty cards" in their products. Instruments that produce high rates of return are, of course, most desirable. This paper describes a study designed to test whether conjoint analysis, as opposed to packing test cards in products, could be used to collect data which would help us eventually to build a tool that predicts response rates for product registration questionnaires based on their attributes.

Sawtooth Software's CVA was used to help design the study and analyze the results. Part of the purpose of this test was to gauge CVA's efficacy in building the desired tool.

INTRODUCTION

Over the last twelve years, we have conducted many "live"[1] tests to measure response rate for our product registration cards. The objective of these tests has been to measure the effect of design features (like number of questions, font size, and layout), and incentives on response and completion rates. However, implementing these tests has always proved to be logistically difficult; plus, it takes a long time to obtain usable results. A more convenient and timely way to collect data was sought.

Since a product registration card can be viewed as being composed of several different attributes, we proposed using conjoint analysis to measure consumers' preferences for those attributes. Survey respondents would be presented with series of scenarios in which they have just bought some consumer product. They would then be asked to rate how likely they were to mail back the (particular) registration card that came with it. However, since this was a completely untested idea for product registration card research, we wanted to perform a "trial run" to gauge the efficacy of it.

We believe this research to be fairly unique on two levels: The actual method of implementation for the conjoint analysis and, this is the only case we know of that attempts to use conjoint analysis to aid in the design of survey instruments.

This paper focuses on describing this conjoint study design and its implementation, and giving a high-level comparison of the data obtained by this study versus the information we have

[1] By "live" test, we mean actual measurement of response rate for a particular card in a particular product. This is accomplished by printing 10,000 cards of the experimental design and distributing them in the product, then counting the returns. Thus, 500 returns is a 5% response rate.

gathered through our "card test" program. We also discuss the role and performance of the CVA software in this research. It is not the intention of this paper to give a detailed, conclusive analysis of all the results of our evaluation, as that phase of this project is not yet complete.

OBJECTIVES

The long-term goal of our product registration card research program is to build a robust statistical model which can predict response rates for card designs based on their attributes. For a product registration card, attributes are the price and type of product in which it is packed, its length (number of questions), the return incentive, font sizes, black & white vs. color, etc.

The abilities of conjoint analysis (and conjoint software) were seen as a way to obtain the necessary data to build the desired modeling tool without the headaches of our usual "card tests" (which could also take many years).

The immediate goal of this particular study was to test the efficacy of a conjoint approach to data collection and analysis as a means to replace our current method of data collection - distributing experimental card designs in actual products. Integral to this process would be evaluation of the conjoint software chosen for this project (CVA).

Given our goals, we felt it was important to keep the test design as simple and straightforward as possible, in order to facilitate our understanding of the data, and the eventual explanation of the project and its methods to our internal sponsors and external clients.

METHODOLOGY

From the outset, we had intended to use a "full-profile" conjoint analysis approach because the research involved a relatively small number of attributes, each with a fairly small number of levels. Hence our choice of CVA as the design and analysis software.

The basic survey method was to mail a packet of materials to members of a consumer mail panel. This packet consisted of a survey instrument, a set of "mocked-up" product registration cards, and pictures of the products in which the cards were supposedly packed. The questionnaire asked each respondent to pretend, product by product, that he or she had just purchased the indicated item and then rate the indicated card as to how likely he or she would be to complete and return the card.

This design and the scale used follow a basic "single-concept" or "card-sort" conjoint design. This was used largely because it seemed to best mimic the actual card return process: a consumer sees only one card in their package, and makes a decision about whether to return it. This was also very compatible with Sawtooth Software's CVA product, which was used to help design the study and perform part of the analysis after data collection. In particular, we wanted to use the CVA simulator's "purchase likelihood model" to build our final "response rate modeling" software tool.

The use of a consumer mail panel was seen as a cost-effective way to achieve a high response rate (usually about 70% for a mail panel). It was recognized at the time that mail panel members might be somewhat more inclined to return product registration cards generally, but it was believed that that could be factored in to the results.

Study Design and Execution

Given that the object was to gauge how well the method would work both for comparing the data gathered to past "live" tests and for future use, several main parameters were set:

- To give good comparability between previously measured response rates and survey results, we tried to include in the study a maximum number of cards that had been tested in products. We found seven such "calibration" cards.
- Each respondent was asked to complete about ten different tasks (in this case, rate a card design for the product described). Ten tasks was deemed about right, given that the probable size of future tests (much larger than this) would dictate that the number of tasks be at least ten to fifteen.
- Each task a respondent saw was a registration card for a <u>different</u> product. This was in order to mimic the actual card return scenario as much as possible: consumers see only one card per product they buy, and the choice they have is to return it or not.

This last point meant that in order to test two cards for the same product (as we had in actual tests) it was necessary to use more than one group of respondents; each group was sent a different set of cards for the same ten products. In conjoint analysis terms, this is a "block design".

The salient attributes of the product registration cards were defined as:

1. The price point and type of product in which the card is packed - Ten levels were used, ranging from $22 Toaster to $900 Refrigerator (see the Study Design in the Appendix for a complete listing).
2. Length - Two levels: <u>Long</u> and <u>Short</u> (see Appendix for explanation).
3. Incentive - Four levels: "Consumer Protection Text" (see Appendix for explanation), a $100,000 cash sweepstakes, and two different extended warranty offers - "Warranty A" and "Warranty B".

These design parameters were chosen largely because they corresponded to the cards that had been previously tested in products. They also represent common features being used in many existing cards.

Once these were defined, we used the CVA software to develop a test design with thirty tasks, ten each for three groups of respondents. Those tasks included the seven cards that previously were tested in products (see the Appendix for the full design). No "A Priori" order was specified for the levels of the attributes, and no pairs of attributes were prohibited. CVA's design had to be modified slightly in order to include all seven of the "calibration" cards. CVA gave the resulting design a D-Efficiency of 92.6.

We chose to mail the survey packages to three groups of 200 mail panel members each, for a total of 600. Each of those groups was demographically balanced to the US and pre-selected to be households which had bought two or more of a wide range of consumer durables in the last six months, and thus would (probably) have seen such registration materials recently.

All cards were created based on our clients' existing cards, using their artwork and logos, to maximize the "reality" of the test. The seven cards that had had their response rates measured via distribution in actual products were copied exactly. For the rest of the cards, the levels of the attributes listed above were standardized, as much as possible, to mimic those of the seven "calibration" cards. Because of the variety of product types used, some slight differences existed in exactly how "Long Card", or "Consumer Protection Text" was implemented across different "products".

Pictures of the products were digitized and printed to include as part of the survey packages.

Both the cards and the pictures of the products were printed in black and white.

The survey instrument consisted of both sides of an 8.5"x14" page, with instructions and ten questions. For each of the ten questions (one for each product/card), the questionnaire asked the respondent to look at the corresponding picture and imagine that he or she had just purchased or been given the item and had found the corresponding product registration card with it. They were then asked to rate how likely they were to return the card. They were instructed to use a scale of 0 to100, where 0 means "definitely would not return the card", and 100 means "definitely would return the card". A partial copy of the questionnaire is attached in the appendix.

Data Processing

The survey packages were distributed in mid-November 1997, with returned questionnaires received in December 1997.

After data entry, the rating scores for each card were averaged across all respondents and the resulting number was interpreted as a likely response rate for that card. These numbers are reported below as the "Average Scores Measured by Survey".

Additionally, the rating scores for each card for each respondent were fed into the CVA software, where utilities were calculated for each level of each attribute (card design element). These were then used by the software to calculate a "purchase likelihood" model. This gives a number from 0 to 100 which we again interpreted as a likely response rate for that card. Those numbers are reported below as the "Purchase Likelihood calculated by CVA software".

It is important to note that, because this was a "block design", for each group of respondents, there were missing data values corresponding to the tasks that that group did not see. In order for CVA to accept the source data file, a score for each task had to exist, with only a fairly small number of missing values per record allowed. In order to accommodate this, we filled in the missing values for each groups' scores with mean scores from the other groups, within six demographic segments.

RESULTS

The overall response to the survey was 77%. This was as expected: mail panel response is advertised as "typically 70%".

We will concentrate our discussion of the data obtained for the seven cards which had been tested in products, as these are those on which we hoped to "calibrate". Results for all cards tested are given in Table D in the appendix.

At our basic level of analysis, we had two main areas of interest for comparison. We wanted to compare the raw survey average scores to our measured response rates in order to measure the effects of the sampling frame and the scale. We also paid close attention to how the raw survey average scores compared to CVA's Purchase Likelihood. The reason for this last was to help gauge how well CVA would predict response rates for cards not yet tested: if its "base case" scenario could mimic the survey data well, then we would feel confident that it was effective in generating accurate Purchase Likelihood scores for card designs that were not included in the study (e.g. - a long card with the $100K Sweeps incentive, packed in the $900 refrigerator). Also, if the survey data proved unusable, we might still be able to use conjoint analysis with actual response rate data from "live" tests as input to CVA.

Table A: Results from Product Tests and the Survey

	A	**B**	**C**
Card Tested	**Actual Response Rate Measured in Product Test**	**Average "Likelihood of Return" Scores Measured by Survey**	**Purchase Likelihood Calculated by CVA**
$900 Refrigerator Long, Consumer Protection Text	41%	86.8%	84.4%
$900 Refrigerator Short, Consumer Protection Text	42.43%	93.5%	91.37%
$320 Color TV Long, Consumer Protection Text	25.27%	83.42%	82.13%
$320 Color TV Short, Consumer Protection Text	28.56%	90.08%	89.34%
$35 Beard Trimmer Long, Consumer Protection Text	3.22%	56.95%	63.36%
$35 Beard Trimmer Long, $100K Sweeps incentive	5.89%	70.19%	68.61%
$22 Toaster Long, No incentive	5.28%	NA	NA
$22 Toaster Long, $100K Sweeps incentive	5.84%	73.4%	70.48%

Looking at Table A, we see that:

- If one considers the data at the "product" level - Toaster, TV, Beard Trimmer - the rank order of the average estimations of card return likelihood given by the respondents (Column B), roughly follows the observed "live" test results (Column A). However, a Wilcoxon Signed Ranks test indicates that the data in Columns A and B are significantly different.

- The CVA Purchase Likelihood numbers (Column C) are a very good fit to the survey data (Column B), as expected. In fact, if we consider the entire set of data from Table D (Appendix), a Paired Samples T-Test for correlated data shows no significant difference between the two columns: $t < 1.0$, $p > .05$. Further support for this is shown by observing the high degree of correlation: $r = .954$, $p < .001$.

- The survey's average scores, compared to what we have measured via actual product tests, are very high. Further, the over-estimations are much larger for low-priced items than for the higher priced items.

It is this last point that indicates how the data collection method influenced the results. We expected that the survey respondents would over-estimate their own likelihood of returning the cards and we were prepared to compensate for it. However, the fact that the degree of over-estimation was not more uniform across all the cards came as something of a surprise. This finding leads to the next observation.

When considering response rates, we usually speak about differences between two cards in terms of "relative percent". That is, if we consider the long versus the short card packed in the $900 refrigerator, there is a difference of 1.4 percentage points in their actual, measured response rates. That, in turn, is a 3.49% difference relative to the long (lower response rate) card.

We had decided beforehand that an important measure to watch in this study would be whether the relative differences between cards as measured by the survey data would be approximately the same as the relative differences between the same cards as measured in product testing.

Table B: Relative Differences Between Cards Tested

Card Pairs Compared	Relative Response Rate Difference Measured in Product Test	Relative Difference in Average Scores Measured by Survey	Relative Difference in Purchase Likelihood Calculated by CVA
$900 Refrigerator Short Card vs. Long Card	3.49%	7.72%	8.26%
$320 Color TV Short Card vs. Long Card	13.02%	7.98%	8.78%
$35 Beard Trimmer Consumer Protection Text vs. $100K incentive	82.9%	23.25%	8.29%

Table B shows that the relative differences between cards measured both in products and the survey are better for some cards than others, but overall are not very comparable. This, of course, is reflective of the inconsistency of the over-estimation shown in Table A. It is interesting to note that CVA has tended to "flatten out" these differences.

Perhaps the most interesting data obtained from this study comes from looking at CVA's utility and importance scores for each of the attributes. Not only does the data agree quite well with our "live" measurements, it tells us some things we did not know but which, in retrospect, make good sense.

Table C-1: Average Importance Calculated by CVA for each Attribute

Attribute	Average Importance of Attribute Calculated by CVA
Product/Price	85.3
Questionnaire Length	5.5
Incentive	9.2

Table C-2: Average Utility Values Calculated by CVA for each Level of each Attribute

Attribute #1 Product/Price	Average Utility	Attribute #2 Questionnaire Length	Average Utility
$22 Toaster	12	Long Card	0
$35 Beard Trimmer	10	Short Card	15
$50 Toaster Oven *****	13		
$180 Ready-To-Assemble (RTA) Entertainment Center *****	7	**Attribute #3 - Incentive**	
$190 VCR *****	33	Consumer Protection Text	9
$350 Gas Grill	28	$100K Sweeps	14
$320 Color TV	32	Extended Warranty A *****	15
$400 Stereo *****	28	Extended Warranty B *****	12
$750 Washer & Dryer *****	36		
$900 Refrigerator	36		

***** Item not yet tested via distribution in an actual product

Historical data from our twelve years of testing cards in products indicates that the type and price of the product in which a product registration card is packed has up to eight times as much influence on response rate as any card design feature. We have also seen that questionnaire length has less influence on response rate than other design features. Although not definitive, the order and magnitude of the average importance for the attributes given in Table C-1 agree well with our historical data.

Even though we do not have definitive numeric data to compare with the data in Table C-2, we believe that the utility scores for the card attribute (design elements) are generally a very good match with what we have learned from our "live" response rate measurement tests. That is, "toaster" has a much lower utility than "color TV", "short card" has a higher utility than "long one", the $100K Sweepstakes incentive has a higher utility than "Consumer Protection Text", and so on. Even better, the data contain some subtleties that conform with our experience and theories.

For example, the utilities indicate that a card packed in the toaster should get a slightly higher response rate than a similar one packed in the beard trimmer. Based on the difference in price, we expected the reverse, but in our product test, our "control card" (a generic one with no incentive) had a higher response for the toaster (about a 5% response rate) than the beard trimmer (about a 3% response rate). At the same time, the $180 ready-to-assemble entertainment center has the lowest utility of all, even though it is more expensive than some of the other items. Even though we have not tested a card in this particular item, this result makes sense (it is a piece of furniture), and is consistent with our experience with similar types of products. Similarly, the $190 VCR has a utility in the same category as the $300 TV, even though the price is much less. This probably reflects the perceived complexity and/or fragility of the item, which are components of the "product type/price" attribute.

CONCLUSIONS

Clearly, the results so far are somewhat mixed. On the one hand, we see that the respondents have rated the cards in such a fashion that the rank order and magnitude of the attributes' importance and the utility scores of the individual attributes are a good fit with our experience.

On the other hand, the respondents not only grossly over-estimated the likelihood of response compared to our experience, but have tended to gravitate toward the upper portion or center of the scale. The result of this is that we have so far been unable to find a "calibration function" which would map our survey data back to our response rate measurement data.

Also, the CVA software performed well to the extent we used it. It was a good design and analysis tool for this study. Although we have not yet finished our analysis, we hope to use and evaluate CVA's more advanced capabilities as we progress.

Although we have much more analysis to do, we feel that an accurate high-level assessment of the study is summed up by the following:

- We generated a large amount of useful information and learned a lot.
- The raw survey data gathered probably cannot be used accurately to predict response rates for a given card, or accurately to predict differences in response rate between cards.
- Conjoint analysis can be used to measure consumers' preferences for product registration card design features, and predict how a given card will perform relative to another.
- In order to finally determine how well using survey methods for collecting data will allow us to predict actual response rates for product registration cards, more study and analysis is needed.

These findings point out the importance of the method of data collection in any research study. Let us briefly examine several aspects of the way we collected the data and how we might have done them differently (or do them in the future).

First is the scale used. In the design phase, using the 0 to 100 rating scale seemed like the most straightforward one, and was consistent with usual conjoint practices. It is possible that use of other types of scales and different ways of asking the questions could give better results.

Also, there is the possibility that this study had some measure of the "number of levels" effect. We had ten levels of the Product/Price attribute and only two to four levels for the other attributes. Perhaps using fewer levels for Product/Price, and thereby having fewer tasks for the respondents to perform would have resulted in more accurate differentiation between products.

Then there is the sample base. We speculated from the start that the mail panel members might be somewhat pre-disposed to returning product registration cards compared to the general populace, but past studies show little evidence to support this. Thus, we felt (and still do) that only a small amount of over-estimation on their part due to this. However, it would be very interesting to repeat this study using a different sampling frame.

The results of this study compelled us to spend a fair amount of time reexamining the data from our response rate measurement tests[2]. As we did this, one thing became apparent: the attribute that comprises the product itself - its price and nature - seems to sometimes combine with the others to form a “compound attribute”. This is evidenced by observing that the relative differences in response rate - as measured in actual product testing - due to questionnaire length or incentive or some other design element were sometimes different for different products. Therefore, it will be necessary to do more “live” testing to understand how all these relationships between attributes behave. Although it is exactly what we were trying to find an alternate method of doing, we could take this same conjoint design and pack real cards in real products, and thereby obtain actual “real-world” data.

Finally, we still believe that conjoint analysis will be the most efficacious technique for generating the necessary information to build the predictive modeling tool we desire. It may be that a Full Profile approach is not the best for this application; that perhaps better results can be obtained through the use of Adaptive, or Choice-Based Conjoint.

We are eager to complete our analysis of the data we have from this study, and to make plans for future conjoint studies in this area.

[2] This includes data from card testing not reported on in this paper.

APPENDIX

Explanation of "Long Card" and "Short Card":

Short Card:

Usually only two panels. The information asked for is only name and address, model and serial number, date of purchase, and sometimes telephone number.

Long Card:

Usually three panels, sometimes four. In addition to asking for name and address etc., there are about a half-dozen questions relating to the purchase - where, how, why, etc. Also, these cards have another 10 to 12 questions asking for the respondent's household demographic and lifestyle information.

Explanation of "Consumer Protection Text":

This is language designed to act as an incentive to return the card. It lists several "important benefits" to returning the card, which vary by product.

Complete study design

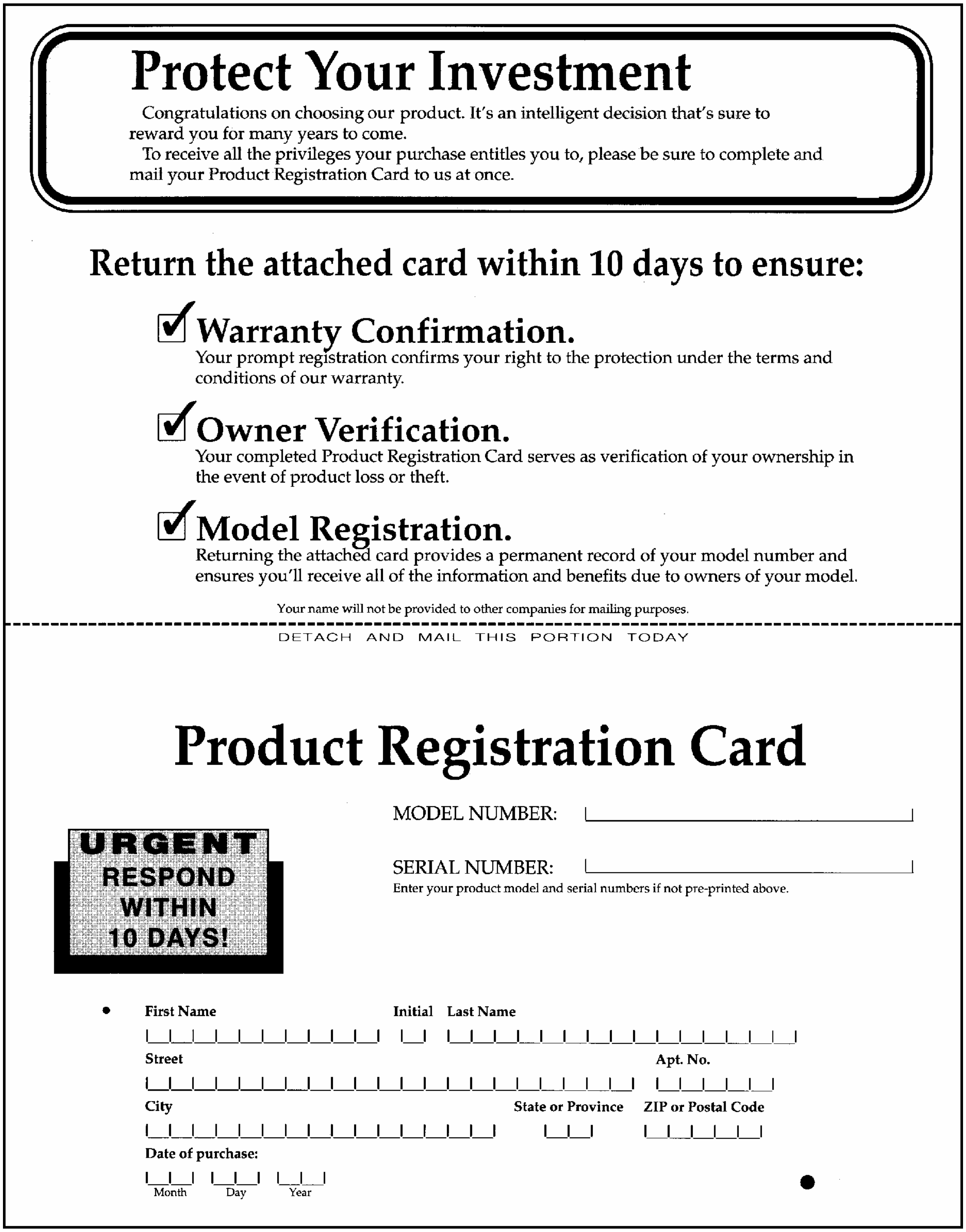
Product	Respondent Group I Card Set	Respondent Group II Card Set	Respondent Group III Card Set
$320 25" Color TV	Short Card, Consumer Protection Text	Long Card, Consumer Protection Text	Long Card, Extended Warranty A
$22 Toaster	Long Card, Extended Warranty B	Short Card, Consumer Protection Text	Long Card, $100K Sweeps
$35 Beard Trimmer	Long Card, Consumer Protection Text	Long Card, $100K Sweeps	Short Card, Extended Warranty A
$350 Gas Grill	Short Card, $100K Sweeps	Long Card, Extended Warranty B	Long Card, Consumer Protection Text
$900 Refrigerator	Long Card, Extended Warranty B	Short Card, Consumer Protection Text	Long Card, Consumer Protection Text
$50 Toaster Oven	Long Card, Consumer Protection Text	Short Card, $100K Sweeps	Short Card, Extended Warranty A
$180 RTA Entertainment Center	Long Card, $100K Sweeps	Short Card, Extended Warranty B	Short Card, Consumer Protection Text
$750 Washer & Dryer Set	Short Card, Consumer Protection Text	Long Card, $100K Sweeps	Long Card, Extended Warranty A
$190 VCR	Short Card, $100K Sweeps	Long Card, Consumer Protection Text	Short Card, Extended Warranty A
$400 Shelf Stereo	Short Card, Extended Warranty B	Short Card, Consumer Protection Text	Short Card, $100K Sweeps

The shaded cells indicate cards that have been tested in actual products.

Table D: Results for all cards included in the study

Product / Card Tested	Average "Likelihood of Return" Scores Measured by Survey	Purchase Likelihood Calculated by CVA
$320 25" Color TV - Short, Consumer Protection Text	90.08	89.34
$22 Toaster - Long, Extended Warranty B	59.94	66.52
$35 Beard Trimmer - Long, Consumer Protection Text	56.95	63.36
$350 Gas Grill - Short, $100K Sweeps	90.14	89.58
$900 Refrigerator - Long, Extended Warranty B	84.86	84.95
$50 Toaster Oven - Long, Consumer Protection Text	59.95	66.18
$180 RTA Entertainment Center - Long, $100K Sweeps	67.96	66.29
$750 Washer/Dryer - Short, Consumer Protection Text	95.18	91.12
$190 VCR - Short, $100K Sweeps	91.82	91.46
$400 Shelf Stereo - Short, Extended Warranty B	90.04	88.34
$320 25" Color TV - Long, Consumer Protection Text	83.42	82.13
$22 Toaster - Short, Consumer Protection Text	82.75	77.45
$35 Beard Trimmer - Long, $100K Sweeps	70.19	68.61
$350 Gas Grill - Long, Extended Warranty B	86.52	80.45
$900 Refrigerator - Short, Consumer Protection Text	93.5	91.37
$50 Toaster Oven - Short, $100K Sweeps	83.63	81.98
$180 RTA Entertainment Center - Short, Extended Warranty B	79.37	74.64
$750 Washer/Dryer - Long, $100K Sweeps	86.02	87.61
$190 VCR - Long, Consumer Protection Text	81.09	81.96
$400 Shelf Stereo - Short, Consumer Protection Text	90.46	87.79
$320 25" Color TV - Short, Extended Warranty A	85.99	90.92
$22 Toaster - Long, $100K Sweeps	73.4	70.48
$35 Beard Trimmer - Short, Extended Warranty A	81.34	80.06
$350 Gas Grill - Long, Consumer Protection Text	77.63	79.08
$900 Refrigerator - Long, Consumer Protection Text	86.80	84.40
$50 Toaster Oven - Short, Extended Warranty A	85.98	82.12
$180 RTA Entertainment Center - Short, Consumer Protection Text	71.23	74.03
$750 Washer/Dryer - Long, Extended Warranty A	88.31	86.95
$190 VCR - Short, Extended Warranty A	94.21	91.29
$400 Shelf Stereo - Short, $100K Sweeps	89.88	90.16

An example of a “Short” card with “Consumer Protection Text”



Protect Your Investment

Congratulations on choosing our product. It's an intelligent decision that's sure to reward you for many years to come.

To receive all the privileges your purchase entitles you to, please be sure to complete and mail your Product Registration Card to us at once.

Return the attached card within 10 days to ensure:

☑ **Warranty Confirmation.**
Your prompt registration confirms your right to the protection under the terms and conditions of our warranty.

☑ **Owner Verification.**
Your completed Product Registration Card serves as verification of your ownership in the event of product loss or theft.

☑ **Model Registration.**
Returning the attached card provides a permanent record of your model number and ensures you'll receive all of the information and benefits due to owners of your model.

Your name will not be provided to other companies for mailing purposes.

DETACH AND MAIL THIS PORTION TODAY

Product Registration Card

URGENT
RESPOND WITHIN 10 DAYS!

MODEL NUMBER: ____________________

SERIAL NUMBER: ____________________

Enter your product model and serial numbers if not pre-printed above.

● First Name | Initial | Last Name

Street | Apt. No.

City | State or Province | ZIP or Postal Code

Date of purchase:
Month Day Year

Partial Copy of the Questionnaire

Here are the instructions to the panel members and the first two tasks. The other eight tasks are similar to these. The real product brand names have been changed to maintain confidentiality of Polk's clients.

Dear Panel Member:

Thank you for taking the time to participate in our survey. It will take you only a few minutes to complete. We hope you will find it fun and interesting.

This survey is about the product registration cards, or "warranty cards", found packed with many household items you may have purchased or been given. It is about ONLY those types of materials. Please DO NOT think about other types of surveys or questionnaires you have been asked to complete in the past.

Enclosed with this survey are two sets of materials: 10 product registration cards and a set of 10 pictures which are stapled together. Please locate these materials and follow the instructions for each question below.

This survey asks you to use your imagination, and is something in which your entire household can take part. If you are not familiar with one of the consumer products in the questions below, or are not a likely user or buyer of the product, please ask another member of your household who may be more likely to use or be familiar with the product to answer that question.

Each situation described below should be considered separately, and has nothing to do with any of the other situations.

Also, please do not check the "I cannot answer. . . " box for a question unless you truly would never buy or own (now or in the future) a product even a little bit similar to the one described.

Please return only this questionnaire. You may discard the other materials when you are finished.

Please complete and return this questionnaire within 10 days

Thank you again for your Help!

1. **Find and look at the picture labeled "Picture A" and the registration card labeled "Card A"**

Please imagine that you recently needed to replace your old television, or just wanted a new one, and have just purchased this new Brand XYZ 25-inch color stereo television. You paid $320 (before sales tax). In the box, packed with the owner's manual, you found this product registration card **(Card A)**.

Please look at Card A and write a number below telling how likely you are to mail back the card, using a scale where:

0 = Definitely Would NOT mail back the card; 100 = Definitely Would mail back the card

Write your answer as a number from 0 to 100: □□□

□ I cannot answer this because no one in this household would ever buy a television.

2. **Find and look at the picture labeled "Picture B" and the registration card labeled "Card B"**

Now imagine that your old toaster died, so you went to a store and found this Brand ABC toaster. You paid $22 for it (before sales tax). When you opened the box at home, you found this product registration card **(Card B)** lying on top of the toaster.

Please look at Card B and write a number below telling how likely you are to mail back the card, using a scale where:

0 = Definitely Would NOT mail back the card; 100 = Definitely Would mail back the card

Write your answer as a number from 0 to 100: □□□

□ I cannot answer this because no one in this household would ever buy a toaster.

CONJOINT ANALYSIS ON THE INTERNET

Jill S. Johnson
The Bonham Group Market Research Company
Tom Leone
MediaOne
John Fiedler
POPULUS, Inc.

CLIENT PROJECT BACKGROUND REPORT

How can we optimize market penetration and maximize revenue in a rapidly changing environment with several new competitors poised to enter the industry? A major telecommunications company asked this question with regard to its high-speed Internet access product. As such, it was necessary to understand consumer demand and price elasticity under various pricing and product configurations.

There were several components that comprised the total cost of the high-speed Internet access service. They included installation costs, equipment costs and the monthly service cost. It was also important to understand the impact of brand on consumer demand.

The issue at hand for the telecommunications company was to develop an intermediate pricing strategy for its consumer segment. The pricing strategy was to be developed as part of a marketing strategy that focused on "conversion" of current on-line subscribers to a high-speed service as opposed to acquisition of "new" on-line subscribers. Therefore, current dial-up subscribers were the primary audience of interest for the study. While non-subscribers were also of interest, they presented a longer-term target segment. Given the rapidly changing landscape of the Internet industry, it was determined that research would be conducted with this consumer group at a later date.

Further, it was necessary for the research to assess perceptions of consumers who reside within the company's service area, where the high-speed Internet access service was available, while excluding current customers.

VENDOR FIELD REPORT

The measurement objectives clearly called for some sort of conjoint analysis. Because there were relatively few attributes and because pricing was a key objective, we focused on either full profile conjoint (Sawtooth Software's CVA™) or Choice-based Conjoint (Sawtooth Software's CBC™). The client's budget was modest; there were only four weeks until the analyzed findings were due. The stimuli were sufficiently complex that telephone interviewing would not be possible. Computer assisted self-administered interviews (CASI) were considered but there were not qualified field agencies located within all of the six geographic areas required by the client. Further the budget would not permit recruiting current Internet users for central location interviews.

The Internet was a logical means to interview people about an Internet service. So we chose to implement the study using Sawtooth Software's CVA Internet Module. It was a learning experience for us and we would like to describe twelve critical steps we identified during the project.

1. Use Sawtooth Software's *CVA System Internet Module*™. The Windows® system uses a template, fill-in-the-blank approach that makes questionnaire development very straightforward (easier to program). We used an Internet service provider (ISP) that gave us the permissions necessary to run the interview on its server. (Some ISPs do not permit users to run programs or to collect and store data on their servers.)

2. Study a universe appropriate for a web-based survey. Despite higher claimed estimates, real web access penetration is only 20%.

3. Obtain an appropriate sample. In our recent study, we focused on only 5% of ZIP codes; these reflected the geographic areas in which our client would soon launch its service. Survey Sampling (www.ssisamples.com) was able to provide names and phone numbers of Internet subscribers within our targeted areas.

4. Recruit and screen participants by telephone. Recruiting can be accomplished by a brief CATI survey, screening potential respondents for Internet access along with topical and security screens. A successful recruit resulted in verifying the respondent's name and obtaining e-mail address. Aiming for a final survey sample of 500, POPULUS obtained the names and e-mail addresses of 1,000 qualified respondents, assuming 50% cooperation for a completed interview.

5. Train interviewers to properly record an e-mail address. Even the best of interviewers are accustomed to record open-end answers for meaning rather than precise wording. Interviewers must be instructed to read back an e-mail address, character by character.

6. Send personalized e-mails within 24 hours to each recruited respondent. Otherwise people can quickly forget what they've promised to do.

7. Create a unique password allowing each respondent access to the web site. The Sawtooth CVA Internet module allows for the creation of passwords. Respondents use the password to begin the survey. If necessary, a respondent can leave the survey and use the password to resume the survey at a later time. However, once a respondent has completed the survey, the password is rendered inoperable, preventing repeated access to the survey site.

8. Use an e-mail package such as MailKing® (http://www.mailking.com), a software program used to send personalized e-mail messages to each respondent. Each e-mail message should contain a hyper link to the survey web site and the respondent's unique password. Each morning, simply load the results of the previous evening's recruiting into a spreadsheet, add a password to each record, and MailKing does the rest.

9. Offer a generous incentive. In our case, respondents were informed that of those who completed the Web survey, one person from each of the six service area cells would be randomly selected to receive a check for $100. Notify winners via e-mail.
10. Assume a 50% cooperation rate. POPULUS sent out 1,057 e-mail messages along with follow-up messages. Of these, 162 (15%) were returned as undeliverable. Completed surveys were obtained from 482 respondents within two weeks of the first e-mail mailing.
11. Monitor the site regularly: daily, even hourly. It's easy to keep clients up-to-date regarding the number of completed web site interviews.
12. Download interim data frequently. Use the Sawtooth DOS CVA program for the conjoint analysis and any statistical package for the rest of the data. Schedule the top-line meeting with your client the day after the survey site is closed.

The project was completed on time and on budget. The client's only concern was a reluctance to utilize the CVA simulator. After some initial resistance was dispelled, the program and data files were zipped up, appended to an e-mail, and installed five minutes later.

CLIENT ANALYTICAL REPORT

Personnel changes at the communications company resulted in having to bring a new manager up to speed on the project. The change came after the data collection and initial topline findings for the study were presented to the internal group who sponsored the study.

At first, the new manager relied heavily on POPULUS to field additional requests for simulations. The Sawtooth CVA simulator is very different from other simulators that had been used at the company to analyze conjoint results. The other simulators the manager was familiar with were mainly Excel spreadsheets where each scenario is entered into a very friendly front-end one at a time and the result calculated. Once trained though, he found that the main benefit of the Sawtooth CVA simulator was that it allowed the running of multiple (up to 30) scenarios at once. Given the volume of simulation requests he received for the project that function become very valuable. The only problem he encountered with the simulator was the inability to print the results of a simulation batch directly after the simulation was run (i.e. the results file had to be first saved to a directory, and then opened in Wordpad, and then printed). Additionally, the results had to be re-entered into a spreadsheet for purchase intent discounting—a step not needed with other simulators he used.

Aside from the minor mechanics of running the simulations, the data was well received and used by various organizations in the company. The data became part of a larger model (using Crystal Ball software) being constructed to optimize the pricing configuration for the high speed data product. In conjunction with internal data on costs to serve, assumptions about the population, retail alliances and other factors a model was built that allowed the sponsor group to configure its pricing to maximize penetration and revenue while maintaining a control on costs.

While some at the company were skeptical of online methodology (due to concerns about representativeness), the sponsoring group for this project expected an online component for the research due to the target audience for the product. The initial telephone recruit combined with 48% response rate served to allay any concerns others may have had. In fact, the telephone recruit was a new twist in the way online research had previously been done by the communications company. For the most part, prior online studies used e-mail to customers or a syndicated online panel to generate traffic to a survey web site. The telephone recruit to an online site will probably become a model for future studies—for this product line—that target online consumers beyond the company's current customers.

TWO ORDINAL REGRESSION MODELS

Tony Babinec
SPSS, Inc.

1. INTRODUCTION

Ordinal response variables are widespread in market research and social and behavioral research. Examples abound:

- Purchase likelihood
- Satisfaction
- Skill level
- Income groups

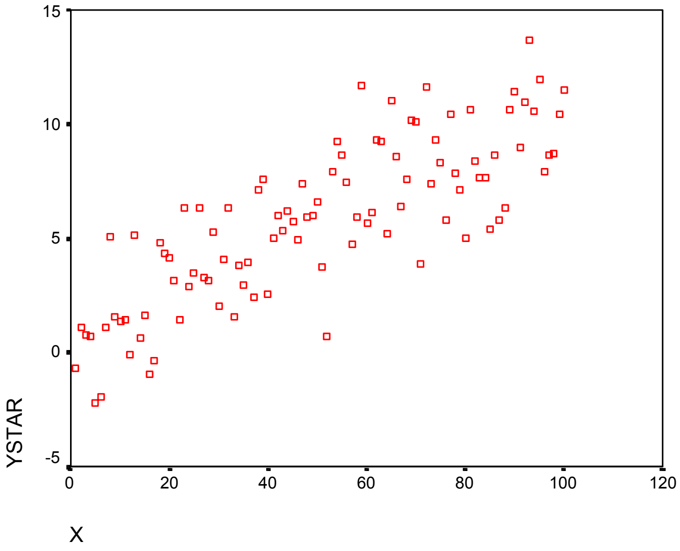
It is probably a safe guess that many if not most researchers scale these variables using sequential integer scores (or perhaps midpoint scores in the case of grouped income) and then analyze them using conventional linear regression. Doing so involves the implicit assumption that the intervals between adjacent categories are equal. The assumption is made that OLS regression "gets close enough to the truth." Sometimes the following "handwaving" argument is made: Linear regression is accessible and easy to use while the added complexity of ordinal logistic or probit models gains you nothing. The point of view of this paper is that there can be serious statistical problems with the naïve use of OLS regression, and that prudence suggests using an approach appropriate to the measurement level of the data. There exist a number of widely available models, including one that deserves more attention, namely, the adjacent-category logit model.

This paper proceeds as follows. Section 2 presents a simple motivating example. Section 3 presents a summary of theoretical results regarding the use of OLS regression for dichotomous and ordinal responses. Section 4 presents Goodman's association model and a logit form. Section 5 contrasts the more widely known cumulative logit model with the adjacent-category logit model. Section 6 applies them both to an empirical realization of the discretized bivariate normal. Section 7 presents an extended example wherein several models are applied to a "real world" regression modeling situation. Section 8 presents conclusions.

In the expositions that follow, discussion of models presents the single predictor case for notational and expository convenience. Section 7 presents an example with multiple predictors. The reader interested in learning more about the model is referred to technical appendices in [Magidson, 1998].

2. A Motivating Example

Here is a simple example inspired by one discussed in [Long, 1997]. Consider the following bivariate scatterplot.



The variable X consists of the integers 1 – 100. Variable YSTAR is generated from the equation:

ystar=1 + .1*x + e

where the intercept is 1, the slope coefficient is 0.1, and the error term is realized from a Normal distribution with a mean of 0 and a standard deviation of 2. Regression of YSTAR on X produces the following estimates:

$R^2 = 0.66$

Standard error of the estimate = 2.0761

Intercept = 0.776(se = 0.418)

Slope = 0.0992(se = 0.007).

It is sometimes argued that an ordinal variable is an observed form of a latent continuous variable. The observed variable contains incomplete information in the following sense: Instead of observing the continuous latent variable, you observe discrete groupings governed by unobserved thresholds. For example, suppose that instead of observing YSTAR you observe Y, an ordinal version of YSTAR created in the following way:

recode ystar

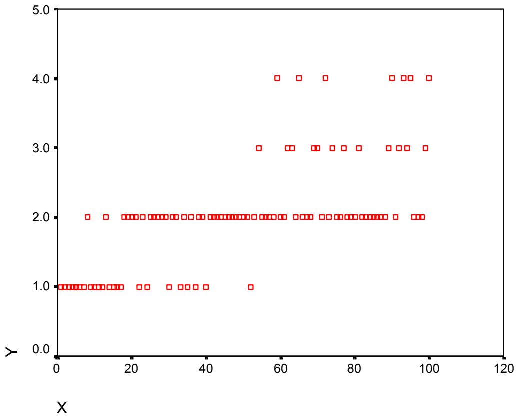
(lo thru 3=1)

(3 thru 9=2)

(9 thru 11=3)

(11 thru hi=4) into y.

This coding scheme assigns sequential integers to the categories, a widespread practice in empirical research. The bivariate scatterplot of Y versus X is as follows:



Note that the vertical axis in the second scatterplot is not to the same scale as in the plot immediately above.

The regression of Y on X produces the following results:

R^2 = 0.438

Standard error of the estimate = 0.6006

Intercept = 1.112(se = 0.121)

Slope = 0.0182(se = 0.002)

Contrasting these results with those obtained with the continuous response variable, you can see that for the grouped Y variable R^2 is attenuated, the standard error of the estimate is wrong, and the estimated regression coefficient is biased heavily downward.

3. WHAT'S WRONG WITH USING OLS REGRESSION ON ORDINAL RESPONSES?

This section reviews the standard regression model and states the theoretical and practical objections to the naïve use of OLS regression when the response variable is dichotomous or ordered categorical.

The simple linear regression model ([Kennedy, 1992]) has the following form:

$Y = \alpha + \beta{*}X + \varepsilon$

where Y and X are variables and α and β are parameters. The assumptions of the standard model are:

$E(\varepsilon) = 0$

$V(\varepsilon) = \sigma^2$

$Cov(\varepsilon_i,\varepsilon_j) = 0$

That is:

- Errors have a mean of 0.
- Errors have a finite, constant variance.
- Errors are uncorrelated across cases.

The above assumptions constitute the typical assumptions made under OLS regression. α and β are conveniently estimated by least squares. If the stronger assumption of normality is made:

$\varepsilon \sim Normal(0,\sigma^2)$

then the OLS estimates are maximum likelihood estimates. In this case, the estimates from regression are best (minimum variance) unbiased and the researcher can use the standard hypothesis testing and confidence intervals framework.

Suppose that you use OLS regression to predict a dichotomous response variable. It is natural to code the response variable 0,1, and to interpret the predicted values from the model as the probability that the response is 1 for that case. Such a model has been termed the *linear probability* model (see, for example, [Aldrich and Nelson, 1984]). The problems with doing so include the following:

1. The residuals do NOT have a constant variance. As a consequence, the estimates from regression are not "best," that is, minimum-variance.
2. The standard errors of the regression coefficients are wrong, so that related hypothesis testing and confidence interval construction are invalid.
3. The predicted values from OLS regression can range outside the interval [0,1], whereas probabilities are bounded by that interval.

4. The linearity assumption inherently imposes constraints on the marginal effects of predictor variables that are not taken into account by OLS estimation.

5. The linearity assumption implies that the marginal effect of a predictor is constant across its range, which can be an unrealistic assumption.

6. The usual R-squared measure is problematic. Cox and Wermuth [Cox and Wermuth, 1992] have shown that for a dichotomous Y and a situation where the observed proportions of success are bounded by [.2,.8] and therefore the linear regression model ought to be a good fit, the maximum possible value of R^2 is 0.36.

Suppose instead that the response variable is polytomous with integer scores. The problems with proceeding as if this variable is continuous include the following points, which parallel many of those stated above for the dichotomous case ([Clogg and Shihadeh, 1994]):

1. The scoring system is arbitrary. In addition, if the first or last category is open-ended, special considerations must be taken.

2. The estimated regression coefficient refers to the effect of a unit change in X on the rank order of Y levels. With respect to the estimated effect of unit change in X on change in Y in its original units, the OLS estimator is biased and has the wrong standard error.

3. Comparability across samples can be affected by the scoring system used.

4. Inferences, both substantive and statistical, depend on the scoring system used.

5. The analysis cannot be fully efficient (minimum-variance).

6. It is possible to get predicted values outside the range of the observed response variable.

7. It is more natural to assume a multinomial distribution for Y, not a normal distribution.

8. Many of the standard summaries of regression analysis – R^2, mean squared error, sums of squares, standardized coefficients, and so on – no longer have the same meaning they have in ordinary regression with a continuous Y.

4. THE ASSOCIATION MODEL AND THE LOGIT FORM

Goodman's association model [Goodman, 1979] was originally proposed to analyze two-way tables of variables with ordered categories. Since its introduction, the association model has been extended and recast in a number of ways. This section reviews the association model and shows a particular form of a logit model that can be viewed as a type of ordinal regression.

For an I x J contingency table with ordered categories, let P_{ij} denote the probability under some model of being in row i and column j. One form of the association model is expressed in terms of row and column scores:

$$P_{ij} = \eta\alpha_i \beta_j \exp(\phi\mu_i\nu_j)$$

where η, $\alpha_{i,}$ $\beta_{j,}$ and ϕ are parameters and μ_i and ν_j can be parameters or fixed scores. Without loss of generality, let η be 1. When both the μ_i and the ν_j are fixed scores, this model gives rise to the uniform(U) association model. When the μ_i are estimated but the ν_j are fixed, this model gives rise to the row effects(R) model. When the μ_i are fixed but the ν_j are estimated, this model gives rise to the column effects(C) model. Finally when both the μ_i and the ν_j are estimated, this model gives rise to Goodman's RC association model. The U, R, and C models can be estimated using conventional loglinear modeling software, but the RC model is not a loglinear model and therefore cannot be estimated using conventional loglinear estimation software.

For identification purposes, restrictions are specified on the μ_i and ν_j. One example of such restrictions is the following:

$\Sigma \mu_i = 0$

$\Sigma \mu_i^2 = 1$

$\Sigma \nu_j = 0$

$\Sigma \nu_j^2 = 1$

Association models can also be understood in terms of local odds ratios in the I x J table:

$\theta_{ij} = F_{ij} F_{i+1,j+1} /(F_{i,j+1} F_{i+1,j})$, $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$.

where the F_{ij} are expected counts under the model. There are (I-1)(J-1) local odds ratios, each referring to the association in a particular 2x2 subtable formed from adjacent rows and columns. The usual independence model says that

$\theta_{ij} = 1$ for all i and j.

The uniform association model is described by the simple condition:

$\theta_{ij} = \theta$, $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$.

The row effects model is described by the condition:

$\theta\ ij = \theta\ i.$, $i = 1, \ldots, I-1$.

The column effects model is described by the condition:

$\theta\ ij = \theta\ .j$, $j = 1, \ldots, J-1$.

Further insight into the association model is given by the formula for the log-odds ratios:

$\log \theta\ ij = \phi\ (\mu_i - \mu_{i+1})\ (\nu_j - \nu_{j+1})$

The uniform association model is obtained when:

$\mu_i - \mu_{i+1} = \Delta$ $(i = 1, \ldots, I-1)$, $\nu_j - \nu_{j+1} = \Delta'$ $(j = 1, \ldots, J-1)$,

where Δ and Δ' are unspecified. The row-effect association model and column-effect association model can be defined in similar terms.

In conventional loglinear modeling, there is a natural connection of logit models with loglinear models. The loglinear model treats all variables as jointly dependent, that is, makes no distinction between response and predictor variables. The logit model makes a distinction between a response variable and a predictor variable. The loglinear model models expected counts. The logit model models ratios of counts. For this reason, all conventional logit models are loglinear models and can be estimated using loglinear modeling approaches.

In similar fashion, consider an ordinal-level response variable predicted by a single x. In this case, consider logits formed from adjacent categories of the dependent variable. For the Uniform association model above, this gives:

$\log(F_{i,j+1} / F_{i,j}) = a_j + \varphi (y_{j+1} - y_j) x_i$, j = 1, …, J-1.

In this model, there are J-1 intercepts and one slope, j, constant across all logits. This model is termed the *parallel logit* model for adjacent categories. This model can be generalized to accommodate the following situations:

- Response variable has two categories or more than two categories.
- Distances between adjacent response categories are assumed equal or not equal.
- There are one or more predictor variables.
- Predictors can be dichotomous, nominal, ordinal, or continuous.

The model is estimated using maximum likelihood and tested using likelihood ratio chi-square tests. This model is contrasted with McCullagh's proportional odds model in the next section.

5. Two Ordinal Regression Models

This section describes two models suited to ordinal response variables – the cumulative logit model and the adjacent-category logit model.

Following Clogg and Shihadeh, let Y take on the values {0, 1, …, K}, for a total of K+1 response categories. Assume that the categories of Y are strictly ordered. Then, Y=k denotes "less" of some attribute than Y=k+1, for k=0,…, K-1. Let P_k denote Pr(Y=k), and assume that each member of the sample has the same probability P_k that Y=k. Assume that the sample is obtained by independently sampling units from some large population where P_k is the proportion with Y=k. Then, the sampled frequencies follow a multinomial distribution.

Cumulative logit model. For Y strictly ordinal, define

$\Pi_k = P_0 + \ldots + P_{k-1} = \Pr(Y<k)$,

With

$1 - \Pi_k = \Pr(Y>=k)$,

for k=1, …, K. The Π_k define the cumulative distribution function of Y. The logits of interest are

$\log[(1-\Pi_k)/\Pi_k]$

which is a contrast of "at or above k" versus "below k" for each k. You can write a logit model as:

$\log[(1-\Pi_k)/\Pi_k] = \alpha_k + \beta_k x$

Clogg and Shihadeh observe that, as written, there is no evident parsimony, nor is it clear whether this form of the model properly deals with ordering. If you impose the restriction

$\beta_k=\beta$

the resulting model is one kind of "ordinal regression" model. McCullagh termed this model a *proportional odds* model, but there exists other models which imply a kind of proportional odds, so this model is perhaps better referred to as the *cumulative logit* model.

The above model with the restriction $\beta_k=\beta$ imposed produces parallel logit regression lines over the range of the x axis. The model imposes a common slope for each logit but allows intercepts to differ. This model has the property that successive odds for the cumulative distribution are proportional, hence the name proportional odds model.

In applying this model, you should assess the assumption that the slopes are identical. One way to do this is via the following heuristic approach. From the categories of the response variable, form a set of dichotomous variables in the following way:

If $Y=0$, $Y_1'=0$.

If $Y>0$, $Y_1'=1$.

If $Y<=1$, $Y_2'=0$.

If $Y>1$, $Y_2'=1$.

…

If $Y<=K-1$, $Y_k'=0$.

If $Y=K$, $Y_k'=1$.

Then, using any conventional binary logistic regression program, estimate logit models for each $Y_{\underline{k}}'$. The parameters in this model are the same as those in the cumulative logit model with no restriction on the β_ks. A qualitative diagnosis of the restricted model of equal β_ks consists of visual examination of the estimated slope values for each of the variables. If the assumption of parallel regressions is true, then each estimated βk is a consistent estimate of β. [Long, 1997] presents a discussion of two formal tests that can be applied: a score test and a Wald test.

Adjacent-category logit model. The response function is the set of logits $\log(P_k/P_{k-1})$ for k=1,…, K. One model that could be used is

$$\log(P_k/P_{k-1}) = \alpha_k + \beta_k x$$

As in the case of the cumulative logit model, it is useful here to impose the restriction

$$\beta_k=\beta$$

which says that the adjacent category logits are parallel, or that the adjacent category odds are proportional. Thus, this model is a type of proportional odds model. If the x scores are treated as fixed, this model can be fitted using conventional loglinear software. However, if both y scores and x scores are to be estimated, then this model is not a loglinear model, and appropriate software must be used.

Some properties of the models. Here are some properties and features of the cumulative logit model.

- Sometimes called the "proportional odds" model because the successive odds for the cumulative distribution are proportional.
- The cumulative logit model can be derived from the notion of an underlying unobserved random variable Z such that $Z - \mathbf{B}^T\mathbf{x}$ has the standard logistic distribution.
- If the model holds for a given response scale, it also holds (in the population) with the same effects for any collapsing of the response categories.
- The expected counts, when summed, do NOT reproduce the marginal distribution of the response variable.
- The likelihood function CANNOT be expressed in terms of sufficient statistics for the regression coefficients. Therefore, you cannot eliminate betas by conditioning on their sufficient statistics, a key requirement for EXACT inference in discrete regression models.

Use of a complementary log-log link gives rise to a proportional hazards model:

- The complementary log-log model can be derived from the notion of an underlying unobserved random variable Z such that $Z - \mathbf{B}^T\mathbf{x}$ has the extreme value distribution.
- This model has a more direct relationship to failure-time models or hazard models.

Here are some properties and features of the adjacent category logit model.

- Can refer to this model as a "proportional odds" model because when betas are equal, adjacent category odds are proportional.
- This model has a close connection to the loglinear model and the association model.
- Therefore, the methods for evaluation of its goodness of fit are even more straightforward than those for the cumulative logit model.

- Also, the equal adjacent odds model has greater flexibility for extension which account for potential lack of fit.
- This model has a close connection to Poisson regression.
- This model has a connection to Anderson's Stereotype regression model. If the Y variable is the row variable, treat its levels as unordered and estimate an unordered polytomous logistic regression model and look for order in the regression coefficients.
- The discretized bivariate normal is well fit by the RC association model, making this model the natural analog to ordinary regression.
- The expected counts, when summed, reproduce the marginal distribution of the response variable.
- The likelihood function can be expressed in terms of sufficient statistics for the regression coefficients. Therefore, you can eliminate betas by conditioning on their sufficient statistics, a key requirement for EXACT inference in discrete regression models.
- Problem of comparability: When you collapse categories, or when you use two "versions" of the same variable that differ in the number of categories, this affects model results.

On balance, the adjacent-category logit model has more to recommend it.

6. Example: Discretized Bivariate Normal

A remarkable property of the association model is that it fits the discretized bivariate normal distribution very well. That is, even though the association model makes the relatively weak assumption that the data arise from the multinomial distribution, the association model is justified in the case where the data arise from bivariate normality. For related discussions, see [Goodman, 1981], [Goodman, 1991], [Becker, 1989], and [Magidson, 1996].

To see this, note that the association model is the discrete analog of the bivariate normal distribution. Here is a restatement of the association model form:

$$P_{ij} = \eta \, \alpha_i \, \beta_j \exp(\phi \, \mu_i \nu_j)$$

and here is the bivariate normal density assuming zero means and unit standard deviations without loss of generality:

$$f(x,y) = \{2\pi(1-\rho^2)^{1/2}\}^{-1} \exp\{-x^2/[2(1-\rho^2)]\} \exp\{-y^2/[2(1-\rho^2)]\} \exp\{\rho/(1-\rho^2)xy\}$$

A simple visual comparison shows that the model terms "line up."

As an example, here is a contingency table formed by collapsing values of two standard normal realizations, X and Y, that are correlated at 0.5, with N=10000. The cutpoints for the two variables are equally-spaced cutpoints at 0, +/- 0.5, +/- 1, +/- 1.5, +/- 2, +/- 2.5.

Y12 * X12 Crosstabulation

Count

		X12												
		1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10	11	12	Total
Y12	12.00					1	2	12	7	10	12	13	4	61
	11.00					3	8	42	43	25	31	15	8	175
	10.00				5	25	40	78	90	100	56	27	15	436
	9.00		1	6	24	59	119	189	238	143	99	43	13	934
	8.00		3	18	62	166	261	306	308	212	116	33	7	1492
	7.00	3	18	49	123	263	401	450	307	181	83	17	7	1902
	6.00	7	22	77	173	346	397	404	250	138	37	6	2	1859
	5.00	7	41	99	211	302	338	277	163	70	17	3	2	1530
	4.00	16	46	98	158	213	181	141	79	19	4			955
	3.00	16	23	59	109	99	74	44	8	10				442
	2.00	5	9	29	31	26	31	12	5	2	1			151
	1.00	4	9	17	13	9	5	4	2					63
Total		58	172	452	909	1512	1857	1959	1500	910	456	157	58	10000

Two models are fit to the data: 1) the adjacent category logit model and 2) the cumulative logit model. Here is a summary of the goodness of fit of the two models:

Independence model: L^2 = 2848.44 on 121df. X^2 = 3109.29

U association model: lack of fit L^2 = 118.07 on 120 df, p=0.282.

Cumulative logit model: lack of fit L^2 = 195.3622 on 120 df, p=.000017.

Note that the adjacent-category logit model fits the data while the cumulative logit model does not. Goodman has shown that the RC model or specially constrained versions of it fit the discretized bivariate normal even better. On the other hand, there is no simple remedy for the cumulative logit model to improve its fit to these data.

7. EXAMPLE: LONG'S ANALYSIS OF GSS DATA

Both [Clogg and Shihadeh, 1994] and [Long, 1997] present analyses of a particular response variable found in the General Social Survey in two years, 1977 and 1989. This section follows Long's analysis and presents new results.

The response variable in the analysis is responses to the statement:

WARM:

"How do you feel about the following statement -- A working mother can establish just as warm and secure a relationship with her children as a mother who does not work?"

Responses are coded in the variable WARM as:

1 Strongly disagree (SD)

2 Disagree (D)

3 Agree (A)

4 Strongly agree (SA)

The combined sample size from the two years is 2,293.

The frequency distribution of WARM is shown here.

Mother has warm relationship: 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	sd	297	13.0	13.0	13.0
	d	723	31.5	31.5	44.5
	a	856	37.3	37.3	81.8
	sa	417	18.2	18.2	100.0
	Total	2293	100.0	100.0	

Note that there are more respondents in the two middle categories than in the end categories. The modal response is "agree."

The predictors and codings used by Long are shown in the following table, along with simple summary statistics.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Mother has warm relationship: 1	2293	1.00	4.00	2.6075	.9282
Sex: 1=male, 0=female.	2293	.00	1.00	.4649	.4989
Race: 1=white, 0=not white.	2293	.00	1.00	.8766	.3290
Year of survey: 1=1989, 0=1977.	2293	.00	1.00	.3986	.4897
Age in years.	2293	18.00	89.00	44.9355	16.7790
Years of education.	2293	.00	20.00	12.2181	3.1608
Occupational prestige.	2293	12.00	82.00	39.5853	14.4923
Valid N (listwise)	2293				

There are three dichotomous predictors:

- Sex of respondent (MALE) with male=1 and female=0.
- Race of respondent (WHITE) with white=1 and nonwhite=0.
- Year of survey (YR89) with 1989=1 and 1977=0.

There are three quantitative predictors:

- Age of respondent in years (AGE)
- Years of education (ED)
- Occupational prestige (PRST)

Marginal two-way analysis of the response variable with each predictor taken separately shows:

- Males tend to disagree while females tend to agree with the response statement (negative effect).
- Whites tend to disagree while nonwhites tend to agree with the response statement (negative effect).
- Respondents in the more recent survey tend to agree while respondents in the older survey tend to disagree (positive effect).
- Older respondents tend to disagree (negative effect).
- Respondents with more years of education tend to agree (positive effect).
- Respondents with higher occupational prestige scores tend to agree (positive effect).

Here are results from running OLS regression, the cumulative logit model, and the adjacent category logit model on these data.

Predictor	OLS	Cumulative logit	Adjacent-category logit
Male	-0.336(-9.17)	-0.733(-9.34)	-0.44(-8.88) / 0.64
White	-0.177(-3.17)	-0.391(-3.30)	-0.24(-3.17) / 0.79
Yr89	0.262(6.94)	0.524(6.52)	0.35(6.82) / 1.41
Age	-0.010(-8.7)	-0.022(-8.71)	-0.01(-8.42) / 0.99
Ed	0.031(4.14)	0.067(4.21)	0.04(4.07) / 1.04
Prestige	0.003(1.73)	0.006(1.84)	0.00(1.76) / 1.00
Intercept	2.78(25.26)	0(.)	0(.)
T1		-2.465(-10.29)	1.10(7.39)
T2		-0.631(-2.7)	1.32(4.86)
T3		1.262(5.37)	0.49(1.22)
-2 ln L(total error)		5763.77	5763.77,6564 df
L^2(whole-model test),df		301.72,6	296.28,6 df
L^2(lack of fit)		5462.051,6558	5463.77,6558 df

Entries for the predictor variables are: *parameter estimate(t-statistic) / EXP(parameter estimate)*. For the cumulative logit model and the adjacent-category logit model, the t-statistic is the square root of the Wald chi-square statistic.

In general, the models broadly agree regarding the sign of each partial effect, and also agree regarding the nonsignificance of Prestige given the other predictors in the model. Regarding interpretation of effects and model fit, the models differ.

The OLS column lists the results from treating WARM as a numeric variable and estimating a linear regression model. In the OLS regression results, the regression coefficients represent the change in the response variable for a unit change in the given predictor variable. By the nature of the model, the partial effect is constant across values of the other variables. There is no report of model fit. As stated above, the usual R^2 and related statistics are suspect.

The cumulative logit column lists the results from estimating a cumulative logit model. The coefficients listed above agree with those reported by [Long, 1997] but the model fit statistics reported above are slightly different from those reported by [Long, 1997]. On the face of it,

- The L^2 for the whole-model test (301.72 on 6 degrees of freedom) indicates that as a set the regressors are statistically significant.
- Individually, the t-statistics indicate that all predictors but Prestige are statistically significant.
- The L^2 for lack of fit (5462.051 on 6558 degrees of freedom) indicates no statistically significant lack of fit.

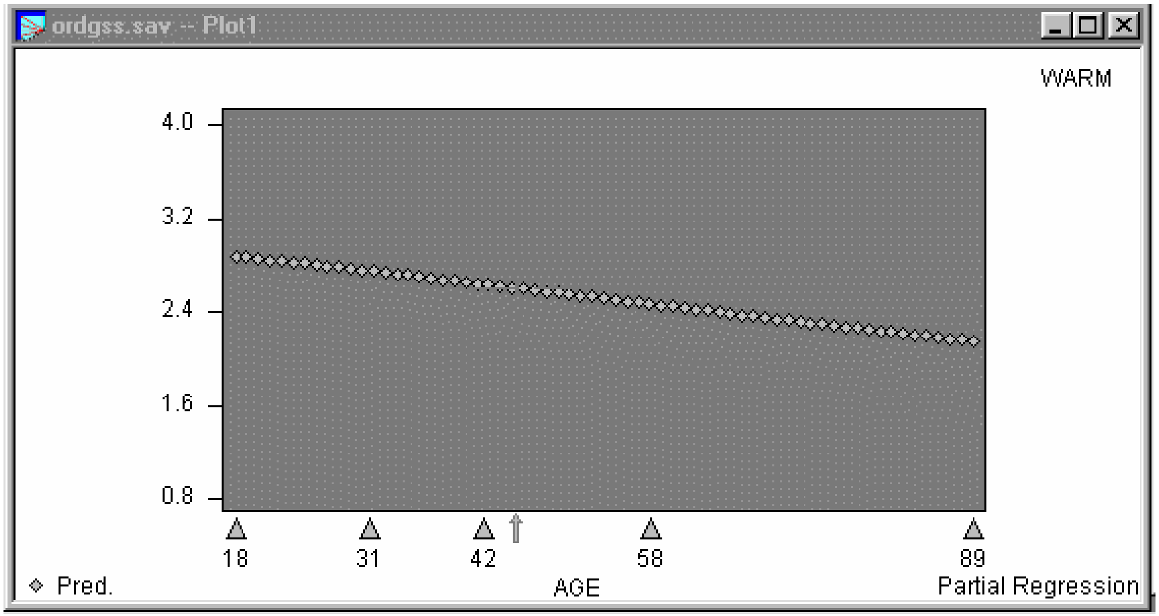
You might be concerned about the relative sparseness of the data and consequently about the interpretation of the fit chi-square. However, a more fundamental objection to the cumulative logit model here is that a test of the proportional odds assumption reveals that it does not hold. Long reports the following test statistics:

- Score test: L^2 equals 48.4 on 12 degrees of freedom, p-value <= 0.001.
- Wald test: L^2 equals 49.18 on 12 degrees of freedom, p-value <= 0.001.

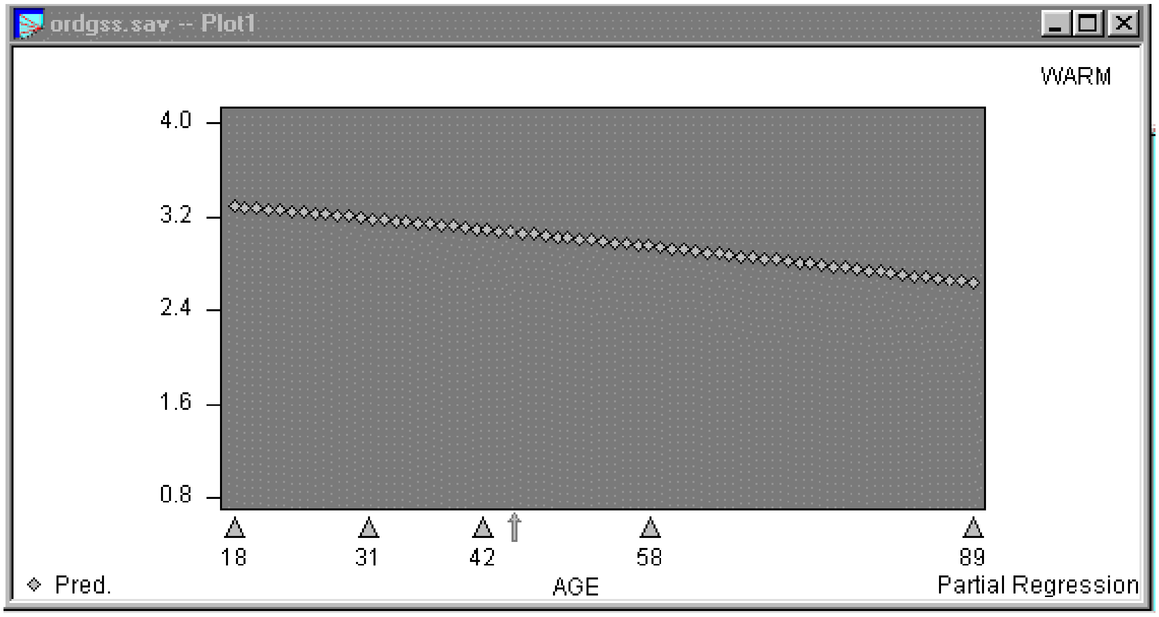
Therefore, the cumulative logit model does not fit the data. It is difficult to know how to proceed. Long concludes: "My experience suggests that the parallel regression assumption is frequently violated based on either an informal test, the score test, or the Wald test. When the assumption of parallel regressions is rejected, alternative models should be considered." Long does not present a further model, but you might consider using a nominal logistic regression modeling routine.

The adjacent-category logit column reports results of fitting the adjacent-category logit model. In this analysis, the response variable, WARM, is treated as consisting of equally-spaced categories. All predictors are treated as fixed with known scores. The estimated regression coefficients, or the EXP transformations of them, can be thought of in much the same way that coefficients are understood in binary logistic regression. An important difference is that in this instance the response variable has more than two categories, so there must be a choice of baseline comparison on the response variable. The model lends itself to rich interpretation; however space allows only a cursory exploration.

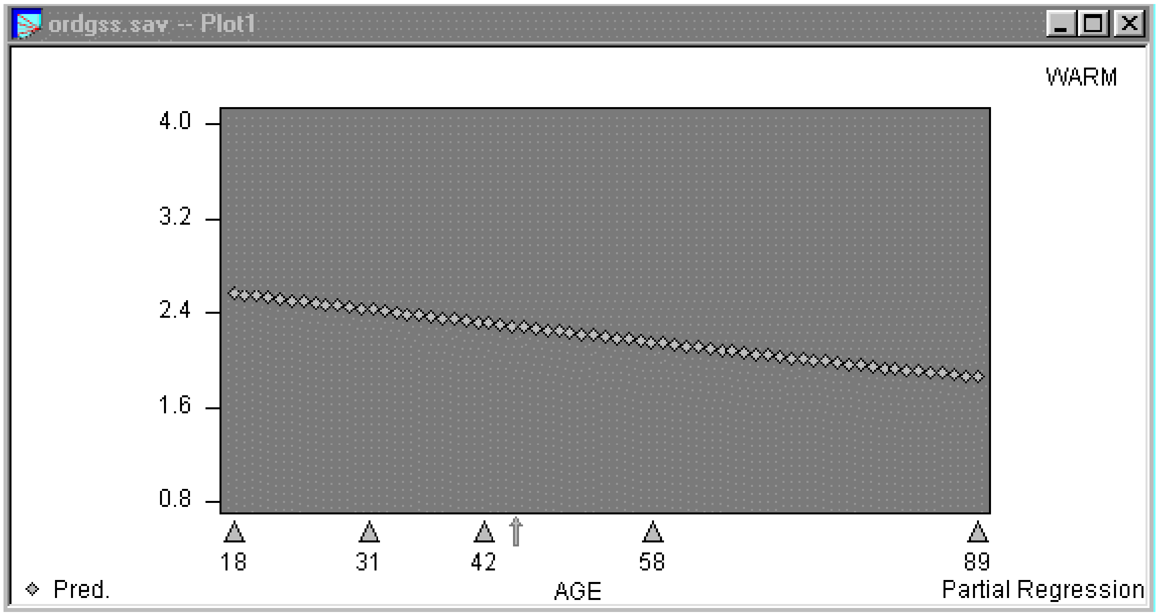
Here is a Partial regression plot of Age. All variables in the analysis are set to their weighted mean. The vertical axis portrays WARM in its metric of 1-4. This plot shows the response along the four-point scale of WARM as AGE varies from its minimum to maximum conditional on the values of the other variables.



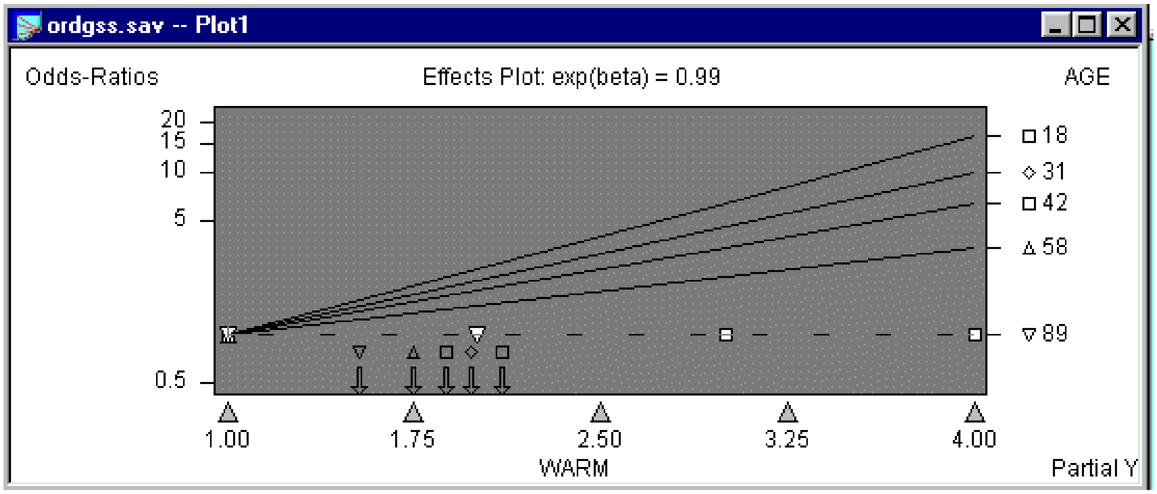
Here is another Partial regression plot of Age. In this plot, quantitative variables are set to their weighted means, while MALE=0, YR89=1, and WHITE=0. Note how the partial regression curve shifts up relative to the first curve shown above.



Here is a third Partial regression plot of age. In this plot, quantitative variables are set to their weighted means, while MALE=1, YR89=0, and WHITE=1. Note how the partial regression curve shifts down relative to the other two curves.



Finally, consider the following partial-Y plot of Age.



The settings of the variables are to the baseline categories that most disagree with the response statement.

Male – Reference category is Male(1).

Yr89 – Reference category is 1977(0).

White – Reference category is white (1).

Age – Reference category is 89.

Ed – Reference category is 0.

Prestige – Reference category is 12.

Warm – Reference category is Strongly Disagree.

The vertical axis is on an odds-ratio scale in a log-spacing. The horizontal axis portrays the categories of WARM. The body of the plot is a set of regression lines. Consider the line of greatest slope, corresponding to Age=18. This line says that 18 year olds relative to 89 year olds are 16 times as likely to Strongly Agree as Strongly Disagree, conditional on the values of the other variables. This horizontal axis also plots the means of representative Age groups (see the symbols and vertical arrows), thereby producing a visual rendering of the effect size of Age conditional on the settings of the other predictors.

The fit statistics for the adjacent-category logit model are as follows:

- The L^2 for the whole-model test (296.28 on 6 degrees of freedom) indicates that as a set the regressors are statistically significant.
- Individually, the t-statistics indicate that all predictors but Prestige are statistically significant.
- The L^2 for lack of fit (5467.49 on 6558 degrees of freedom) indicates no statistically significant lack of fit.

In using the adjacent-category logit model on continuous predictors, you might be concerned about data sparseness and its effect on the use of chi-square to assess goodness of fit. One suggestion is the following: An informal test of sparseness is to compare the likelihood ratio chi-square statistic with the pearson chi-square statistic. If they are "different" in value, you should be concerned about sparseness in your data. In the adjacent-category model fit above, the residual L^2 is 5467.49 on 6558 degrees of freedom while the residual X^2 is 6610.05 on 6558 degrees of freedom. Given this result, you might use the Hosmer and Lemeshow decile fit statistic (references include [Hosmer and Lemeshow, 1989], Magidson [1998]). For the model above, the decile fit statistic is 17.75 on 8 degrees of freedom, p=0.023, which suggests some lack of fit.

To improve fit in this example, you might consider trying the following, either separately or in concert:

- Relax the assumption that the categories of WARM are equally spaced.
- Explore and model interactions between the predictors.

Relaxing the assumption of equally-spaced categories on WARM does improve the fit. The Decile Fit statistic for this model is now 13.86 on 8 degrees of freedom, p=0.085. The estimated response category scores in this model are:

1	Strongly disagree
1.75	Disagree
3.04	Agree
4	Strongly agree

That is, there is some suggestion that the Disagree category is closer to the Strongly Disagree category than to the Agree category.

If instead you decide to explore interactions for these data, you would find that the interaction between Age and Education improves the fit in a model with all main effects in except Prestige, which was omitted because it was nonsignificant above:

- L2 for the whole model is 296.67 on 6 degrees of freedom.
- L2 residual is 3871.64 on 4338 degrees of freedom; X2 for the same model is 4324.92.
- The decile fit statistic is: L2 = 8.17 on 8 degrees of freedom, X2 = 8.17 also.
- The LR chi-square for the interaction effect given the other variables in the model is 3.51 on 1 degree of freedom, with a p-value of 0.061.

8. Conclusion

There are strong theoretical objections to using OLS regression with ordinal response variables. The cumulative logit model was proposed by Peter McCullagh in 1980, and has found its way into use in biological and social research. This paper has emphasized some of the limitations of the cumulative logit model, but in truth there was no good alternative model. However, a line of development beginning with Leo Goodman's work in the late 1970s, and continuing down to the present, has led to the development of an alternative general modeling approach that indeed is more flexible, produces interpretable results, and makes relatively weak assumptions that can often be met in practice. This paper reanalyzed data originally presented in [Long, 1997] and showed that the adjacent-categories logit model can work well in situations where the cumulative logit model does not fit the data. It is hoped that this paper will help spur the wider use of this new modeling approach.

REFERENCES

Agresti, Alan. 1990. Categorical Data Analysis. New York: John Wiley.

Agresti, Alan. 1996. An Introduction to Categorical Data Analysis. New York: John Wiley.

Aldrich, John H., Nelson, Forrest D. 1984. Linear Probability, Logit, and Probit Models. Thousand Oaks: Sage.

Becker, M. 1989. On the Bivariate Normal Distribution and Association Models for Ordinal Categorical Data, *Statistics and Probability Letters* 8 (1989) 435-440.

Clogg, Clifford C. 1982. Using Association Models in Sociological Research: Some Examples, *American Journal of Sociology*, 88, pp. 114-134.

Clogg, Clifford C., Shihadeh, Edward S. 1994. Statistical Models for Ordinal Variables. Thousand Oaks: Sage Publications.

Cox, D.R., Wermuth, N. 1992. A comment on the coefficient of determination for binary respones. *American Statistician,* 46, 1-4.

Demaris, Alfred. 1992. Logit Modeling: Practical Applications. Thousand Oaks: Sage.

Goodman, Leo A. 1979. Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories. *Journal of the American Statistical Association*, 74, 537-552.

Goodman, Leo A. 1981. Association Models and the Bivariate Normal for Contingency Tables with Ordered Categories. *Biometrika*, 68, 347-355.

Goodman, Leo A. 1991. Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data. *Journal of the American Statistical Association,* December 1991, Vol. 86, No. 416, 1085-1111.

Hosmer, D.W., Lemeshow, S. 1989. Applied Logistic Regression. New York: John Wiley and Sons.

Ishii-Kuntz, Masako. 1994. Ordinal Log-Linear Models. Thousand Oaks: Sage.

Kennedy, Peter. 1992. A Guide to Econometrics. Cambridge: MIT Press.

Long, J. Scott. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks: Sage Publications.

Magidson, J. 1996. Maximum Likelihood Assessment of Clinical Trials Based on an Ordered Categorical Response. *Drug Information Journal,* Vol. 30, pp. 143-170.

Magidson, J. / Statistical Innovations Inc. 1998. GOLDMineR 2.0 User's Guide. Chicago: SPSS Inc.

Magidson, J., Babinec, A. 1998. A General Alternative to Linear Regression. Boston: Statistical Innovations Inc.

McCullagh, Peter. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society B,* 42, No. 2, pp. 109-142.

McKelvey, Richard D. 1975. A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *Journal of Mathematical Sociology,* Vol. 4, pp. 103-120.

Stewart, Mark B. 1983. On Least Squares Estimation when the Dependent Variable is Grouped. *Review of Economic Studies* (1983) L, 737-753.

Winship, Christopher, Mare, Robert D. 1984. Regression Models with Ordinal Variables. *American Sociological Review*, Vol. 49 (August:512-525).

FORECASTING SCANNER DATA BY CHOICE-BASED CONJOINT MODELS

Markus Feurstein
Vienna University of Economics & BA
Martin Natter
Vienna University of Economics & BA
Leonhard Kehl
GDS Data Service GmbH

ABSTRACT

Choice Based Conjoint analysis is considered the tool of choice for pricing in the U.S. Most Choice Based Conjoint studies are performed on the basis of aggregate models, given insufficient data for individual-level estimation of part-worths. It is known, however, that households can differ considerably. Therefore, managers are interested in the underlying segment structures and make use of a priori defined segments (e.g., choice frequencies) or apply segmentation procedures like k-means on the basis of additional data. DeSarbo, Ramaswamy and Cohen (1995) have proposed a latent class model that allows to consider heterogeneity within Choice Based Conjoint analysis without the necessity to collect additional data. Up to date, it is unclear how different segmentation approaches perform in terms of forecasting (external) real world aggregate shop data. In this contribution, we compare the performance of various segmentation approaches not with an experimental holdout sample but with aggregate scanning data. Our analysis shows that the ignorance of heterogeneity in the aggregate model leads to biased forecasts. In our study the latent class model turns out to perform significantly better than the other approaches considered. Furthermore, we demonstrate that the methodology of modeling heterogeneity has a strong impact on the optimal attribute level. We find that the consideration of a global correction vector is important. We show that this correction vector should be updated frequently while the price effects are valid for longer periods of time.

Keywords: Pricing Research, Choice Based Conjoint Analysis, Market Segmentation, External Validity

1. INTRODUCTION

Choice Based Conjoint (CBC) analysis (Louviere and Woodworth 1983) is one of the most frequently used methods for pricing decisions. The popularity of this method as compared to ranking based conjoint analyses is due to several advantages (see, e.g., Pinnell 1995, or DeSarbo, Ramaswamy and Cohen 1995) of this methodology:

- data collection: simulated purchase decisions are more realistic than rankings or ratings
- the derived part-worth utilities reflect impact on product choice rather than a change in ratings or rankings

- product specific attributes or levels can easily be accommodated and brand specific utilities can be estimated
- CBC analysis allows for a none-choice or other option
- conjoint design: CBC is more flexible in designs than traditional conjoint analysis

Most studies (e.g., Green, Krieger and Agrawal 1993) on the validity and performance of conjoint approaches rely on internal validity measures. The interview data are split into two parts, where the first is used for model estimation and the second (holdout sample) is used for model validation. Although predictive validity is of high practical relevance, holdout judgements are examined in only 9% of the projects (Wittink, Vriens, Burhenne 1994). Holdout samples are, however, only able to capture validity of preference or simulated choice, but not of actual behavior or real market shares. This is also the reason that one of the most useful concepts to validate and compare methodologies, i.e. Monte Carlo Analysis (see, e.g., Vriens, Wedel and Wilms 1996), does not provide final confidence into conjoint methodologies. Because of the fact that conjoint analysis is heavily used by managers[1] and investigated by researchers, external validity is of capital interest. To test external validity of conjoint analysis, in addition to experimental data collected in interviews, real world data sources (like POS scanning data) are necessary. As compared to the extensive use of CBC models, surprisingly few studies have been published that test external validity of this methodology. There is, however, a prominent group of researchers (Carson et al. 1994; Neslin et al. 1994; Winer et al. 1994) who have indicated the need for additional research concerning external validity of conjoint analysis. DeSarbo, Ramaswamy and Cohen (1995) propose to perform a cross-validation assessment of their approach and other alternatives to market segmentation. This is our primary objective.

In section 2, we describe the design of our validation study. This is followed by the investigation of different approaches to model consumer heterogeneity. In section 4, we study the impact of different segmentation models on optimal prices. The paper concludes with a discussion of the results and presents some remaining research topics.

2. Design of the Validation Study

In our study, the validity of different segmentation approaches is tested for 8 different brands of mineral water. In the market investigated, a country specific law allowed domestic firms to package mineral waters in glass packages only. In 1996, a market liberation took place and it became legal to use synthetic packages for mineral waters. It was expected that some of the major firms would introduce a 1.5 liter synthetic package. The producer under consideration, who is one of the major players in that market, considered to introduce a 1.5 liter package too and wanted to analyze different pricing scenarios for the new and the old package. With pricing of different packages as the main purpose of this conjoint study[2] - a CBC analysis was designed with the following attributes and levels: (a) brand names (A1-A8), (b) package sizes (B1-B2) and (c) prices (C1-C9). As different prices were used for 1 and 1.5 liter packages, interaction

[1] Wittink and Cattin (1989) document 1062 commercial applications of conjoint analysis of 66 firms in the United States for the period 1981-1985. For Europe, Wittink, Vriens and Burhenne (1994) report 956 uses of conjoint analysis by 59 firms for the period 1986-1991.

[2] The survey study of conjoint analysis by Wittink, Vriens and Burhenne (1994) identifies pricing as the number one purpose of conjoint studies in Europe.

effects between (b) and (c) were modeled. The personal interviews with a total of 128 respondents were collected at three different locations (shopping malls) using the Sawtooth interview software. Each respondent got shown 28 choice sets with five concepts per set plus a 'none-choice or other' option.

The respondents were asked to pick one of the concepts shown in a first choice and one of the four remaining concepts as the second choice. A second choice opportunity was only presented when the first choice was different from the 'none-choice or other' option. Apart from the choice experiments additional demographic and product specific information was collected: age, sex, household size, usage frequency and the region of the respondent.

Evaluation Methodology

From the choice data, we build probabilistic choice simulators and forecasted the market shares of all available products. As the 9 price levels chosen need not necessarily correspond with the actual shop prices, we interpolate the price utilities.

As an external validation of the approaches discussed, we use real world scanning data, which consists of 95 weekly price and sales data of the 1 liter packages of all eight brands and three 1.5 liter packages. To allow for a comparison between the segmentation models, we split the scanning data into two parts with 48 (quarters 1,3,5,7) and 47 (quarters 2,4,6,8) observations, respectively. The first part is used for the calculation of a correction vector (which accounts for external effects, see, e.g., Golanty 1995) of the conjoint models and the second part is used for validation of all models. The results reported concern the validation data only.

We measure the validity of the different approaches with the out-of-sample Variance Accounted For,

$$\mathrm{VAF}=1-\mathrm{MSE}(Sj)/\sigma^2(Sj),$$

where MSE(Sj) denotes the mean squared error between model-forecasts and the actual shares of brand j, and $\sigma^2(Sj)$, the standard deviation of the shares of j. A model with VAF=1 explains all variance in the data. If average estimation and validation market shares differ considerably, VAF can also become negative. Often management is not only interested in a high correlation between true and estimated market shares but in good estimates of the real market shares. Therefore, we include the VAF measure and do not only focus on R^2, which is insensitive to different levels. For the sake of brevity, we present market share weighted measures and not those of individual brands. In addition to the variance explanation measure, we report the price elasticity[3] ($\varepsilon_{S,P}$) calculated at the average price of each brand and the attribute importance.

[3] Price elasticity is calculated numerically from a price reduction of 1% from the regular price.

3. Consumer Heterogeneity and External Validity

External Effects

In a market with new products, other effects than price, brand or package may lead to significant changes of shares. Building up channels of distribution may cause shifts in market shares (Golanty 1995). Life-cycle theory suggests that products follow a certain pattern of ups-and-downs during their time on the market. However, conjoint models do not account for such effects. Due to the cross-sectional time-series nature, scanner panel data are an interesting source to improve choice modeling (Winer et al. 1994). Usually a (static) global correction vector is applied to adjust for external effects. If, e.g., market shares are known from the ACNielsen Retail Index (NRI), this correction vector can be calculated as the fraction between average shares of the NRI and the shares forecasted by the CBC model at average observed prices. In Figure 2, we demonstrate the effect of different correction schemes for one specific brand. The first curve shows the real market share of the validation periods while the second curve displays the predicted shares of choice for the actual shop prices. One clearly sees that without the consideration of external effects neither the level nor the evolution of its share are met. This holds for all other models discussed below. The highest VAF measure without the consideration of a correction vector equals VAF=8.7% (see Table 6) for the latent class model although the corresponding R^2 reaches 61.8% for the same model. It must be noted that the level of the shares determines the optimal attribute (here price) level. Therefore, the use of CBC **share-of-preference estimates should in general not be taken as forecasts of the market shares** without adjustment of the external effects. An approach to cope with that problem is the calculation of a correction vector.

To adjust shares of choice of the estimated latent class model to market shares we determine a correction vector from the first in-sample quarter and multiply this vector[4] with all the remaining latent class forecasts. Figure 2 shows that this improves the forecasts of the latent class model (third curve). However, the forecasting accuracy decreases over time as this brand loses market shares in later periods. When the correction vector is updated quarterly market evolution effects are better under control (fourth curve of Figure 2). In the following, we compare external validity of various segmentation models and take external effects into account by a quarterly updated correction vector.

Aggregate Level CBC

Allowing only one alternative to be chosen per concept CBC interviews try to mimic real shopping situations. However, as compared to ranking or rating based conjoint approaches, this methodology leads to very ‘’sparse’’ data. Therefore, choice data are most often analyzed by first aggregating data (Johnson 1997). Aggregation over respondents assumes that all respondents are similar since aggregate methods cannot account for real differences among the respondents. With respect to the large amount of literature (cf. Hagerty 1985) concerning consumer heterogeneity this assumption seems to be inappropriate. Maximum likelihood estimation of this CBC model shows that the aggregate model accounts for VAF=52.9% of the variance in real

[4] The use of a factor has no effect on the price elasticity.

shop data. Calculation of the importance values indicates that price (49%) is the most important attribute, followed by brand (41%) and package (10%)[5].

The average price elasticity (weighted by market share) of the aggregate model (Table 6) is -1.80. Due to aggregation one would expect that the aggregate model rather understates price effects. However, visual inspection of aggregate forecasts versus the shop data yields that the price effect are overestimated for some brands. This model can be seen as a benchmark for the other approaches that do consider heterogeneity between respondents.

Individual Level CBC

Due to the mentioned data problem, only few CBC studies use individual-level estimates (number of respondents) to construct a market simulation model. Therefore, we expect that individual estimates will not be able to accurately forecast aggregate shop data. Surprisingly, the external validity for the individual level analysis (VAF=59.8%) is 6.9% points higher than for the aggregate model. This indicates that the gain due to modeling heterogeneity exceeds the effect of over-fitting sparse data. Hence, variety among individuals seems to be an important issue that should not be omitted. Segmentation models try to overcome the weaknesses of the above approaches, ignorance of heterogeneity and over-fitting, respectively.

CBC with A Priori Segments

With regard to data aggregation, the study of Wittink, Vriens and Burhenne (1994) reports that 39% of the projects use a priori segments and that 50% of the firms have used a priori segments at least once. Although predefined segments are the most frequently used segmentation approaches in practice, it is not obvious which set of background variables should be used as segmentation criteria. We study two commonly used segmentations, regional and usage based clusters.

Regional Segmentation

We use the different regions in which the respondents were interviewed as a priori segments and estimate distinct choice models for each region. This seems interesting because significant differences in the market shares of the individual brands can be identified in the three different areas. In accordance with the number of regions, we fix the number of segments, S, to S=3 . Table 1 contains the attribute importance of the 3 regions including a weighted average over the segments. To members of region 3 brand is the attribute with highest importance. Price is the most important attribute in segment 2. In this region package is of diminishing importance and brand is less important than on average. Segment 1 is close to the average respondents, only package is higher than average. The predictive accuracy of the regionally aggregated CBC models (see Table 6) is 3.9% points better than our benchmark model but 3 points below the individual-level model.

[5] Note that these aggregate importances are different from the weighted averages that we report for the segment analyses in Tables 1,2,4 and 5.

Segmentation by Usage Frequency

Alternatively to the region-based a priori segments, we also test another frequently undertaken segmentation: based on the additional information about usage frequency of mineral water, we defined the following 3 segments: heavy users (who use mineral water several times a week), normal users (weekly usage), and light users (less than once a week).

The attribute importance calculated from the CBC-part-worths (see Table 2) indicates that price is most important in all segments and that brand is more important to frequent users than to normal and light users. The external validity of this approach (VAF=53.8%) is only marginally better than for the aggregate model. Thus, for the market analyzed, a priori segmentation on the basis of usage frequency, does not properly separate households.

CBC with K-means Segments

A frequently undertaken approach is to determine segments on the basis of additional information (sex, age, and household size) collected during the interviews. All variables were dummy coded. Here, the k-means clustering procedure is used to determine the segments. For the purpose of comparability, we fixed the number of segments to 3.

Aggregate and segment-specific percentages of all attributes are shown in Table 3. Sex only has a low discriminatory influence on the cluster solution. It can be seen that age is the main discriminating attribute followed by the size of the household, a respondent belongs to. Respondents of segment 1 are between 30 and 50 years old, and typically live in households with 3 or 4 persons. Members of segment 2 are aged higher than 50 years and stay in a 2 persons household. Young (age<30) persons belong to segment 3. The probability that segment 3 respondents live in a 4 persons household is more than twice as high as for the average respondent. Because of the fact that the k-means segments can be identified through only one variable (age) this solution is identical to an a priori segmentation which uses age as background variable.

The segment specific attribute importance (see Table 4) indicates that brand and package are (relative to the average) important for members of segment 2 and that price is most important to members of segment 1 and 3 (aged<50) but not to segment 2 members.

The k-means segmentation approach improves external validity as compared to the aggregate model by 3.2 percentage points and lies between the two a priori approaches. The collection of additional data used for this and the a priori approaches discussed above, increases interview time and costs of such studies. Therefore, a segmentation method without additional data requirements would be desirable. The following approach meets these requirements.

The Latent Class CBC Model

In ranking or rating based conjoint models the individual-level part-worths can be used as a basis for clustering. As it is often inefficient to estimate individual part-worths with a limited number of choices, this approach is usually not recommended for CBC studies. DeSarbo, Ramaswamy and Cohen (1995) propose to use a latent class version of CBC to overcome the limitations of aggregate analyses or a priori segmentations. The authors generalize the Kamakura, Russell (1989) scanner data response methodology to a latent class CBC model considering within subject replications over choice sets. They propose to use a maximum likelihood procedure to estimate the segment specific parameters of the equation[6]. For the purpose of comparability we use a three class solution[7].

Importance values of brand, price and package are presented in Table 5. Brand – with an importance value of 68 – is the driving force in the first latent class, whereas price has a strong impact on the preference of members belonging to the second class. The third class lies in between the first two classes regarding all three attributes. Weighted with cluster sizes price is the most important attribute followed by brand and package. Package is the least important attribute in all classes, with the second class showing the highest relative importance value for package. Surprisingly the mean price elasticity (see Table 6) is comparable to those of the aggregate model. However, the forecasting accuracy for the latent class model is higher for ten out of eleven brands. Therefore, the price elasticity of the aggregate model is biased leading to wrong price decisions for most brands (see next section). Figure 1 shows real shop data of a 1 liter brand and forecasts of the latent class and the aggregate model. While the latent class model nicely follows ups-and-downs of the real shop data the forecasts of the aggregate model understate the price effects for this specific brand. Table 6 shows that the latent class model produces the most valid forecasts among the approaches analyzed. It improves forecasting accuracy as compared to the aggregate model (on the average) by 13.5% points.

Impacts on the Optimal Price

Table 6 shows that different approaches to model heterogeneity can result in very different levels of forecasting accuracy and price elasticity. In our study, the manufacturer wanted to apply CBC to optimize prices of the 1 and 1.5 liter packages. Therefore, we investigated the sensitivity of the price-optimum on different segmentation models. As only the cost-data for one brand
(1 and 1.5 liter package) are available, we restrict our sensitivity analysis to this brand. To cover the true costs and optimal prices we report only deviations (in percentages) from the optimal price of the latent class model. For the optimization prices of the competitive brands were fixed at their regular price. The results (see Table 7) indicate that the choice of the segmentation approach considerably affects optimal prices under realistic cost assumptions. Differences between the optimal prices of the aggregate and the individual-level CBC are as high as 34% for 1.5 liter packages. This indicates that the optimal prices are very sensitive to the segmentation model (cf. Vriens, Wedel and Wilms 1996). As the latent class model has the highest forecasting accuracy we recommended the optimal prices of this model.

4. Conclusion

In our study we examined the performance of different segmentation models by use of external validation based on POS scanning data. Due to external effects it is necessary to introduce a correction scheme in order to map the CBC shares of preference to market shares. The fractions between the shop data and the model's forecasts at average prices serve as a correction vector for the next quarter. Without any correction the out-of-sample performance is less than VAF=10% for all models under consideration. Therefore if one is interested in forecasts of market shares

[6] For our analysis we use the maximum likelihood estimation procedure implemented in the Sawtooth CBC LC software.
[7] Winer et al. (1994) find that latent class approaches seem to handle household heterogeneity fairly well with a small number of segments.

instead of shares of choice external effects must be taken into account. Our results show that scanning data are useful to improve the external validity of CBC models by introducing dynamics of the market to the static conjoint models through quarterly updated correction vectors.

The aggregate model, which is widely used in practice, performed worst in terms of external validity. Contrary to our expectations, forecasts based on individual-level part-worth estimates, were significantly better than aggregate forecasts. This emphasizes the necessity of disaggregate models. The forecasting accuracy of CBC models with a priori defined segments depends on the set of background variables that are used as segmentation criteria. The first a priori model which was based on regional segments was slightly better than the aggregate model. The second a priori segmentation model that we have tested was based on usage frequency. This model hardly improved aggregate forecasts. Determining segments via the k-means clustering procedure on the basis of sociodemographic (sex, age, household size) information only slightly improved validity as compared to the aggregate model. The highest forecasting precision was achieved by the latent class model proposed by DeSarbo, Ramaswamy and Cohen (1995).

Although all models rely on the same CBC data, the type of approach to modeling consumer heterogeneity has a significant influence on the external validity and on the optimal attribute level. We have demonstrated that for the same data source, the optimal prices can differ by more than 30% when different segmentation approaches are applied.

In summary, our studies induce the following practical rules for forecasting market shares from CBC models: The heterogeneity of households is best captured by the latent class model which results in the highest forecasting accuracy. In order to capture external effect within static CBC analysis a quarterly updated correction vector should be applied. In this contribution we examined the external validity for one product only. In future it would be interesting to see papers that investigate the impact of consumer heterogeneity and external effects on external validity for different products, numbers of respondents and attributes.

REFERENCES

Carson, R.T., J.L. Louviere, D.A. Anderson, P. Arabie, D.S. Bunch, D.A. Hensher, R.M. Johnson, W.F. Kuhfeld, D. Steinberg, J. Swait, H. Timmermans, J.B.Wiley (1994), "Experimental Analysis of Choice," Marketing Letters 5:4, 351-368.

DeSarbo, W.S., V.Ramaswamy, S. Cohen (1995), "Market Segmentation with Choice-Based Conjoint Analysis," Marketing Letters 6:2, pp. 137-147.

Golanty, J. (1995), "Using Discrete Choice Modeling to Estimate Market Share, Journal of Marketing Research," Vol. VII, pp. 25-28.

Green, P.E., A.M. Krieger, M.K. Agrawal (1993), "A Cross Validation Test of Four Models for Quantifying Multi-attribute Preferences, Marketing Letters," Vol. 4/4, pp. 369-380.

Hagerty, M.R. (1985), "Improving the Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis," Journal of Marketing Research, Vol. XXII, pp. 168-184.

Johnson, R. M. (1997), "ICE: Individual Choice Estimation," Sawtooth Software Inc., Working Paper.

Kamakura, W.A., G.J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," Journal of Marketing Research, XXVI, pp. 379-390.

Louviere, J.J., G.G. Woodsworth (1983), "Design and Analysis of Simulated Choice or Allocation Experiments: An Approach Based on Aggregate Data," Journal of Marketing Research, 20, pp. 350-367.

Neslin, S., G. Allenby, A. Ehrenberg, S. Hoch, G. Laurent, R. Leone, J. Little, L. Lodish, R. Shoemaker, D. Wittink (1994), "A Research Agenda for Making Scanner Data More Useful to Managers," Marketing Letters 5:4, 395-412.

Pinnell, J. (1994/1995) "Multi-Stage Conjoint Methods to Measure Price Sensitivity," Sawtooth News, Ketchum, ID: Sawtooth Software, Inc., edited by Weiss, S., Vol. 10/2, pp. 5-6.

Vriens, M., M. Wedel, T. Wilms (1996), "Metric Conjoint Segmentation Methods: A Monte Carlo Comparison," Journal of Marketing Research, Vol. XXXIII, pp. 73-85.

Winer, R.S., R.E. Bucklin, J. Deighton, T. Erdem, P.S. Fader, J.J. Inman, H. Katahira, K. Lemon, A. Mitchell (1994), "When Worlds Collide: The Implication of Panel Data-Based Choice Models for Consumer Behavior," Marketing Letters 5:4, 383-394.

Wittink, D.R., P. Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," Journal of Marketing 53, 91-96.

Wittink, D.R., M. Vriens, and W.Burhenne (1994), "Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections," International Journal of Research in Marketing 11, pp. 41-52.

APPENDIX

Table1: Regional segments: aggregate and segment specific attribute importance

Attribute	weighted average	s1 (33.3%) region1	s2 (34.4%) region2	s3 (32.3%) region3
brand	45	44	41	51
package	9	13	1	13
price	46	43	58	36

Table 2: Usage segments: aggregate and segment specific attribute importance

Attribute	Weighted Average	s1 (13.4%) heavy	s2 (39.0%) normal	s3 (47.6%) light
brand	40	44	38	40
package	10	10	6	13
price	50	46	56	47

Table 3: K-means: aggregate and segment specific variable levels

Attribute	aggregate	s1 (47%)	s2 (27%)	s3 (26%)
sex=female	0.74	0.78	0.71	0.74
sex=male	0.24	0.21	0.28	0.26
age <30	0.26	0.00	0.00	1.00
Age <50	0.47	1.00	0.00	0.00
age >50	0.27	0.00	1.00	0.00
1 persons household	0.07	0.06	0.11	0.06
2 persons household	0.36	0.23	0.62	0.36
3 persons household	0.23	0.30	0.17	0.19
4 persons household	0.21	0.35	0.02	0.19
>4 persons household	0.08	0.05	0.05	0.19

Table 4: K-means: aggregate and segment specific attribute importance

Attribute	weighted average	s1 (47%) age 30-50	s2 (27%) age >50	s3(26%) age <30
Brand	43	39	49	44
Package	10	8	15	8
Price	48	54	36	49

Table 5: Latent class: aggregate and segment specific attribute importance

Attribute	weighted average	s1 (42%)	s2(44%)	s3(14%)
Brand	46	68	21	58
Package	7	2	13	5
Price	47	30	66	38

Table 6: Market share weighted Variance Accounted For (VAF) and R^2 for all models with and without use of a correction vector in % and price elasiticities (of the segments) for all brands

Model	VAF	VAF	R^2	R^2	$\varepsilon_{S,P}$
Uses correction vector	no	yes	no	yes	
Aggregate Level CBC	7.4	52.9	59.7	66.7	-1.80
Individual Level CBC	2.1	59.8	67.0	74.0	-1.00
A priori (regional) segmentation CBC	0.0	56.8	63.5	72.3	-1.50
A priori (usage intensity) segmentation CBC	0.0	53.8	60.0	68.0	-1.57
K-means (criteria: age, sex, size of household)	0.0	56.1	61.8	71.2	-1.46
Latent Class CBC	8.7	66.4	72.5	77.2	-1.79

Table 7: Deviations (in %) of optimal prices (1 and 1.5 liter) from the latent class optimal price for one specific brand

Model	deviation from LC 1 liter	deviation from LC 1.5 liter
Aggregate Level CBC	0%	-14%
Individual Level CBC	27%	20%
A priori (regional) segmentation CBC	20%	-9%
A priori (usage intensity) segmentation CBC	10%	-11%
K-means (criteria: age, sex, size of household)	17%	-7%
Latent Class CBC	0%	0%

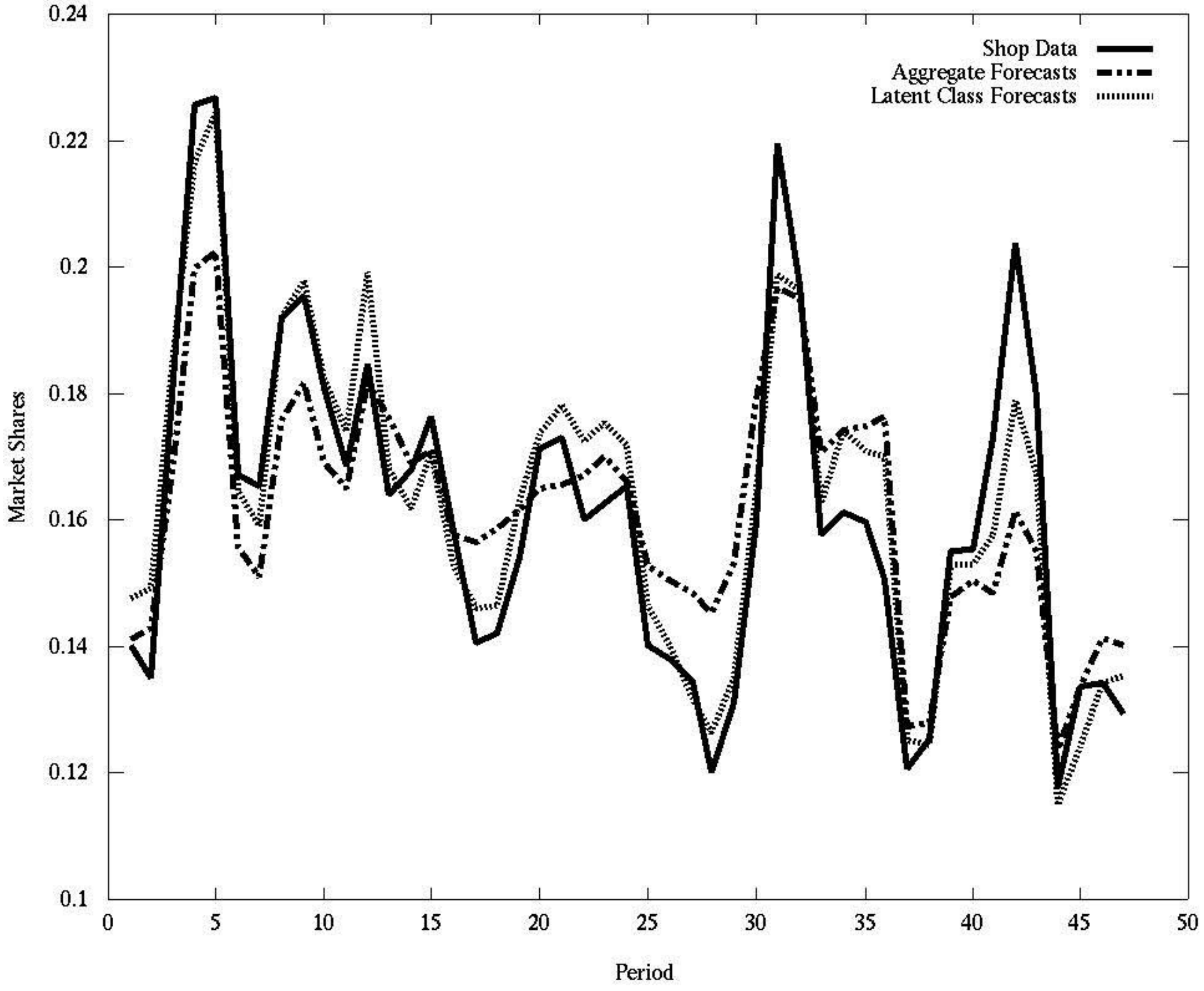


Figure 1:
Real shop data and forecasts of the latent class CBC and the aggregate CBC model for one brand

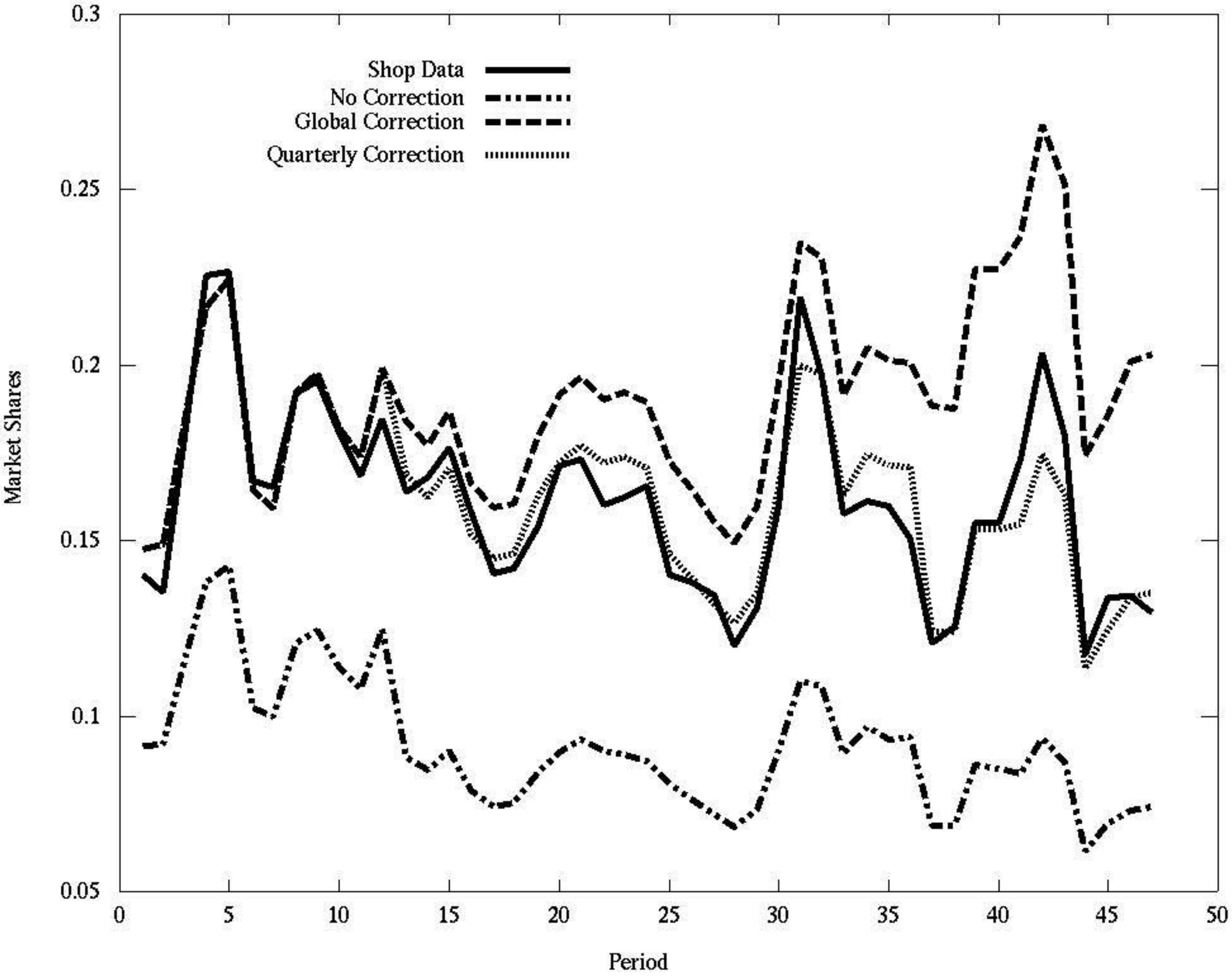


Figure 2:
Real shop data, shares of choice, global update and quarterly update of the correction vector for the latent class model

PREDICTING ACTUAL SALES WITH CBC: HOW CAPTURING HETEROGENEITY IMPROVES RESULTS

Bryan K. Orme
Sawtooth Software, Inc.
Michael A. Heft
Daymon Associates

INTRODUCTION

Conjoint analysis has been used extensively over the last three decades in marketing, but few published studies have demonstrated that it can predict actual sales. We believe the sparse evidence is not because conjoint cannot predict real world behavior (if that were the case, why would conjoint analysis continue to be so popular?), but that the data are guarded by organizations with no real incentive to publish the results. We report on a study wherein shoppers at grocery stores were given a CBC interview, and the results used to predict actual sales within the same stores.

The second focus of this paper is to demonstrate that capturing heterogeneity (differences in preference between groups or individuals) can improve predictive validity. Traditionally, CBC has been analyzed in the aggregate, by pooling all respondents and developing a summary set of effects (utilities) to reflect the market. Lately, methods that capture heterogeneity by modeling utilities at the group-level (Latent Class) and even at the individual level (Hierarchical Bayes and ICE–Individual Choice Estimation) have become available and been heralded as better models. After showing that Lclass and ICE do a better job at predicting actual sales for our study than aggregate level logit, we'll spend the remainder of this paper investigating why.

THE GROCERY STORE STUDY DESIGN

Six-hundred respondents were intercepted within five grocery stores and completed a computerized survey programmed using Ci3. The interview facilitators approached every *n*th customer and assisted with running the survey. A randomized choice experiment that included scanned images of the products was programmed into the questionnaire. Three different choice designs, each covering a different product category, were included. Respondents were asked to indicate which product categories they often purchased, and were randomly selected to complete a CBC interview for a product category for which they qualified. For proprietary reasons, we cannot reveal the grocery store chain nor the categories and brands that were studied. We can say that the purpose of the conjoint research was to determine pricing strategy.

The completed interviews per category were as follows:

Category I	246
Category II	205
Category III	149

Three attributes were included in each design:

1 Brand (picture of the package)

2 An unimportant "decoy" attribute

3 Price (a conditional, customized range for each brand)

Each respondent completed 15 choice tasks, and "None" choices were permitted. The "decoy" attribute was not used in modeling, but helped disguise the purpose of our research (pricing) for the respondents.

Four price points were chosen for each brand, such as:

1 25% lower than average price

2 12.5% lower than average price

3 12.5% higher than average price

4 25% higher than average price

Prices were rounded to the nearest 9-cent increment to better reflect the way these products are actually priced on the shelf.

THE VALUE OF GOOD PRICING INFORMATION

Pricing is not only a highly sensitive topic in marketing, but also one that has a major profitability impact. In spite of its importance, pricing is a difficult topic for most managers. A recent McKinsey & Company survey asked managers from over 300 North American companies whether they had done any research to measure or predict price elasticity in the previous year (Clancy and Shulman, 1994). Only 15 percent reported doing any kind of primary research. Consistent with those findings, a survey by marketing consultants Clancy and Shulman, found that only about 12 percent of all American companies do any serious pricing research, and one-third of those have no strategy with which to use the research (Clancy and Shulman, 1994).

Pricing in the supermarket environment is even more difficult since the average supermarket is dealing with over 30,000 stock keeping units (SKUs). Pricing decisions must consider cost, marketing strategy and profitability. In the highly competitive consumer package goods arena, supermarkets must also consider the explicit signal that individual "marker items" convey to consumers regarding the general price levels of that store chain. Thus, the price of a very few frequently purchased items can create a consumer price image that will influence future choice of that chain as a regular shopping declaration.

A specific price issue, that of the optimum pricing relationship between different brands, is one of the most frequent questions that consultants to retailers are asked to answer. Our client commissioned this project because of that need. As a bonus, the data were very useful for validating different methods of analysis.

THE VALIDATION SALES DATA

Actual units sold as recorded by checkout scanners for the last 52 weeks (reported by week) were provided by the client. Here are some details on each category:

Category I:

> A food product ranging in price from $1.29 to $2.49. This product was the fastest moving (in terms of unit sales) with about 80 units sold per week in each store. Prices changed throughout the 52-week period, with all brands holding constant for about 8 weeks at a time. When brands went on sale, there was only modest evidence of stocking-up behavior. Three brands were studied in total.

Category II:

> A non-food product ranging in price from $4.19 to $9.19. This product was the slowest moving, with about 5 units sold per week in each store. Prices changed throughout the period, with prices holding constant for between 8 to 24 weeks at a time. Only modest evidence of stocking-up behavior was detected. Five brands were studied in total.

Category III:

> A food product ranging in price from $1.69 to $3.89. This product sold about 60 units per week in each store. Prices remained constant for most of the weeks, with certain brands going on sale often throughout the year for only a week or two at a time. When brands went on sale, sales volume temporarily and dramatically increased (by a factor of as much as 6x), reflecting stocking-up behavior. Six brands were studied.

Market simulators based on conjoint analysis reflect steady-state, long-range demand. We should not expect conjoint to accurately predict transitional pricing periods for our grocery store categories. Significant stocking-up behavior can occur, and the degree of that behavior depends on factors such as the shelf life, physical size, households' purchase frequency of the product, and how often the product goes on sale.

For each category, different validation scenarios were chosen, with preference given to periods in which prices were held constant for many consecutive weeks. For category III, only one such stable period could be accumulated. Seven other validation scenarios were constructed from the other two categories. Of the 156 total weeks of sales data available to us (52 weeks for three categories), we discarded 45 transitional weeks wherein prices had recently changed. Viewed from the other perspective, we retained 111/156 = 71% of the observations.

DIFFERENT METHODS FOR ANALYZING CHOICE DATA

We developed market simulators using three different methods for modeling CBC data:

1. Aggregate Logit: A single set of utilities summarizes the preferences of the entire sample. Both main-effects and interaction terms can be modeled.
2. Latent Class (LC): The method simultaneously divides the sample into market segments that differ in preferences, and estimates utilities summarizing the preferences for each group. Both main-effects and interaction terms can be modeled.
3. Individual Choice Estimation (ICE): A set of utilities is estimated to reflect the preferences for each individual. The software currently only provides for main-effects.

Another method besides ICE has proven to be useful for estimating individual-level utilities from choice data: Hierarchical Bayes (HB). Though we did not try using HB for our data set, we expect it would have also worked very well given that ICE and HB have been shown to provide very similar results (Huber 1998).

VALIDATION PERFORMANCE

Five CBC simulators were developed for each product category:

1. Aggregate logit, main-effects only (Logit ME)
2. Aggregate logit, main-effects plus Brand x Price interaction (Logit BxP)
3. Latent Class, main-effects only (LC ME)
4. Latent Class, main-effects plus Brand x Price interaction (LC BxP)
5. ICE, main-effects only (ICE)

The LC simulators used six segments each. ICE utilities were computed using LC vectors as starting points with no additional iterations. The ICE utilities were not calibrated, but used in their raw form. All models employed logit simulations (Model 2), with no external effects and the scale parameter (exponent) set to unity.

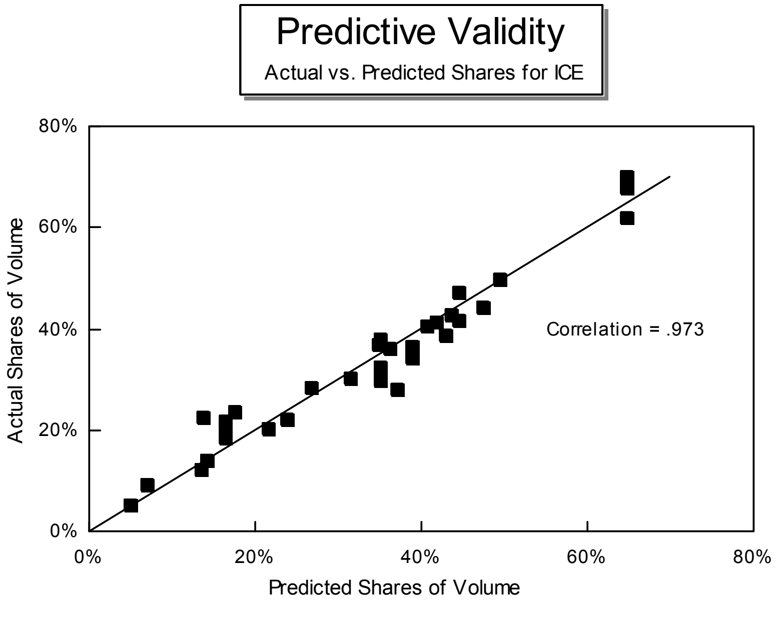
The actual and predicted shares for all validation scenarios are summarized in Table 1, using two measures of fit: MAE (Mean Absolute Error) and Correlation.

Table 1
Predictive Validity

	MAE	Correlation
Logit ME	4.65%	0.905
Logit BxP	3.92%	0.951
LC ME	3.14%	0.967
LC BxP	3.31%	0.967
ICE	2.87%	0.973

ICE and LC appear to predict actual sales better than aggregate logit. A statistical test (F-test) of the differences in predictive ability suggests that the only significant differences (at or above the 95% confidence level) are for the LC and ICE models relative to Logit ME. Ours is not the first study to conclude that capturing heterogeneity improves predictions. Studies using synthetic data (Johnson, 1997a), respondent data with holdout choices (Johnson, 1997a; Huber and Orme, 1999), and real sales data (Natter *et al.,* 1998) have also demonstrated that recognizing heterogeneity improves results. Plotting the best model (ICE) versus real market shares shows a remarkable correlation of .973 between predicted and actual shares (we've added a 45-degree line to represent the line of perfect prediction.)

Figure 1



We were delighted with how accurately the conjoint simulations predicted the shares, especially in light of our relatively modest sample sizes. For the majority of the observations (categories I and III), no adjustments (scale or external effects) were made. For category II, shares for the premium and discount brands were modeled separately to adjust for a significant difference in the amount of shelf space given to the premium versus discount brands.

For our study, adding interaction terms to aggregate logit improved the predictions, but it still fell short of the simpler main-effects LC and ICE models (see Table 1). We didn't investigate additional terms to account for similarity effects and cross-elasticities, but speculate that they would have improved predictions for aggregate logit. Also note that adding interaction terms to the LC model did not improve validity, and the ICE model with no interactions terms was the most successful. This finding suggests that disaggregate approaches may somehow account for complex effects (such as interactions) with main-effects models. We'll provide more evidence on this later.

The Role of the Scale Factor

At the onset of this project, we expected we would need to adjust the scale factor (exponent) to accurately predict market shares. The scale factor controls the flatness or steepness of share estimates. The more random noise in the conjoint responses, the flatter the share predictions (and vice-versa). We were pleased (and a little surprised) that no significant tuning was justified. Our respondents evidently made choices in the CBC tasks with roughly the same degree of attention as buyers in general give to the real world purchases. This may not always be the case for other product categories and situations where conjoint is used to predict market shares.

Though there were small differences in the implied scale factor (based on the variance of the predicted shares) between the aggregate logit, LC and ICE models, one cannot argue that these account for the differences in predictive validity for our study. We experimented with tuning the exponent for the aggregate logit models, but that resulted in insignificant overall improvements in MAE and correlation. Also, computing correlations between share predictions and actual sales (as shown in Table 1) largely, but not entirely, controls for scale.

Our findings agree with other researchers who have suggested (Johnson, 1988) or demonstrated (Brice, 1997) that the Share of Preference (logit) model generally predicts actual market shares better than the more extreme First Choice model. (The First Choice model is equivalent to a Share of Preference model with a very high exponent.) Though the Share of Preference scaling worked best for our study, there may be specific instances (high involvement purchases, relatively noisy conjoint data) where the First Choice model excels.

NOTES AND CAVEATS FOR PREDICTING SALES WITH CBC

Why did our CBC simulations predict sales so well?

1. We exercised considerable control by interviewing respondents in the same stores that contributed the validation sales data.
2. Brands in Categories I and III were equally available on the shelves (roughly same shelf space) in all five stores. CBC interviews mimic this presentation, since each brand is equally represented (in terms of computer screen real estate) in choice tasks. Category II had very unequal representation on the shelf between the premium and discount brands, so we separately modeled shares among them.
3. Conjoint analysis reflects stable, long-term demand. None of the products we studied were new introductions, and we carefully selected validation periods by accumulating sales data across weeks in which the prices had remained constant for enough time to reasonably stabilize demand. As mentioned before, we retained 71% of the total sales data available to us.

Additionally, we speculate that the following helped:

4. Pictorial representation of the products.
5. Conditional pricing, to reflect realistic prices.
6. We chose categories that had few brands and included only the most commonly purchased package size of each.

Now that we have demonstrated that ICE and LC outperformed aggregate logit in predicting actual sales for our study, we'll spend the remainder of this paper discussing why capturing heterogeneity improves predictive validity.

IIA (INDEPENDENCE FROM IRRELEVANT ALTERNATIVES)

The logit model (described later) is governed by a property called IIA (Independence from Irrelevant Alternatives). The property dictates that products in a market simulation take share from other products in proportion to their respective shares. At first, this property was thought beneficial (Johnson, 1997b), but it poses problems for accurate market representations. Unless terms are specifically modeled to directly capture complex effects, aggregate level logit models ignore unequal substitution patterns between products, differential cross-elasticities, and interactions.

THE "RED BUS/BLUE BUS" PROBLEM

IIA is often illustrated with the "Red Bus/Blue Bus" example. In that example, different modes of transportation are available, such as cars and red buses. The example reflects a classic line extension problem wherein the bus company decides to repaint half of its buses blue in hopes of increasing bus ridership. Although we wouldn't expect that move to significantly increase ridership (since color is very unimportant to potential bus riders), aggregate level logit models will unrealistically inflate the net bus ridership due to the IIA property.

Our grocery store data can be used to demonstrate this principle, and how capturing heterogeneity helps resolve the IIA problem for classic line extension problems. One of the three product categories we studied was a product that came in a plastic bottle. In addition to brand and price, we included a “decoy” attribute that specified whether the bottle was round or square. The shape of the bottle was virtually unimportant to our respondents, both on average, and within all six LC segments we analyzed.

The simulation results below show the Share of Choice under main-effects aggregate level logit for five fictitious products, each at their average price:

Brand and Form	**Share of Choice**
Brand A, Square	16.7%
Brand B, Round	17.9
Brand C, Square	8.4
Brand D, Round	28.6
Brand E, Square	28.4
	100.0%

Suppose the brand manager for Brand A wanted to know if offering his product in both a square and a round bottle would significantly increase its net share. We can use the conjoint simulator to respond to that question. But depending on whether we capture heterogeneity or not, we’ll get a vastly different answer. The simulation below is the same as above, but with Brand A also offered in a round bottle:

Brand and Form	**Share of Choice**	
Brand A, Square	14.4%	
Brand A, Round	*13.7*	(Net Brand A = 14.4 + 13.7 = 28.1)
Brand B, Round	15.5	
Brand C, Square	7.2	
Brand D, Round	24.7	
Brand E, Square	24.5	
	100.0%	

The aggregate level logit simulator suggests that the line extension will increase the net share to Brand A from 16.7% to 28.1%, or a 68% increase. If you were the brand manager, would you believe it?

Capturing heterogeneity with LC or ICE results in more realistic answers. Using the same simulation scenarios as above, we calculated the net increase to Brand A for the line extension under 2- through 8-group LC solutions, and with ICE.

Table 2
Line Extension Example
Relative Share Increase for Brand A

Method	Increase
Aggregate Logit	+68%
2-Group LC	+47%
3-Group LC	+37%
4-Group LC	+31%
5-Group LC	+24%
6-Group LC	+25%
7-Group LC	+19%
8-Group LC	+23%
ICE	+11%

The share inflation problem is reduced considerably as we recognize more heterogeneity. ICE (which fits a set of utilities for each individual) shows the least amount of share inflation–and we'll argue is probably the most realistic answer.

Why does capturing heterogeneity reduce share inflation for similar products? Consider the case of a LC solution. Utilities customized within each group of respondents fit respondent choices better than a single aggregate set of logit utilities. When logit/LC utilities better fit the data, they become larger (in absolute value). The shares of preference for products *within* each segment become more extreme because the utilities have greater variance. (This is the same familiar situation as increasing the exponent in Sawtooth Software simulators.) As utilities become more extreme, simulations begin to behave more like the First Choice model within each segment, which is immune to IIA problems. When results are averaged across segments, the overall scaling is very close to the original aggregate level logit, so the sensitivity of the market simulation model remains largely unaffected.

CROSS-ELASTICITIES

Another weakness of aggregate-level logit is its inability to account for cross-elasticities with the standard main-effects or main-effects-plus-interactions models. (One can model cross-elasticities with aggregate-level logit, but that requires including many more parameters to be estimated.)

Cross-elasticity is defined as the relative percent change in quantity demanded of brand A resulting from a percent change in price of brand B. Recall that with aggregate-level logit, IIA dictates that when a brand lowers its price, it steals share from other brands in proportion to the other brands' shares. In other words, the cross-elasticities are held constant.

We can use the line extension example from the previous section to illustrate cross-elasticity. Recall that we ended up with six total products (after Brand A was released in both the square and round bottle). What happens if Brand A, square bottle lowers its price by 20%? Aggregate logit simulations reveal the following:

Table 3
Constant Cross-Elasticities under Aggregate Main-Effects Logit

Product	Share at Avg. Price	Brand A, Square Lowers Price by 20%	Percent Change in Share
Brand A, Square	14.4	21.7	+51%
Brand A, Round	13.7	12.6	-8%
Brand B, Round	15.5	14.2	-8%
Brand C, Square	7.2	6.6	-8%
Brand D, Round	24.7	22.6	-8%
Brand E, Square	24.5	22.4	-8%

The results are consistent with IIA (constant cross-elasticities). In reality, we'd expect the square form of Brand A to take relatively more share away from the round form of the same brand than from the remaining products. ICE simulations suggest that behavior:

Table 4
ICE Cross-Elasticity Example

Product	Share at Avg. Price	Brand A, Square Lowers Price by 20%	Percent Change in Share
Brand A, Square	9.3	17.0	+83%
Brand A, Round	9.3	7.0	-25%
Brand B, Round	15.6	13.6	-13%
Brand C, Square	6.4	5.4	-16%
Brand D, Round	33.1	31.2	-6%
Brand E, Square	26.3	25.8	-2%

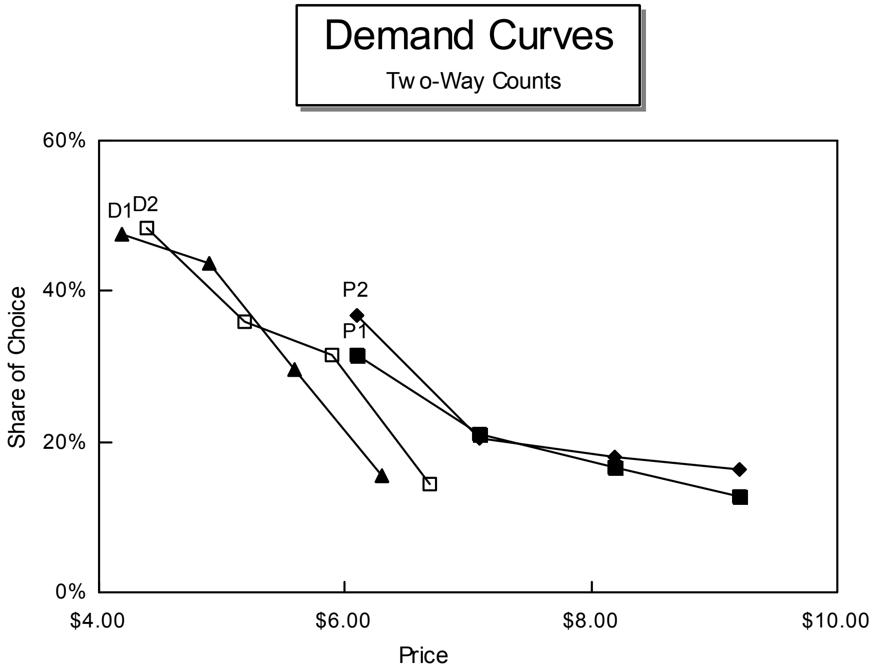
If we account for respondent heterogeneity (with LC or ICE), some degree of differential cross-elasticity can be captured and modeled using only main-effects models.

INTERACTIONS

One of the celebrated benefits of aggregate-level logit is the ability to model interactions, such as those between brand and price. However, if interactions result from differences in preference between segments or individuals, these can also be captured by recognizing heterogeneity with LC or ICE *without having to include interaction terms*.

The following graph displays two-way Count probabilities from CBC, representing the probability of each brand being chosen at each of its price points. Brands P1 and P2 are two premium brands. D1 and D2 are two discounted brands.

Figure 2



Each brand was shown at customized price points, to reflect its realistic price range. The data suggest an interaction between brand and price. The premium brands' shares do not appear to be as elastic as the discount brands. A log-log regression of share of choice on price reflects the following elasticities, and confirms that observation:

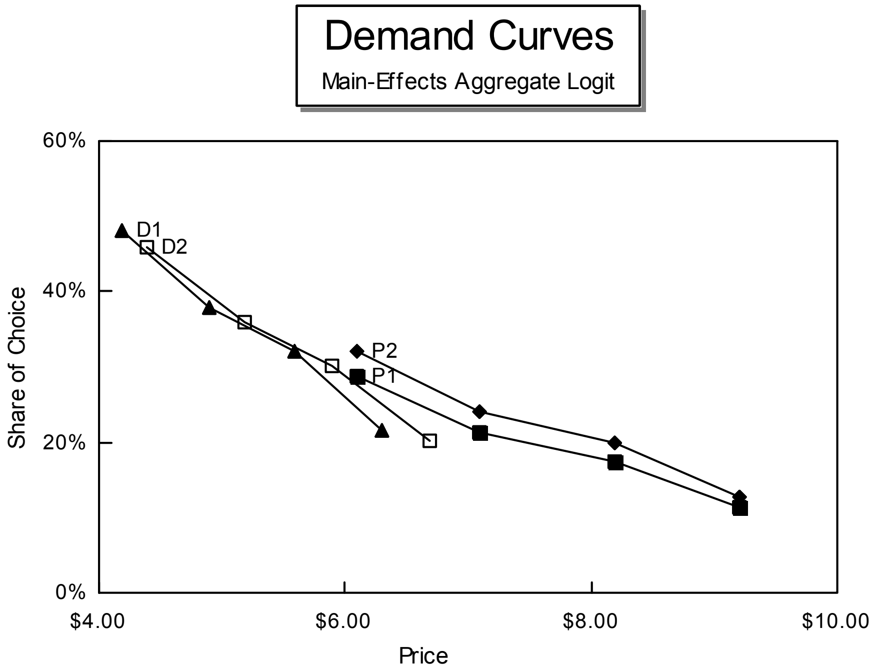
Table 5
Computed Elasticities from Two-Way CBC Counts

	Elasticity
P1	-2.15
P2	-1.89
D1	-2.72
D2	-2.65

We can also generate demand curves based on the same data using conjoint simulators. To do so, we conduct sensitivity analysis. For the example above, four products are entered in the market simulator. To determine the demand curve for a brand, we compute its share of choice at each price point, holding the other brands constant at their average prices.

Under main-effects only simulators, a single set of price utilities is calculated to reflect the average impact of price on choice, everything else (including brand) held constant. Computing demand curves for our example (which appears to exhibit an interaction between brand and price) using aggregate logit would be improper:

Figure 3



The interaction effect is lost. Table 6 displays the elasticities calculated from the aggregate logit main-effects only model next to those from Counts.

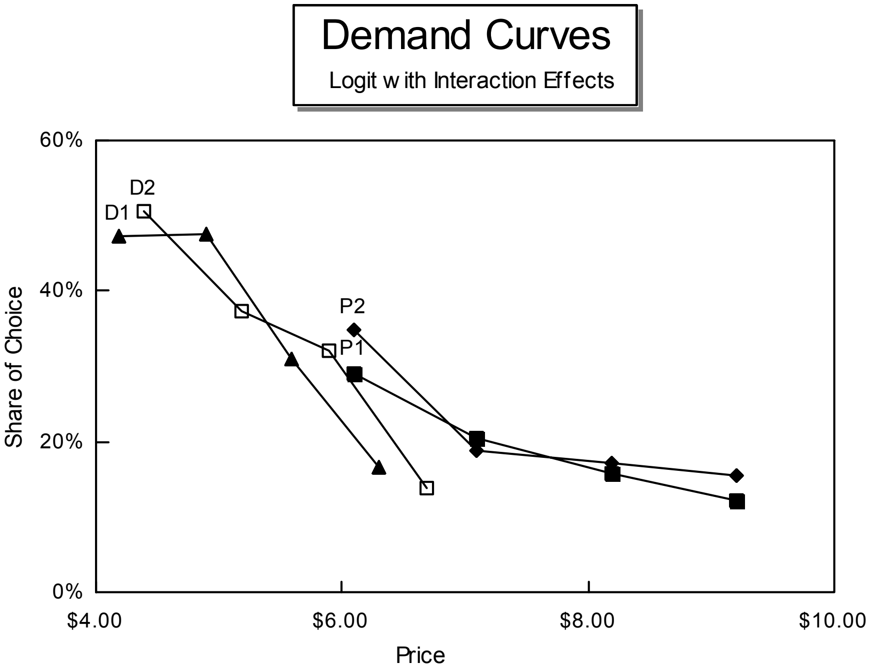
Table 6
Elasticities

	Two-Way Counts	Aggregate Logit ME
P1	-2.15	-2.16
P2	-1.89	-2.11
D1	-2.72	-1.86
D2	-2.65	-1.88

The elasticities are nearly constant under main-effects aggregate logit–the only discrepancies due to the average difference in height of the demand curves. (Given parallel demand curves, the higher share brands have lower elasticities, because a loss of each share point reflects a smaller percentage of the base share than for smaller share brands.)

Adding interaction terms between brand and price to the aggregate level model results in a much better fit to the underlying Counts data:

Figure 4



And the elasticities closely match the counts data:

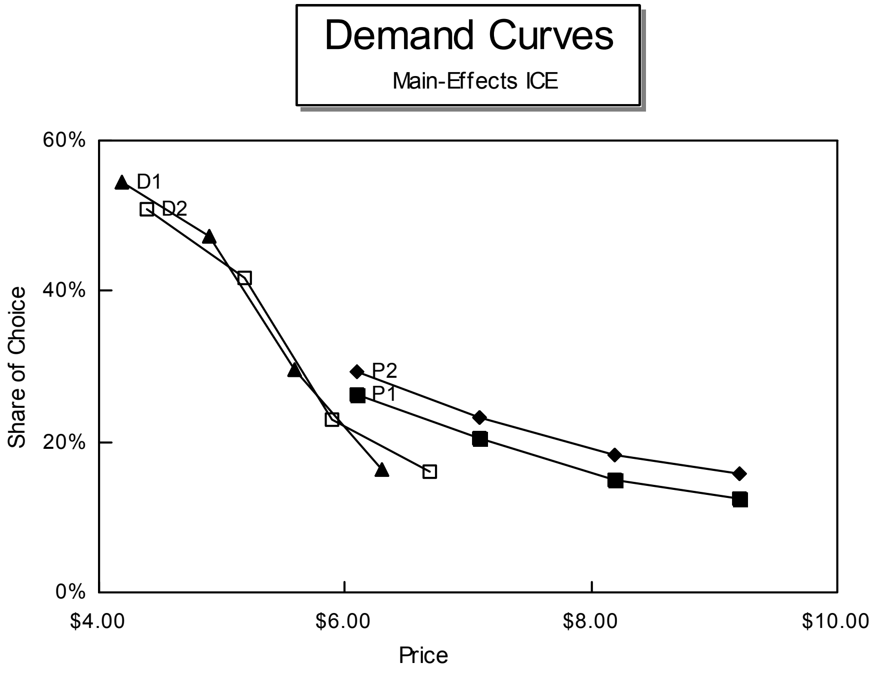
Table 7
Elasticities

	Two-Way Counts	Aggregate Logit: Main-Effects Only	Aggregate Logit with Interactions
P1	-2.15	-2.16	-2.07
P2	-1.89	-2.11	-1.85
D1	-2.72	-1.86	-2.57
D2	-2.65	-1.88	-2.85

We've seen that aggregate logit requires interaction terms to adequately model the relationship between brand and price. The main-effects only model failed to recognize the difference in elasticity between brands under aggregate logit, but how will the same model perform if we recognize heterogeneity?

Demand curves for the four brands under ICE (main-effects only) are as follows:

Figure 5



Even though we used main-effects only (did not include an interaction term for brand x price) the demand curves reflect that the premium brands are less price sensitive than the discount brands. Note that the price curves developed under main-effects ICE are generally smoother than the counts data and the logit with interactions model. In our opinion, ICE seems to cut through a lot of the noise (we don't have a particularly large data set) and do a fairly good job in reflecting the underlying relationships suggested by the Counts data. Table 8 reflects elasticities for all the models presented thus far.

Table 8
Elasticities

	Two-Way Counts	Aggregate Logit: Main-Effects Only	Aggregate Logit with Interactions	ICE: Main-Effects Only
P1	-2.15	-2.16	-2.07	-1.86
P2	-1.89	-2.11	-1.85	-1.53
D1	-2.72	-1.86	-2.57	-2.97
D2	-2.65	-1.88	-2.85	-2.86

How is it that ICE can reflect differential price elasticities using only a main-effects model? It is because the same respondents who strongly prefer the premium brands are also less price sensitive. Respondents who prefer (or are willing to settle for) the discount brands are generally more price sensitive. Those individuals who choose the premium brands in simulations will be less likely to change to the discount brands due to price movements, and vice-versa. It is important to recognize that *any* main-effects conjoint model that captures heterogeneity (traditional conjoint, ACA) can detect interactions when conducting sensitivity simulations.

We are not necessarily advocating modeling demand curves with main-effect disaggregate models rather than counts tables or aggregate models that directly model the interaction between brand and price, but pointing out that such models may often do an adequate job of reflecting more complex relationships without the additional terms added.

SUMMARY AND CONCLUDING REMARKS

Our data show that, under favorable conditions, CBC can accurately predict market shares for packaged goods at the grocery store. This extremely encouraging result calls for replication and verification. Our data also demonstrate that recognizing heterogeneity can improve results.

Aggregate-level logit has been faulted for its IIA properties. Specifically, it can fail when products with differing degrees of similarity are included in simulations. Corrections for product similarity such as Huber, Orme and Miller's RFC model (Huber, Orme and Miller, 1999) can improve aggregate level logit simulations, but it is best to begin with an underlying model that is less susceptible to the "Red Bus/Blue Bus" problem.

Many failings of IIA can be avoided by adding complex terms to aggregate logit models (i.e. interaction terms, cross-elasticities, availability terms). These models can become very complex and risk becoming over-fitted (too many terms relative to observations), fitting a good deal of noise along with true effects. It puts a heavy burden on the analyst to choose the right combination of complex terms to maximize predictive validity for aggregate logit models. But before the advent of disaggregate choice modeling techniques, more complex specifications were often needed to model the marketplace adequately.

Using LC and especially ICE can reflect complex effects (differential substitution, cross-effects and interactions) and achieve very accurate predictions using parsimonious main-effects models if such effects can be largely accounted for by differences in preference among underlying segments or individuals. Our research adds to a growing body of evidence that suggests this is often the case.

References

Brice, Roger (1997), "Conjoint Analysis A Review of Conjoint Paradigms and Discussion of the Outstanding Design Issues," *Marketing and Research Today*, November, 260-66.

Clancy, Kevin J. and Robert S. Shulman (1994), Marketing Myths that are Killing Business, New York: McGraw Hill, Inc., 201.

Huber, Joel, Bryan K. Orme and Richard Miller (1999), "Dealing with Product Similarity in Conjoint Simulations," *Sawtooth Software Conference Proceedings*.

Johnson, Richard M. (1988), "Comparison of Conjoint Choice Simulators—Comment," *Sawtooth Software Conference Proceedings*, 105-108.

Johnson, Richard M. (1997a), "Individual Utilities from Choice Data: A New Method," *Sawtooth Software Conference Proceedings*, 191-208.

Johnson, Richard M. (1997b), "Getting the Most from CBC–Part 2," *Sawtooth Solutions*, Spring, 2-3.

Natter, Martin, Markus Feurstein and Leonhard Kehl (1998), "External Validity of Segmentation Based CBC-Analysis," Marketing Science Conference, Fountainbleu, France.

USING SCANNER PANEL DATA TO VALIDATE CHOICE MODEL ESTIMATES

Jay L. Weiner
The NPD Group, Inc.

ABSTRACT

The purpose of this paper is to compare three forms of multinomial logit model estimation. The models include main-effects only, main-effects with interactions and brand-specific parameter estimation.

INTRODUCTION

Frequently, we are asked to simulate market share changes based on potential strategic marketing decisions. Discrete choice analysis offers the ability to estimate market share and simulate changes in market share based on changes in price, or other product attributes. The primary purpose of this study was to assess whether the client could raise its prices in the market without adversely affecting its market share. The client selected price points representing a range of ten percent around the current average price charged. The resulting model permitted the simulation of market share if the client raised its prices.

The scope of this paper is to examine the effect of various forms of analysis, main-effects only (choice based conjoint), choice based conjoint with interactions and brand-specific price effects with brand-specific price cross effects on the accuracy of predicting market shares. In addition, this paper will discuss the effect of using a constant sum (next 10 purchases) versus the more traditional choose one or “next” purchase. Wurst (1997), hypothesized that using the next 10 purchase occasions may be better suited to certain purchase situations. The most likely application of this should occur in product categories where consumers make frequent or multiple purchases. The best application would be in product categories where consumers tend to purchase a variety of brands, such as breakfast cereal.

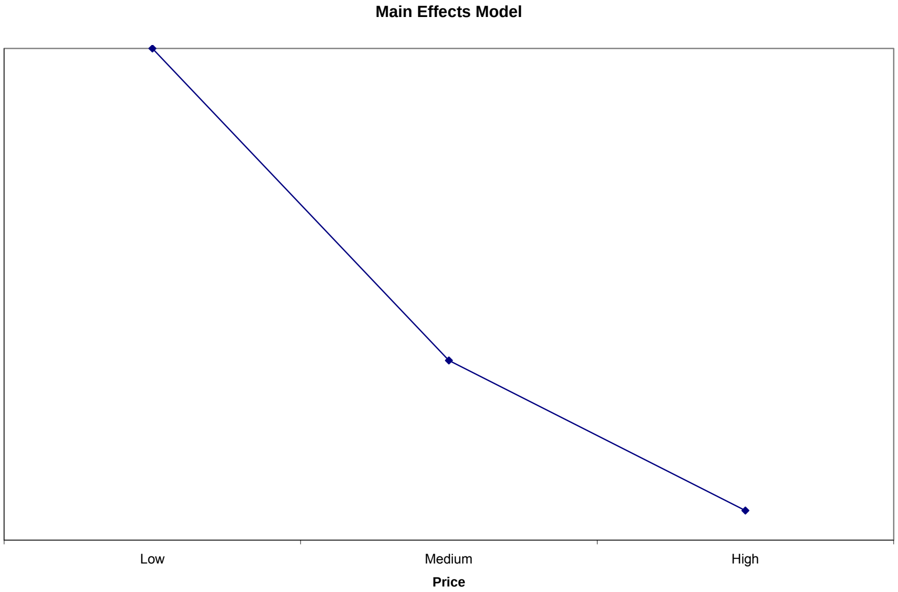
The product used in this study is a fast-moving consumer package good where consumers tend to exhibit strong brand loyalty. The average purchase cycle is 46 days (assumed interval). This means that on average, consumers make 8 purchases per year in this category. It seems logical that respondents can think about their next ten purchases in this product category. The market share estimates from the model are compared to the share reports from the Information Resources, Inc. (IRI) Infoscan Reviews Database.

METHODOLOGY

Respondents were shown a series of choice sets and asked to indicate how many of their next ten purchases would be one of eight brands (no single choice data were collected). To model the single choice, the brand with the largest proportion of each respondent’s choice was used as the first choice. In the event of a tie, the brand chosen was selected at random. The data were analyzed using Multinomial Logit (SAS PHREG). Three different models were estimated for both the first choice and next ten purchases (six models in total). The first method of analysis was a main-effects only (Choice Based Conjoint). Second, the data were analyzed using a main-effects plus brand price interactions model. Finally a model was estimated for brand-specific price effects and brand/price cross effects. The key advantage of this final approach is to allow the model to deal with the Independence of Irrelevant Alternatives (IIA) problem.

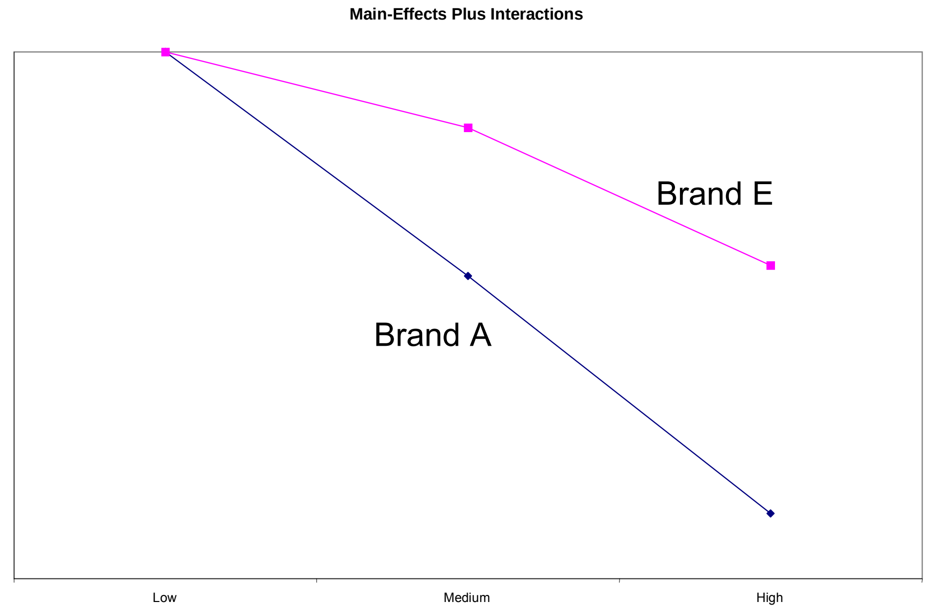
MAIN-EFFECTS ONLY APPROACH

For true choice based conjoint, each brand in the study would appear with the exact same price points. For example, if there were three levels of price (6.95, 7.95 and 8.95), all brands would be presented at these same levels. It is possible to allow some products to have different price points by labeling. For example, the price points could be considered low, medium and high. For premium priced products these price points might be 7.95, 8.95 and 9.95 while price leaders might have price points 5.95, 6.95 and 7.95. It is possible to estimate parameters for each brand intercept and each level of price. The estimates for each level of price will be identical whether all brands were allowed to have the same level of price, or if labeling was used. In this study, each brand had three unique price points. For the purposes of estimating a main-effects only model, the price parameters were estimated as 3 levels (low/medium and high) and not using a vector model. The price curve for this study is shown below.

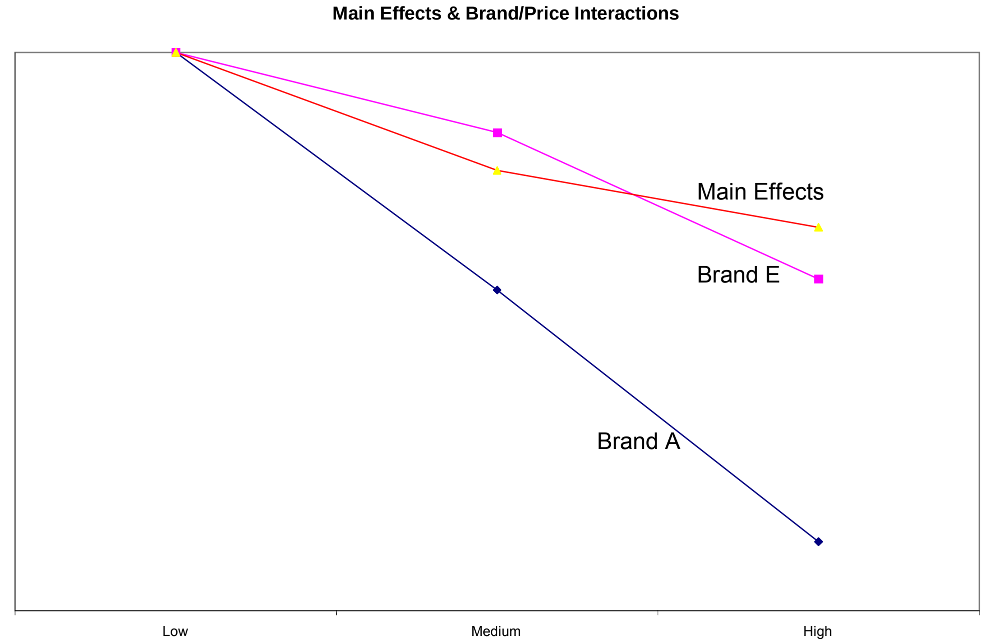


INCLUDING INTERACTIONS

The primary drawback of the main-effects only approach is that it assumes that all brands have the same price sensitivity curves. Johnson & Olberts (1996) propose to include the modeling of interactions in choice models to estimate the effects of price on specific brands. By including interactions of the model, we can estimate the effects of the low price for each brand in the study. This allows us to determine which brands' buyers are more sensitive to changes in price and to allow the model to estimate the effects of labeling. The price curves for two brands (the market leader BRAND A, and a premium priced product BRAND E) are shown below.



It is clear from this chart, that the brands do indeed have different price curves. Brand E buyers are far less sensitive to changes in price than Brand A buyers. This should translate into higher margins for the manufacturers of Brand E. To fully understand the benefit of including interactions in the model, the following chart shows the main-effects for price and the interactions for Brands A & E.



Brand-Specific Price Effects

The price curves from this model are identical to the main-effects plus interactions model. The main drawback of the two previous approaches to estimating choice probabilities comes during market simulations. In both the main-effects-only and the interactions model, any loss in market share is proportioned to the competitive brands based on each brand's current market share. In reality, brands tend to pull share from other brands in the consumers evoked set. It is more likely that each brand will lose to those brands that consumers perceive to be more similar to that brand, independent of market share. By estimating brand-specific price effects, it becomes possible to include brand/price cross effects. This permits the estimation of the effects of brand A's price on the share of brand B. This means that the model can specifically account for share changes in Brand B when Brand A raises or lowers its price. These effects are often considerably different than allocating share based on the proportion of market share of other brands. This benefit comes at the cost of having to estimate considerably more parameters. This means that more choice scenarios are needed.

Predicting Current Market Shares

To predict current market share, the current average price charged across markets (as reported by the client) was used. The IRI purchase panel data projects actual market share from store sales for a twelve month period that includes the time frame in which these data were collected. The IRI data also includes all sales during this period that were offered on some sort of promotion; either discounted on the shelf or purchased using a coupon. The study did not attempt to model the effects of sales promotion.

The Results

	IRI Reported Share	Main Effects Model First Choice	Main Effects Model "Next 10" Purchases	Main Effects Model with Interactions First Choice	Main Effects Model with Interactions "Next 10" Purchases	Brand Specific Price Effects Model First Choice	Brand Specific Price Effects Model "Next 10" Purchases
BRAND A	15.0%	13.7%	11.5%	18.6%	18.7%	17.1%	17.5%
BRAND B	4.2%	10.4%	6.5%	11.1%	8.1%	11.2%	7.5%
BRAND C	1.5%	7.8%	5.1%	9.5%	7.1%	8.7%	5.7%
BRAND D	7.7%	6.5%	4.8%	6.1%	5.6%	9.8%	9.2%
BRAND E	8.1%	2.7%	2.4%	5.0%	5.1%	7.5%	7.5%
BRAND F	9.9%	5.4%	4.2%	5.6%	5.2%	5.3%	5.8%
Store BRAND	14.7%	13.4%	11.8%	14.9%	15.1%	15.3%	17.7%
Some Other	38.9%	40.0%	53.7%	29.3%	35.1%	25.1%	29.1%

Deviations From IRI Reported Share

	Main Effects Model First Choice	Main Effects Model "Next 10" Purchases	Main Effects Model with Interactions First Choice	Main Effects Model with Interactions "Next 10" Purchases	Brand Specific Price Effects Model First Choice	Brand Specific Price Effects Model "Next 10" Purchases
BRAND A	-1.3%	-3.5%	3.6%	3.7%	2.2%	2.5%
BRAND B	6.2%	2.3%	6.9%	4.0%	7.0%	3.4%
BRAND C	6.2%	3.6%	8.0%	5.6%	7.1%	4.2%
BRAND D	-1.2%	-2.9%	-1.6%	-2.1%	2.1%	1.5%
BRAND E	-5.3%	-5.7%	-3.1%	-3.0%	-0.6%	-0.6%
BRAND F	-4.6%	-5.8%	-4.3%	-4.7%	-4.6%	-4.1%
Store BRAND	-1.3%	-2.9%	0.1%	0.4%	0.6%	3.0%
Some Other	1.1%	14.8%	-9.6%	-3.8%	-13.8%	-9.8%
Average Error	3.4%	5.2%	4.6%	3.4%	4.8%	3.6%

Discussion

With the exception of the main-effects only model, asking respondents to indicate their next 10 purchases improves the accuracy of the model. Greater errors seem to be associated with smaller share brands. In this study, the “first choice” main-effects only model predicted market share quite well. Adding the “next 10 purchases” increases the degrees of freedom and improves the market share prediction for main-effects with interactions and the brand specific price effects model.

The inclusion of “Some Other Brand” in the study makes the respondent task seem more reasonable, but also contributes the greatest errors when predicting market share. To determine if the inclusion of “Some Other Brand” detracts from the model, the analysis was replicated excluding this brand from the model. Predicted shares were re-scaled to sum to the shares of the models brands (about 62% of the market).

Results Excluding “Some Other Brand”

	Main Effects Model First Choice	Main Effects Model "Next 10" Purchases	Main Effects Model with Interactions First Choice	Main Effects Model with Interactions "Next 10" Purchases	Brand Specific Price Effects Model First Choice	Brand Specific Price Effects Model "Next 10" Purchases
BRAND A	1.2%	4.2%	1.1%	1.7%	1.1%	0.2%
BRAND B	6.5%	4.3%	5.6%	3.6%	5.0%	2.3%
BRAND C	7.8%	7.1%	6.6%	5.4%	5.5%	3.3%
BRAND D	1.0%	1.5%	2.5%	2.3%	0.4%	0.3%
BRAND E	5.2%	5.0%	4.0%	3.3%	1.9%	1.5%
BRAND F	4.3%	4.5%	5.1%	5.0%	5.6%	5.0%
Store BRAND	5.0%	4.6%	1.7%	0.0%	2.3%	0.4%
Average Error	4.4%	4.4%	3.8%	3.1%	3.1%	1.8%

SUMMARY

The main-effects only model seems to be quite robust and efficient for fast-moving consumer package goods. It is clear, however that there are different price curves for each brand in the study. To gain key insight into these effects, estimating interactions works well. Finally, if it is important to simulate where lost share goes (IIA related issues), the brand specific model is required. In this case, the additional degrees of freedom gained from using "next 10" purchases enhance the predictability of the model. Including "Some Other Brand" may make the respondent task seem more reasonable, but the results from this study indicate that market share predictions are significantly improved by not modeling this option.

REFERENCES

Johnson, Richard M. and Kathleen A. Olberts (1991), "Using Conjoint Analysis in Pricing Studies: Is One Price Variable Enough?" Advanced Research Techniques Forum Proceedings, American Marketing Association.

Wurst, John C., Brian Griner and Warren F. Kuhfeld, (1997), "Integrating Frequency of Purchase into a Discrete Choice Framework to Enhance Estimation of Market Shares," Advanced Research Techniques Forum Proceedings, Chicago, American Marketing Association.

SHOULD CHOICE RESEARCHERS ALWAYS USE 'PICK-ONE' RESPONDENT TASKS?

Jon Pinnell
MarketVision Research, Inc.

ABSTRACT

When conducting discrete choice experiments, researchers frequently present respondents with sets of configured products (concepts) and ask respondents to pick the one they most prefer or are most likely to purchase. The prevailing wisdom is that this replicates what consumers do in the marketplace. This paper reports the findings from a controlled experiment which compares alternative preference elicitation methods which seek to increase the information content of the response as well as make the tasks more natural. We find that the first choice tasks do quite well at conveying information, and seem superior to the metric methods tested. We continue to be intrigued by rank order (or second choices) – which provide a number of positive findings in this study.

BACKGROUND

Two converging areas of research have caused some researchers to question the current standard of 'pick one' respondent tasks in preference modeling – specifically discrete choice modeling.

First is the desire to use a task which allows respondents to provide more information than a simple indication of the most preferred product. This might include rankings or a more precise (metric) measurement as opposed to the only ordinal indication of the most preferred concept.

Second is allowing for variety seeking behavior or occasion heterogeneity. For example, a consumer might generally drink a particular brand of beer, but have a different brand or taste preference when entertaining. Other common examples include physicians and the choice of a medication to prescribe, which might differ by patient; or computer professionals who have differing preferences based on the function for which the machine is purchased, for example word processing versus graphical rendering.

Several researchers have suggested that instead of asking respondents to pick their most preferred product from a set, we should allow them to indicate a relative preference for each presented product or even a likelihood of purchase.

EMPIRICAL DATA

The empirical findings draw exclusively from an experimental research project conducted from a MarketVision Research methods development budget. In total 400 consumers were interviewed using MarketVision's 'Gateway' Consumer Research Center inside Universal Studios Florida. Four variations of choice modeling were tested, each varying the response that consumers provided. The four methods included two non-metric methods and two metric methods. The four methods tested were:

Non-metric

First Choice

Rank Order

Metric

Allocation

Scaling

Each respondent randomly received two of the four methods in random order. The tasks relied on randomly constructed choice designs and computer-aided self-interviewing. In total 7 attributes were studied, with a total of 33 attribute levels. Each task was constructed of four concepts.

The topic of the research was hotels. Hotels were a logical choice for three reasons:

1. Hotel/Lodging is a category for which people frequently have differing criteria for selection depending upon the usage occasion (pass through traveling versus destination stays).
2. The respondents at Universal are currently in the market for hotels, as the majority of guests travel from out of market.
3. MarketVision Research does a great deal of research in the broad area of travel, tourism, entertainment, and leisure.

Each of these four alternative methods are detailed below:

First Choice

This is the classic choice based method. Respondents are presented with several alternatives (four in our study) and are asked which one they most prefer or are most likely to purchase. As has been detailed elsewhere, these question types are relatively inefficient. First, choices contain relatively little information. We know which one concept the respondent prefers but we learn nothing about by how much they prefer it. Second, choices, despite being processed quickly by respondents, seem intuitively expensive. Respondents are forced to read four concept descriptions and only provide their preferred concept, potentially losing much of the information they processed. These problems aside, first choice is still the most common approach to choice modeling among the four that we tested.

Full Rank Order

Some researchers, trying to increase the amount of information respondents report for each choice tasks have asked respondents to give either a second choice (in addition to their first choice) or even a full ranking of the concepts presented. Intuitively, this is a pleasing solution. If we believe that much of the time of a choice task involves the respondent reading and processing the concepts, it should be a relatively simple task to indicate a second choice. At the same time, this can greatly increase the amount of information derived from each choice task, based on the number of pairwise inequalities created.

	Pairwise inequalities from		
	First choice (a) from 4	First and Second choice (a, c) from 4	
	a > b	a > b	
	a > c	a > c	
	a > d	a > d	
		c > b	
		c > d	
Pairwise inequalities	3	5	(+67%)

Unfortunately, investigations into rank order tasks, and second choices especially, have shown that they produce utilities different from first choice utilities (Pinnell and Huber, 1996 and Johnson and Orme, 1996). It isn't entirely clear, though, that just because rank tasks produce different utilities, that they produce inferior utilities. It might be the case that when respondents know that they will be asked additional questions, they consider their first choice more carefully. In fact, the author has advocated asking second choices but only using first choices to develop the utilities.

Allocation

One of the methods that we have included under the metric heading we will refer to as Allocation. This is analogous to the traditional constant sum question. Under this method, respondents were given a certain number of points to allocate across the four alternatives. No constraints were placed on their allocation except that the total points allocation must equal a fixed number, ten in this study. Though not seemingly difficult, among our sample a number of respondents failed to allocate 10 points. The choice screen even included a counter with a points used and a points available field. We gave each respondent a second chance to reallocate their points, but then skipped them out of the task if they still failed to allocate ten points properly.

It is obvious in pre-testing studies of this type that a number of respondents have difficulty with this question type. We have seen this to be the case even with what would be considered well-educated respondents, like physicians. We sometimes wonder if the respondent is spending more time processing the choice concepts presented or trying to add to ten. That said, it should be a relatively easy task to identify the respondents who had difficulty and remove them from the analysis. The metric nature of the data among the majority of the respondents who understood the task should outweigh the small proportion that is discarded.

It also might be the case that low-tech alternatives provide a reliable work-around. For example, Marder provides stickers to respondents and asks them to allocate their stickers to concepts.

Scaling

The other metric method tested relied more heavily on a choice framework. In the Scale method, respondents were presented with the four concept alternatives and asked to indicate their most preferred concept. Next, respondents were asked to indicate their least preferred concept. We then assigned a value of 100 to the most preferred concept and a value of 0 to the least preferred concept. Respondents were asked to scale the two interior concepts within the 0 and 100 range provided by the best and worst. Intuitively, this is a pleasing approach because we explicitly determine the most preferred concept, and by how much it is preferred. At the same time, some might argue that the scaling of this approach is too arbitrary.

EVALUATION CRITERIA

We are always somewhat troubled by the yardstick used to measure competing methods. Instead of being able to answer definitively that one method is preferred, we search for patterns of congruence, bias or cost. To that end, we report our findings using a host of criteria. Each is briefly introduced below:

Cost—Time Required Per Task

From the respondent's perspective, as well as the researcher's, the time it takes to evaluate a particular choice set and provide the required response(s) is one of the most noticeable differences between methods considered. We evaluate time in terms of median response time in seconds for each task.

Model Fit

Model fit represents how well the researchers' tools are able to explain the respondents' answers. We suggest that a method whose answers are more easily explained (has more variance explained) is superior to one whose answers are less easily explained.

Utility Congruence

We investigate how similar the part worth utilities derived from metric methods are to traditional first choice utilities derived from logit. To conduct this analysis, we used both cross task comparisons as well as inferring a 'first choice' from the two ratings methods (based on the concept receiving the highest rating).

Information Content

In addition to exploring if the utilities estimated from various methods converge (utility congruence), we also explore the amount of information derived from the four methods. The premise behind this criterion is similar in nature to the scale parameter in logistic regression. We explore the information content of various utilities through an inferred scale parameter as well as the average of the absolute value of t ratios of utilities that we develop.

Reliability

We evaluate two forms of reliability: the reliability of a method over tasks as well as the reliability of a method over random sub-samples of respondents. The ability of a method to produce reliable utility estimates over each variation makes it a relatively more appealing method, ceteris paribus.

Cross-Task Predictive Validity

In addition to the ability to produce reliable and consistent utility estimates, a method that corroborates or replicates other choices should be viewed as superior.

Predictive Validity of Hold-Outs

Finally, we included a set of ratings-based hold-out tasks by which to provide another measure of the predictive validity of each method. It might stand to reason that the metric methods will do better at predicting metric hold-outs.

FINDINGS: COST—TIME REQUIRED

It should be clear, ceteris paribus, that a respondent task with a lower cost is preferred to one with a higher cost. We consider cost to be the time it takes a respondent to answer each task.

The median times per task (in seconds) are reported below for each of the four alternative treatments. We had anticipated the four methods to require different times. Based on our prior expectations, we included a different number of tasks for each of the methods – hoping to equalize the time required. The number of tasks included in the questionnaire for each method is shown below.

Number of Tasks Included in Design

First Choice	10
Full Rank Order	7
Allocation	7
Scaling	5

As it turns out, our prior estimates of time were overly optimistic for both of the metric (Allocation and Scaling) tasks. The actual median times per task for each method are shown in the following table. The underscore represents how many tasks could have been asked for each method using roughly the same amount of time that 10 first choice tasks took.

Median Time Per Task
(in seconds)

	First Choice		**Rank Order**		**Allocation**		**Scaling**	
Task Number	Task Time	Cum. Time	Task Time	Cum. Time	Task Time	Cum. Time	Task Time	Cum. Time
1	14	14	25	25	37	37	49	49
2	12	26	16	41	20	57	37	86
3	11	37	16	57	16	73	32	118
4	10	47	15	72	16	89	30	148
5	10	57	14	86	15	104	27	175
6	10	67	12	98	13	117		
7	10	77	12	110	12	129		
8	10	87						
9	10	97						
10	9	106						

We were somewhat surprised by the time required for the metric methods – especially the Scaling approach. In the time it takes respondents to provide ten first choices, they can respond to six Allocation and only three Scaling questions. It has been the conventional wisdom that techniques which allow a researcher to collect more information (provided unbiased) from a single reading of a choice task would be extremely beneficial.

Since many of our comparisons are going to compare alternative methods to first choice – the implicit gold standard, we have erred on the side of including more, rather than fewer, tasks in the time equalized number of tasks for the competing methods. The number of tasks retained for time equalized comparisons are shown below.

Number of Time Equalized Tasks

First Choice	10
Full Rank Order	7
Allocation	6
Scaling	3

Most of the data and findings presented below rely on time equalized tasks. That is, when we compare part worth utilities from First Choice to Scaling, for example, we are using utilities derived from only the first three Scaling tasks for those comparisons.

Unless otherwise noted, the data reported are time equalized.

FINDINGS: MODEL FIT

Our second criterion relates to how well respondents' answers can be explained by the profiles themselves. We evaluate this for the metric methods with a traditional R^2 value. Keep in mind that we are not too interested in the R^2 values themselves, rather we care about each method's relative explanatory power.

Our focus initially is between the two metric methods. To conduct these analyses, we centered the dependent variable separately for each respondent. The resulting R^2 values derived from linear regression for each of the two (time equalized) metric methods are shown below. We also show an index arbitrarily setting the Scale finding to a value of 100.

Model Fit – Metric Methods

	R^2	Index
Scale	0.18	100
Allocation	0.08	44

With both of these methods, respondents provided a response to all four of the concepts presented in each task. In the scale method the most preferred concept was required to have a value of 100 and the least preferred was required to have a value of 0. Only the ratings of the interior two alternatives were allowed to vary. Conversely, in the allocation method respondents were constrained only by the requirement that the total sum of the ratings be 10.

One can imagine a case in which a particular respondent task could cause all respondents to simplify their information processing so much that they only pay attention to one attribute. This (or similar) condition could produce a higher model fit, but actually contain less information. For now, we assume that this is not that case (which we explore in depth in the next section).

Given that, it would appear that respondents had some difficulty with the Allocation task relative to the Scaling task. Even though assigning values of 0 and 100 might appear arbitrary, we infer that it helped structure the task for the respondents.

FINDINGS: UTILITY CONGRUENCE

In the previous section, we assumed that the four methods would produce similar utilities. Below we investigate whether that was a safe assumption or not. From each of the four methods, we develop independent estimates of part worth utilities. We then calculate a simple correlation between the utility estimates. It is not clear that a low correlation indicates a method is less good, only different. That said, the utilities for the methods show a high level of consistency.

Similarity of Part Worth Utilities

	First	**Rank**	**Scale**	**Allocation**
First	---	.945	.890	.920
Rank	.945	---	.889	.917
Scale	.890	.889	---	.904
Allocation	.920	.917	.904	---
Average	**.918**	**.917**	**.894**	**.914**

Overall, we see the correlations are generally high, with the lowest values coming from the Scale method. Recall that this table represents time equalized utilities, and the Scale method was the most expensive method, so only three tasks are being used in these calculations. It has been shown that respondents become more reliable in later tasks (relative to earlier tasks) so comparisons of the scale questions are likely being hurt by this time equalization. To illustrate this point, we have reproduced part of the previous table based on the utilities using all the tasks (not just time equalized). Here we see the previous differences, which were relatively small to begin with, decrease.

Similarity of Part Worth Utilities
(Not based on time equalized tasks)

	First
First	---
Rank	.945
Scale	.928
Allocation	.922

We take these findings to suggest that the biggest difference from comparing the methods is not necessarily between the methods, but the time each method takes and the number of tasks asked in that time.

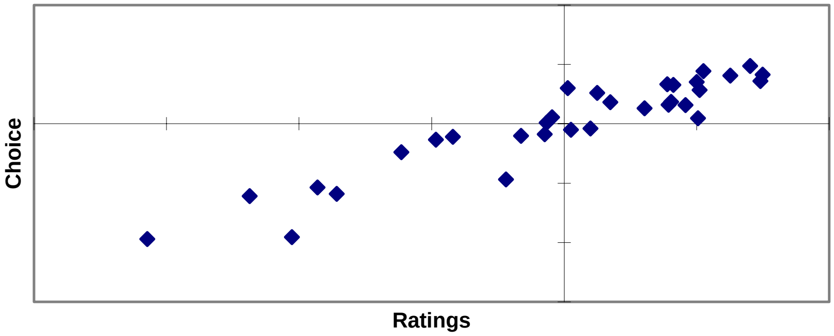
The previous section has been comparing the utilities from different methods. Implicit in this comparison are two types of differences. The most obvious difference is the fundamental difference between the methods. The second, and less obvious, difference is that part of the utilities developed are between respondent comparisons (while some are within respondent). Since one of our key questions is how much more information is contained in the utilities derived from the more involved methods, we should also look to make an entirely within respondent comparison.

Put another way, do the metric methods communicate different information (low correlation) or more precise information than first choices?

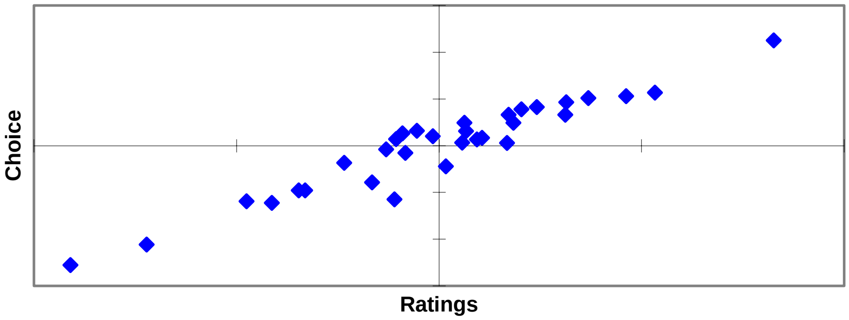
To investigate this, we inferred a first choice for each of the metric methods. We then compared the utilities from the inferred first choices (fit with a logit model) to the ratings utilities (as we had developed with linear regression).

These utility plots are shown below:

SCALE (TE) (Ratings Vs. Choice) (r = 0.936)



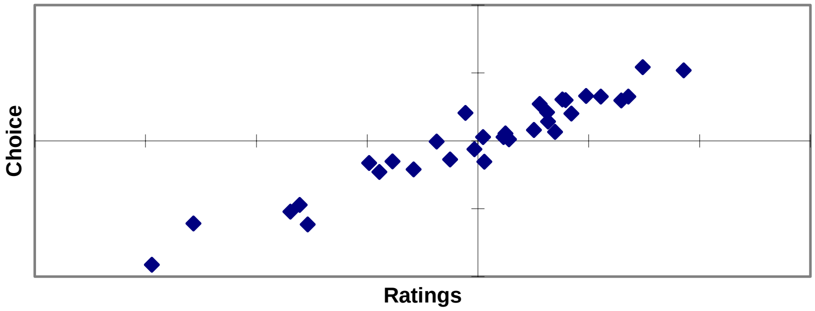
ALLOCATION (TE) (Ratings Vs. Choice) (r = 0.956)



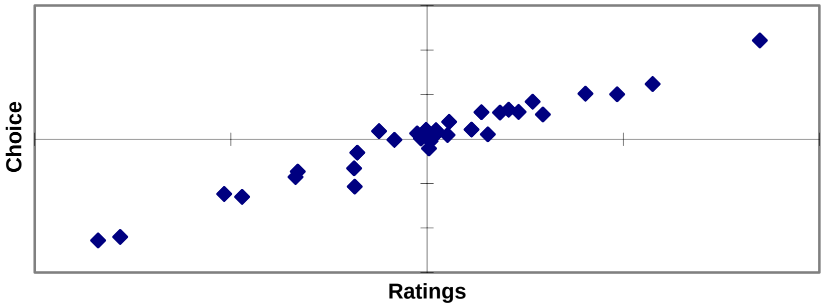
We find three things interesting from these charts.

- First, for both methods, the correlation between the inferred first choice utilities and metric ratings are high. And they are similarly high when comparing the methods to each other.
- Second, it would appear that the scale method appears noisier – but recall that since we are using time equalized tasks, this is based on only three first choices. If we look at non-time equalized tasks, the correlation between the utilities increases to 0.964 for five Scale tasks and 0.975 for seven Allocation tasks. These are right in line with the results of the correlation of the first choice utilities from the Rank tasks and the utilities developed using all of the ranking data (r = 0.966).

SCALE (Ratings Vs. Choice) (r = 0.964)



ALLOCATION (Ratings Vs. Choice) (r = 0.975)



- Third, the allocation tasks appear less noisy, but two things do stand out: a few leverage points and curvature. To illustrate the curvature, we calculated the slope between the choice utilities and the ratings utilities for each method, separately for positive rating utilities and negative rating utilities.

	Scale	**Allocation**
Slope for negative rating utilities	.05736	.53583
(std. err.)	(.0044)	(.0344)
Slope for positive rating utilities	.05739	.49019
(std. err.)	(.0040)	(.0249)

From this we see that there is no indication of curvature in the Scale tasks, while there is some evidence among the Allocation questions. The evidence is even stronger if the top utility is dropped.

The part worth utilities derived from the inferred first choice have a high level of congruence with the utilities derived using linear regression and the metric responses. Given that, we have gone back to compare first choice logit utilities from all four methods. These simple correlations are shown in the following table.

Utility Congruence
Based on Inferred First Choice

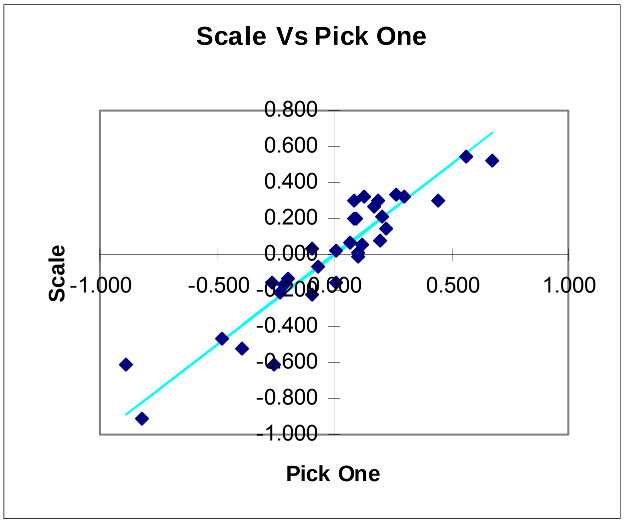
	First	**Rank**	**Scale**	**Allocation**
First	---	.954	.823	.929
Rank	.954	---	.870	.934
Scale	.823	.870	---	.851
Allocation	.929	.934	.851	---
Average	**.902**	**.919**	**.848**	**.905**

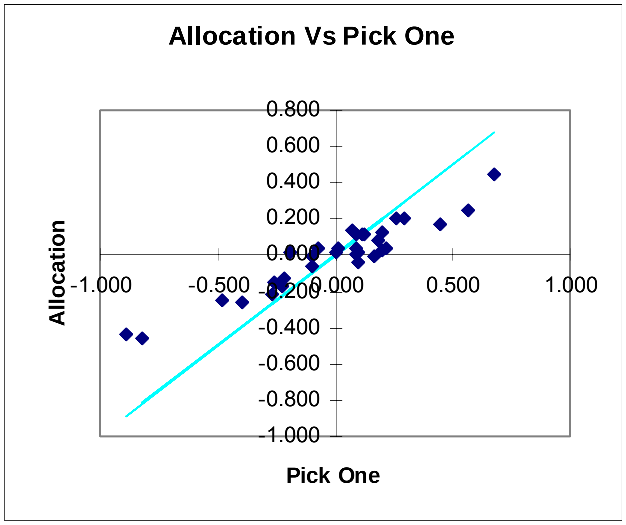
As we saw in the correlations between linear regression, scale utilities and first choice logit utilities, the scale method appears substantially inferior. Using all tasks (not time equalized), the utilities are far more similar.

Utility Congruence
Based on Inferred First Choice
(Not based on time equalized utilities)

	First	**Rank**	**Scale**	**Allocation**
First	---	.954	.924	.948
Rank	.954	---	.932	.941
Scale	.924	.932	---	.905
Allocation	.948	.941	.905	---
Average	**.942**	**.942**	**.920**	**.931**

Overall, we see that the four methods produce very similar results when the time equalization is removed. However, a single correlation can hide a multitude of sins. We see a different picture when we plot the inferred first choice utilities from each of the metric methods against the utilities from the first choice model itself. These are shown below.





The solid line represents y = x. Note how flat the utilities from the allocation tasks are relative to the scale tasks and the first choice tasks. To illustrate this point further, we conducted a linear regression (forcing the intercept to be zero) to test if the slopes are truly different. Those results are shown below.

Utility Congruence
Based on Inferred First Choice
Forcing the Origin
(Not based on time equalized utilities)

	Correlation With First	Slope (First = 1.00)
Rank	.954	0.872 (.049)
Scale	.925	0.938 (.069)
Allocation	.929	0.544 (.039)

Based on the high and similar correlations, we see that utilities appear to be representing the same preference structure. The difference in the slope for the Allocation task is particularly noteworthy. These slopes can be interpreted as the scale parameter is, suggesting the utilities from the Allocation tasks are noisier.

Another way to interpret the noise of part worth utilities is to examine the t-ratios of the parameters. The following table shows the average of the absolute value of the t-ratios from utilities derived from the inferred first choice for the four methods.

Avg. Abs. t ratio

First	4.040
Rank	3.004
Scale	2.137
Allocation	2.033

Recall that these are based on the time equalized tasks, so the first choice is based on 10 tasks, the ranks 7, the allocation 6 and the scale only 3 tasks. Still the three inferred first choices from the Scale tasks appear more precise than 6 inferred first choices from the Allocation tasks.

This look at average t-ratios of inferred first choices might be discarding a great deal of information that was collected in all but the first choice tasks. The following tables shows the average of the absolute value of the t ratios when all the information in the tasks is used to develop the utilities.

Avg. Abs. t ratio

	Inferred First Choice	**Full Info**
First	4.040	4.040
Rank	3.004	4.221
Scale	2.137	2.416
Allocation	2.033	2.394

Surprisingly, the two methods with metric level data show only marginal improvement when the other data is included in the utility calculation. The Full Rank order tasks, however, catapult to the most precise of these four methods using this criterion.

FINDINGS: TASK ORDER RELIABILITY

As mentioned above, it has been shown that respondents in choice tasks become more reliable on successive tasks. We wished to explore the task order reliability of the utilities developed from each of the four methods. We based this comparison on inferred first choice only.

Inferred First Choice

	Early to Middle	Middle to Late	Early to Late	Average
First	.629	.854	.698	0.727
Rank	.625	.696	.547	0.623
Scale	.752	.558	.596	0.635
Allocation	.734	.638	.584	0.652

It is not immediately clear what to make of this table in a cross method comparison, but one generalization might be that choice methods tend to increase in reliability over time, while the metric methods decrease in reliability over time. We offer a conjecture that respondents in the metric methods become more concerned with the task itself than with what the task is trying to measure (their preference). Overall, these data would suggest the First Choice method to be most reliable over tasks.

FINDINGS: SAMPLE REPLICATE RELIABILITY

Another form of reliability is the reliability of the derived utilities across independent sample replicates. For each of the four methods, we developed three random divisions of respondents and estimated utilities for each group independently. Three separate random divisions were used, and the results averaged. This was done for the inferred first choice as well as taking advantage of all of the information provided. The following table shows the average reliability for the inferred first choices.

Reliability Across Sample Replicates
of Inferred First Choice

	First	**Rank**	**Scale**	**Allocation**
Reliability	0.8664	0.8577	0.7236	0.6472

This table shows the choice methods producing more reliable results across sample replicates than the metric methods with the First Choice and Rank Order in a dead heat. The metric tasks appear to not do well in this comparison. Again, they are penalized by the time each takes as well a not taking full advantage of the information collected.

If we reproduce the table shown above, this time using all of the information from a method (still using time equalized tasks), we see only modest improvements in the reliability. This is shown in the following table.

Reliability Across Sample Replicates
Using Full Information

	First	**Rank**	**Scale**	**Allocation**
Reliability	0.8664	0.9205	0.7430	0.6633

The most striking difference is the reliability for the Ranking questions. The improvements in the metric methods are each around .02, confirming the previous finding of little additional information in the metric tasks.

FINDINGS: CROSS-TASK PREDICTIONS

In addition to the reliability of the utilities developed from the four methods, we are also interested in the validity of the utilities. In controlled experiments such as this, we often look to internal predictive validity to provide such a measure.

Since this paper has largely focused on seeing what incremental value the metric data has over just choice data, we can use the two metric methods to predict which concept out of four the respondents would choose when making their first choice. We can compare this to the actual choice that they made. In addition, we can evaluate how well the metric data does relative to using the utilities developed only from the inferred first choices.

By way of comparison, we can use the utilities developed from the first choice of the Rank section to provide a benchmark. Using those utilities, we can correctly predict 49.4% of the choices that respondents made in the first choice section (recall that this is a partly within, partly between respondent comparison). This benchmark is arbitrarily given an index value of 100. The findings from the inferred first choice and the full information for Scale and Allocation are shown below.

Cross Task Predictive Validity
Predicting First Choice Tasks

	% Hits	**Index**
Rank (first choice)	49.4	100
Rank (full information)	51.5	104
Allocation (inferred first choice)	50.2	102
Allocations (full information)	50.1	101
Scale (inferred first choice)	48.7	99
Scale (full information)	48.5	98

We are forced to conclude that neither metric method is superior to the other on this criterion. The full information utilities from the Rank tasks offer the best cross task predictions.

FINDINGS: PREDICTIVE VALIDITY OF METRIC HOLD-OUTS

Prior to any of the four experimental methods, respondents evaluated four metric hold-out tasks. These tasks resembled a full profile ratings based conjoint question and were fixed for all respondents. The tasks were repeated again at the end of the survey.

To analyze these holdouts, we will calculate a correlation between the predicted rating of the holdout (based on utilities) and the actual rating. As a benchmark, the correlation between the pre and post holdouts was 0.99.

As done above, we first calculated the predictive validity correlation using the utilities from the inferred first choice model. These are shown in the following table.

Correlation with Metric Hold-Out Ratings
Using Inferred First Choice

	First	Rank	Scale	Allocation
Using First Choice Utilities to Predict Ratings HO	0.6021	0.9952	0.7551	0.6047

Using this criterion, it would appear that the best first choice utilities come from when the respondents are asked to provide full rank orders. Again, using the inferred first choice alone, we might be sacrificing the information contained in the metric data of the Scale and Allocation methods. We calculated the same predictive validity correlation using the utilities using the full information available in the methods. These are shown in the following table.

Correlation with Metric Hold-Out Ratings
Using Full Information

	First	Rank	Scale	Allocation
Using Full Information Utilities to Predict Ratings HO	0.6021	0.9313	0.9173	0.9110

Notice that the correlation from the rank utilities decrease with the addition of the full rank data. However, in either instance (first choice or rank order), the rank order data provides the highest correlation.

Clearly, the metric data help the metric methods predict metric holdouts. They still under-perform the rank method, but outperform (in both instances) the inferred first choice only results.

CONCLUSIONS

The metric methods (especially Scaling) appear very expensive relative to first choice. For the time required, they offer little information above the first choice utilities. This appears to be true for both the cross task comparison as well as when we inferred the first choice from the metric tasks themselves.

Respondents process first choice methods quickly, which we knew, and first choices contain little information, which we also knew. Many of our comparisons measured the alternative methods against first choices. However, first choices seem to consistently provide reliable utility estimates. Their only shortcoming might appear to be predicting metric holdouts.

We continue to be intrigued by what appears to be the processing differences between First Choice and the Rank Order tasks. We know that second choices are flatter than first choices. We know that they increase the time it takes to make a first choice. At the same time, we wonder if people don't make better first choices if they know that they will also be asked to make second choices as well.

Recommendations

We are left with a conclusion that first choices are probably used as much as they are for a good reason. Some readers might be left with a concern over occasion heterogeneity. We would offer the following suggestion based on our experience.

First, we need to separate two forms of research. The first centers around product design, pricing or branding – the common uses for discrete choice modeling. The second is focused on optimizing a category portfolio. In the former, we are interested in understanding the preference for attributes and levels. In the later, we have a set of competing concepts and want to understand the combination that provides the strongest line-up. For portfolio line-up optimization research, it strikes us that an allocation task in a fully competitive environment might still be beneficial. In the standard product design research application, though, we are inclined to continue to recommend choice (non-metric) methods.

It has been suggested that prologues can be used with a standard pick-one tasks to introduce various scenarios. These have normally been suggested in order to keep the respondent's attention or to set a frame of reference. Normally these varying prologues are not delineated through the analysis phase.

In the past, the author has used prologues with success. However, we typically ask a set of first choice questions within each specific prologue. We then conduct the analysis separately for each 'prologued' occasion. We have concluded that the success of using prologues to delineate occasions is determined by how clearly the occasions can be enumerated, communicated, understood, and agreed upon by both the respondent and the researcher *a priori*. We continue to be more intuitively pleased by this approach than by pooling potentially several different occasions to develop one set of utilities.

References

Johnson, Rich and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint?" Presented at ART Forum; Beaver Creek.

Pinnell, Jon and Joel Huber (1995), "The Effectiveness of Second Choices in Experimental Choice Studies." Presented at INFORMS Marketing Science Conference; Gainesville, FL.

ASSESSING THE RELATIVE EFFICIENCY OF FIXED AND RANDOMIZED EXPERIMENTAL DESIGNS

Michael G. Mulhern, Ph.D.
Mulhern Consulting

INTRODUCTION AND BACKGROUND[1]

In recent years, considerable research attention has focused on the design of choice-based conjoint experiments. Bunch, Louviere and Anderson (1994) compared design strategies for symmetric main effects experiments with generic attributes and found a "shifting" or foldover strategy worked well. Anderson and Wiley (1992) and Lazari and Anderson (1994) catalog statistically efficient experimental design plans for main and cross-effect experiments. Kuhfeld, Tobias and Garratt (1994) discuss optimization of a wider variety of designs using computer search routines. Huber and Zwerina (1996) identify criteria to enhance design efficiency criteria both within and between choice sets. Johnson and Orme (1996) investigated how many choice sets a respondent can answer without fatigue or other biases and found the answer to be at least 20.

In reviewing this research, there is another design question whose resolution could benefit practitioners. When compared to catalog derived experimental designs (which generate a fixed number of profiles), what is the relative efficiency of randomized designs as the number of choice sets increases? Stated somewhat differently, what is the relative efficiency of fixed vs. random designs for a given number of choice sets?

Choice data are often collected via paper and pencil questionnaires so fixed rather than random experimental designs are required. Consequently, the purpose of this research effort is to assess the relative statistical efficiency of fixed and randomized designs as we increase the number of total choice sets for the randomized designs.

There is some evidence that the total number of choice sets required may differ for symmetric and asymmetric designs. A symmetric design is an experimental design where each attribute contains the same number of levels. An asymmetric design contains attributes with varying numbers of levels (Addelman, 1962). Huber and Zwerina (1996) point out that for many design specifications orthogonality and level balance can not be optimized simultaneously in asymmetric designs. As a result, whether an experiment is symmetric or asymmetric may impact the number of choice sets required.

[1] The author would like to thank Keith Chrzan for his valuable contribution to this paper.

MOTIVATION AND RESEARCH PLAN

The motivation for this paper came from a practical research problem faced by the author. When bidding on a study that required paper and pencil data collection, it was obvious that a randomized design was not viable. A fixed design was required but the question immediately arose - How many choice sets should be allocated across respondents?

In this context, a fixed experimental design is one in which the analyst predetermines the number of choice sets prior to fielding the study. Typically, there is a small number of versions (i.e. blocks or subsets of all choice sets) developed to reduce the risk of information overload and respondent fatigue. A randomized design creates the versions or blocks of choice sets based upon the respondent number and frequently generates many more choice sets to be shown across all respondents than a fixed design. They are typically slightly less efficient than fixed designs for estimating specific effects but are fairly robust in the estimation of all effects (Sawtooth Software, 1995).

Fewer total choice sets would allow for smaller samples and reduced study costs. At the same time, too few choice sets would result in an inefficient design and potential problems in model development. A tally of a number of practicing researchers yielded a wide variety of answers to the basic question of "How many choice sets are enough?"

Phase 1: Symmetric Choice Experiments: Randomized vs. Fixed Design

The research consists of two phases. In each phase, a constant number of unique choice sets were generated with a catalog experimental design to serve as a starting point. The number of simulated respondents was then varied to increase the total number of observations.

In the initial stage, the relative efficiency of fixed symmetric designs is compared to randomized designs for a fairly large choice experiment.

Bunch et al. (1994) describe a “shifting” strategy for creating highly efficient designs for fixed symmetric experiments. By relying on the law of large numbers, a randomized design with enough unique choice sets will become highly efficient. In the initial phase, the research question is how quickly (i.e. how many choice sets are required) does the randomized design become nearly as efficient (i.e. provide a relative efficiency approaching 1.0) as the fixed symmetric design.

Design efficiency measures the goodness of a design relative to hypothetical orthogonal designs. These measures focus on the statistical properties of the estimators; specifically, the variance and covariance of the parameter estimates. Efficiency measures the relative precision of the parameter estimates.

Several measures of design efficiency are available. Previous research has indicated that all provide similar results. D-efficiency is the most widely used among practitioners and is the measure of efficiency used in this research. Further information regarding the various measures of experimental design efficiency can be found in Kuhfeld (1997) and Kuhfeld, Tobias and Garratt (1994).

Phase 2: Asymmetric Choice Experiments: Randomized vs. Fixed Design

In the second phase, the relative efficiency of fixed asymmetric designs is compared to randomized designs for large and small asymmetric choice experiments. The question is how does the relative efficiency of randomized vs. fixed designs vary as the number of choice tasks increases for the randomized design.

Huber and Zwerina (1996) point out that level balance and orthogonality cannot usually be jointly optimized in an asymmetric choice experiment. In this context, the term balanced refers to level balance, not utility balance. Because imbalance increases the variance in the parameter estimates, an orthogonal fixed design with relatively few choice sets will probably not be balanced. However, in randomized designs involving many more choice sets, it is possible to satisfy both orthogonality and level balance. Thus, randomized designs with large numbers of choice sets are likely to be more efficient than smaller asymmetric fixed designs.

ANALYSIS PROCESS

The steps performed in the analysis were as follows:

1. Generate experimental designs and artificial respondents' random choice data for differing numbers of observations.
2. Create a logit command file and estimate multinomial logit parameters.
3. Calculate the determinants of the information matrices.
4. Determine the relative efficiency by estimating the ratio of the determinants.

Generate experimental designs for differing numbers of choice sets.

As noted above, fixed symmetric, fixed asymmetric, and randomized designs were generated. The fixed symmetric designs were generated using catalog experimental designs and the attribute shifting strategy proposed by Bunch et al. Shifting creates additional profiles by incrementally increasing each level of each attribute. As each level is "shifted," new profiles are created. Once a constant number of unique choice sets were generated for the fixed designs, the number of respondents was varied to create differing numbers of observations for the randomized designs. The randomized designs were then tested for their relative efficiency vis a vis the fixed designs.

Create a logit command file and estimate multinomial logit parameters.

The Systat Logit program was used for model estimation.

Calculate the determinants of the information matrices.

SPSS was used to calculate the determinants.

Determine the relative efficiency by estimating the ratio of the determinants.

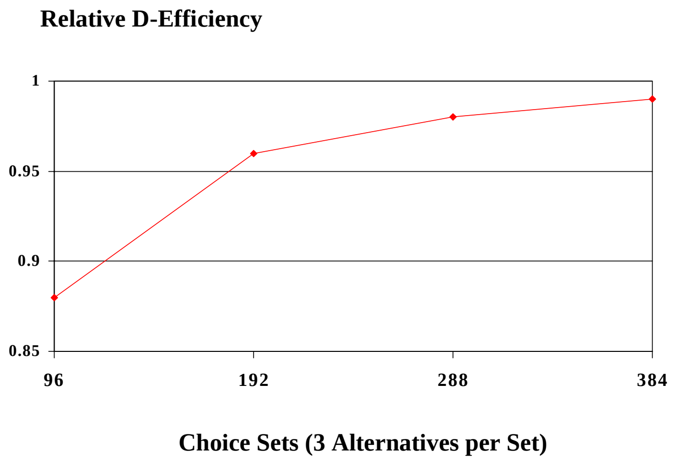
Several forms of efficiency are available to test the goodness of a design, but previous research has shown that they lead to similar conclusions. This research follows the majority of prior research in using D-efficiency as its criterion of efficiency. SPSS was employed for efficiency calculations.

Results

Phase 1: Symmetric Choice Experiments: Randomized vs. Fixed Design

In this phase of the study, the relative efficiency is the ratio of D-efficiency $_{\text{random}}$ divided by D-efficiency $_{\text{fixed symmetric}}$. Three alternatives were included in each choice set. Figure 1 indicates that even with a large symmetric (10 attributes each with 4 levels) choice experiment, the randomized design is 95% as efficient as the optimal fixed design with approximately 190 choice sets. Additional choice sets offer only modest improvements in relative D-efficiency.

Figure 1:
Symmetric Choice Experiment: Relative
Efficiency of Randomized Design



Obviously, design efficiency will not be the driver of sample size decisions for symmetric choice experiments. It is also clear that, for smaller symmetric designs, equivalent relative efficiency would require even fewer choice sets.

Phase 2: Asymmetric Choice Experiments: Randomized vs. Fixed Design

In this phase, the goal was to assess the relative efficiency of a randomized design vis a vis a fixed design for a large asymmetric choice experiment. The choice experiment consisted of 1 attribute at 7 levels, 4 attributes at 6 levels, and 3 attributes at 5 levels. Four alternatives per choice set were incorporated into the experimental design. A catalog design yielded 49 unique choice sets/tasks for the asymmetric choice experiment described above. Parameters were estimated for 20-23 simulated respondents. Similar to the initial phase of the study, the relative efficiency is the ratio of D-efficiency $_{\text{random}}$ divided by D-efficiency $_{\text{fixed asymmetric}}$.

Figure 2:
Asymmetric Choice Experiment: Relative Efficiency of Randomized Design

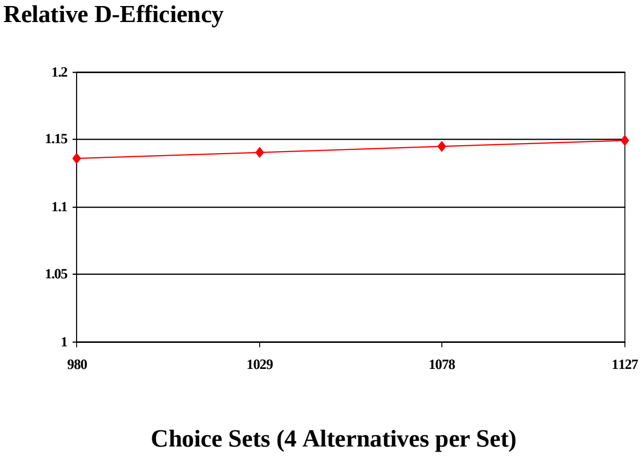


Figure 2 indicates that the randomized design is approximately 14% more efficient with 980 choice sets than the fixed asymmetric design with 49 choice tasks. Additional choice sets provide minimal improvement. Note that with even as few as about 1000 choice sets, which might be 50 respondents each receiving 20 tasks, that the randomized design is 15% more efficient than the fixed design.

Although empirical results are lacking, it is estimated that the randomized design with similar D-efficiency to the fixed design would require in the range of 60-70 unique choice sets.

When a smaller asymmetric choice experiment was run, it was clear that designs with fewer attributes and levels require fewer choice sets.

LIMITATIONS

With most exploratory research, study limitations abound. This study is no exception. This work is limited in that only a small handful of possible experimental designs were explored. In addition, no prohibitions were allowed in any of the designs. Further, only main effects were estimated. Further, simulated, not real, respondents were used to generate the parameter estimates. Finally, only aggregate level estimation was considered in this paper. Neither individual nor segment level estimates were derived.

WHAT WE LEARNED

In this exploratory effort, the relative efficiency of randomized vs. fixed experimental designs was investigated for selected choice experiments. The number of choice sets is determined by the relative efficiency of competitive designs. In this study, D-efficiency was the measure employed.

The total number of choice sets required is contingent upon the size of the choice experiment and the symmetrical or asymmetrical nature of the experiment. Larger, asymmetrical choice experiments require more choice sets.

For the large symmetrical choice experiment, the randomized design is 95% as efficient as the optimal fixed design with approximately 190 choice sets. For the large asymmetrical choice experiment investigated in this study, it was found that the randomized design is approximately 14% more efficient with 980 choice sets than the fixed asymmetric design with 49 choice tasks.

Given the relatively small numbers of choice sets required to achieve reasonable D-efficiency, design efficiency is not likely to be the driver of sample size calculation for a choice based conjoint study. When estimating sample size for a choice based conjoint study, the analyst needs to be less concerned with experimental design efficiency than with other key elements of the process. One critical element is the accuracy of the choice model. A larger number of observations (whether achieved by varying the number of choice sets, alternatives per choice set, or respondents) is one way to increase the probability of developing an accurate choice model (that is, a model that generates valid and reliable effects and therefore, choice share estimates). More observations will reduce the standard errors and increase the potential for a greater number of effects to be statistically significant. This is likely to increase model accuracy.

REFERENCES

Addelman, Sidney (1962) “Symmetrical and Asymmetrical Fractional Factorial Plans,” *Technometrics*, 4 (February), 47-58.

Anderson, Don, W. Kuhfeld and M. Garratt (1996) “Advanced Principles of Experimental Design,” Tutorial presented at the American Marketing Association’s Advanced Research Techniques Forum.

________ et. al. (1993) “Estimating Availability Effects in Travel Choice Modeling: A Stated Choice Approach,” Transportation Research Record 1357, TRB, National Research Council, Washington D.C., 51-65.

________ and J. Wiley (1992), “Efficient Choice Set Designs for Estimating Availability Cross Effects Models,” *Marketing Letters*, 3:4, 357-370.

Bunch, David S, J.J. Louviere and D.A. Anderson (1994) “A Comparison of Experimental Design Strategies for Choice Based Conjoint Analysis with Generic-Attribute Multinomial Logit Models,” Working Paper, UC, Davis, Davis, CA.

Childress, Robert L. (1974) *Mathematics for Managerial Decisions*. Englewood Cliffs, NJ: Prentice Hall.

Huber, Joel and Klaus Zwerina (1996) “The Importance of Utility Balance in Efficient Choice Designs,” *Journal of Marketing Research*, 33 (August), 307-17.

Kuhfeld, Warren (1997) “Efficient Experimental Designs Using CVA Design Software,” Sawtooth Software Conference Proceedings, Sequim WA: Sawtooth Software.

_______, R.D. Tobias and M. Garratt (1994) “Efficient Experimental Design with Marketing Research Applications,” *Journal of Marketing Research*, 31 (November), 545-557.

Lazari, Andreas G. and D. Anderson (1994) "Designs of Discrete Choice Experiments for Estimating Both Attribute and Availability Cross Effects," *Journal of Marketing Research*, 31 (August), 375-83.

Sawtooth Software (1995), CBC User Manual, Version 1.2, Sequim, WA.

COMMENT ON MULHERN

Richard M. Johnson
Sawtooth Software, Inc.

I'm very pleased by Mike's results.

When I first thought about planning what eventually became the CBC System, I realized that there were two ways to create study designs:

1. including a large catalog of "fixed" designs, or
2. providing an algorithm for randomized designs.

There were many reasons for favoring randomized designs, but one big negative: I knew that randomized designs could be less efficient than fixed designs. My own computations suggested that they would be only 5 or 10% less efficient than fixed designs if indeed it was possible to produce a fixed design that was both orthogonal and balanced. But I didn't investigate the case of attributes with differing numbers of levels, which we encounter much more frequently in real life.

So I was pleased to see Mike's confirmation that randomized choice designs are nearly as efficient as fixed designs when all attributes have the same number of levels. But I was surprised and delighted to learn that randomized designs appear to be ***more*** efficient than fixed designs when attributes have different numbers of levels.

This seems to make sense, especially when the different numbers of levels are relatively prime, such as 7, 5, and 3 in Mike's example. In that case it is usually not possible to come up with a fixed design that is also both orthogonal and balanced, while also having only a modest number of choice sets.

However, Mike's results do stop short of answering an important related question. Often CBC is used with paper and pencil questionnaires rather than computer-assisted interviews. In those cases it's not practical to give every respondent a unique set of choice tasks, so there are usually a few fixed versions of the questionnaire. Often CBC is used to generate those versions. The question is how many versions there should be. The problem is made more difficult by the fact that in aggregate analyses of choice data one often wants to measure interactions as well as main effects.

Stimulated by Mike's paper, I did a little analysis to answer this question. I used CBC to create questionnaire designs of varying length, under three conditions:

Each attribute with 3 levels, and 3 alternatives per task

Each attribute with 5 levels, and 5 alternatives per task

Each attribute with 9 levels, and 9 alternatives per task

I investigated standard errors for different numbers of versions. Assuming that each version had 20 choice sets, and there were a total of 300 respondents divided among different numbers of versions, here are the maximum standard errors in each case:

Versions	3x3	5x5	9x9
2	.041	.057	.137
3	.043	.055	.128
4	.039	.051	.124
5	.038	.050	.122
6	.039	.050	.121

All of these standard errors assume the same total of 300 respondents. Of course, they could all be reduced by increasing the sample size. What is relevant here is that the standard errors when there are only two versions aren’t much larger than when there are more. In all of these cases, a total of only 40 choice sets (2 versions, each with 20 tasks) seems to be enough to do the job. Of course, these were high-quality randomized designs, based on CBC’s complete enumeration method that does a good job of both orthogonality and balance. And, to be safe, I’d recommend doubling that number.

So my conclusion is similar to Mike’s: randomized designs are surprising robust. Not only are they efficient for measuring main effects as Mike found, but relatively few versions are needed to estimate two way interactions for paper and pencil questionnaires.

FULL VERSUS PARTIAL PROFILE CHOICE EXPERIMENTS: AGGREGATE AND DISAGGREGATE COMPARISONS

Keith Chrzan
IntelliQuest, Inc.

INTRODUCTION

Conjoint analysis has proven to be a versatile tool for applied researchers (Wittink and Cattin 1989, Green and Srinivasan 1978, 1990). A constant theme in the development and subsequent refinements of conjoint analysis has been to get more results with less input. Orthogonal designs that yield individual respondent level models were a boon, but they proved difficult for respondents once the number of attributes grew too large. Adaptive conjoint analysis (ACA) produced more or less the same quality of results, but with simpler inputs from respondents that allowed more attributes to be measured. Choice-based conjoint analysis (Louviere and Woodworth 1983, Louviere 1988) expanded the array of modeling options available to the researcher by making interactions easier to measure, and by allowing complex and important entities like alternative-specific effects and cross-effects to be quantified. The marriage of a simple respondent task with powerful modeling capability has been an elusive goal among practitioners.

However, separate developments along these lines have recently been introduced for choice-based conjoint analysis. Partial Profile Choice Experiments (PPCE) simplify the respondents' task by using partial profile stimuli. Producing disaggregate choice utilities via hierarchical Bayesian modeling or with the Sawtooth Software ICE product is another powerful modeling capability recently added to the marketing researcher's tool kit. One goal of this study is to test further the quality and validity of using partial profile stimuli. Some of the tests involve the combination of partial profile stimuli as inputs with disaggregate utilities as outputs. This marriage of getting more (disaggregate utilities) from less (partial profile stimuli) is the second goal of this study.

PARTIAL PROFILE CONJOINT ANALYSIS

Partial Profile Choice Experiments (PPCE), were intended to redress the purported inability of full profile conjoint analysis to accommodate large numbers (i.e. more than about six) of attributes (Chrzan and Elrod 1995). In the past few years PPCE has been tested and extended (Chrzan, Bunch and Lockhart 1996, Chrzan and Skrapits 1997, Chrzan 1998). Salient among its benefits, PPCE

- uses experimental designs constructed from a simple recipe to accommodate literally any number of attributes, trading off all attributes against one another while controlling for profile and attribute order effects;
- features experimental designs constructed in blocks that contain only as many choice questions as there are attributes in a study; since a given respondent needs to complete only a single experimental block, which limits the size of the respondent's task;

- reduces unexplained error by offering respondents choices among alternatives that differ on just three or five attributes, regardless of the number of attributes in the study (limiting the complexity of the respondent’s task reduces unexplained error and more than compensates for the non-orthogonality of PPCE designs);
- allows analysis *via* standard multinomial logit software;
- produces utilities that are calibrated by logit, and that are thus scaled appropriately for logit-based simulation modeling;
- produces utilities that are free of biases potentially caused by the differential absence and presence of attributes in the choice sets.

For example, in a study of nine attributes of personal computers, a standard full profile choice based conjoint analysis question would look like this:

If two personal computers were alike in all other ways, which would you rather buy?

Personal Computer A	Personal Computer B
PC brand 1	PC brand 3
233 MHz microprocessor	266 MHz microprocessor
3.2 gigabyte hard drive	6.4 gigabyte hard drive
48 MB RAM	32 MB RAM
20X CD-ROM	14X CD-ROM
1.44 MB disk drive	ZIP drive and 1.44 MB disk drive
17” monitor	15” monitor
56K modem	56K X2 modem
Price: $1,299	Price: $1,499

A partial profile question for the same study will look like this:

If two personal computers were alike in all other ways, which would you rather buy?

Personal Computer A	Personal Computer B
PC brand 1	PC brand 3
233 MHz microprocessor	266 MHz microprocessor
48 MB RAM	32 MB RAM
20X CD-ROM	14X CD-ROM
1.44 MB disk drive	ZIP drive and 1.44 MB disk drive

or even like this:

If two personal computers were alike in all other ways, which would you rather buy?

Personal Computer A	Personal Computer B
PC brand 1	PC brand 3
233 MHz microprocessor	266 MHz microprocessor
48 MB RAM	32 MB RAM

The PPCE question requires the respondent to trade off five or fewer of the attributes at a time. Of course, subsequent choice questions will use different subsets of attributes.

The Quality of PPCE Designs

As noted, a single PPCE question requires respondents to trade off five or fewer attributes. If there are 10 (or 20 or 50, or more) attributes in the study, PPCE provides the respondent with a much easier task than would full profile tradeoffs of all attributes. This easier task should reduce error in the response process. Such "unexplained error" is proportional to the inverse of the multinomial logit scale parameter μ (Ben-Akiva and Lerman 1985). In effect, this error shows up as relatively shrunken coefficients in logit models: if model A has more unexplained error than model B, its coefficients tend to be smaller across the board. A method for identifying a model's relative scale parameter appears in Swait and Louviere (1993).

Partially counterbalancing this advantage is the fact that PPCE designs are not orthogonal, so their estimation is inefficient relative to full profile designs. One measure of efficiency, called D-efficiency, is a function of the multinomial logit covariance matrix (Bunch *et al.* 1994, Hosmer and Lemeshow 1989). The relative design efficiency of PPCE is

$$\left[\frac{\det(I(\beta))_{PPCE}}{\det(I(\beta))_{FP}}\right]^{1/p}$$

where I(β) is the "normalized" information matrix and p the number of parameters in the model (Bunch *et al.* 1994).

Combining this D-efficiency measure with the notions of differences in scale noted above yields:

$$\left[\frac{\det(I(\beta,\mu_{PPCE},X_{PPCE}))}{\det(I(\beta,X_{FP}))}\right]^{1/p}. \qquad (1)$$

PPCE will be more efficient than full profile overall if the value of expression (1) exceeds 1.0.

INDIVIDUAL CHOICE ESTIMATION

Recent advances in the use of hierarchical Bayesian models has enabled researchers to estimate individual respondents' utilities from choice-based conjoint analysis data (Allenby and Ginter 1995). ICE is the Sawtooth Software product for estimating individual respondent utilities from choice-based conjoint experiments (Sawtooth Software 1997). ICE has been found to improve aggregate share predictions (Huber *et al.* 1998). Another benefit is that individual respondent utilities are valuable for analysis of *a priori* segments and for the generation of *a posteriori* segments via cluster analysis.

EMPIRICAL TESTS

A large scale commercial study illustrates the design and analysis of a PPCE and allows various comparisons of PPCE and full-profile choice-based conjoint analysis:

1. aggregate model comparisons
   a) experimental design efficiency
   b) equality of model parameters
   c) predictive validity (mean absolute error of prediction)
2. disaggregate model comparisons: predictive validity (hits, mean absolute error of prediction)

EMPIRICAL STUDY

Design

Nine attributes of personal computers were identified as important in previous research on the category. A total of 1,358 personal computer users completed a web-based interview. Respondents were members of IntelliQuest's Consumer Web Panel. Eight of the attributes had three levels each and one had just two levels.

The partial profile experiment was constructed using the recipe appearing in Appendix I. The 36 partial profile questions were blocked into two cells of 18 questions each; each partial profile question displayed just five attributes being traded-off.

Thirty full profile choice questions were generated in a two-step process. After generating an efficient design using the Trial Run experimental design software (SPSS 1997), a shifting strategy was used to produce the additional profiles required to comprise the choice sets (Bunch *et al.* 1994). The 30 full profile questions were blocked into two cells of 15 questions each, and each full profile question of course included all nine attributes. All choice questions included two alternatives. No "other" or "none" alternatives appeared in any of the choice sets.

Each cell included two unique holdout questions, which featured just five attributes in the partial profile cells and all nine attributes in the full profile cells. Each of the eight holdout questions contained three alternatives. Holdouts and the 15 to 18 model calibration choice sets were interspersed in random order.

Respondent counts, holdout identity and experimental conditions for the four cells were as follows:

		Number	Calibration	
		Of	Choice	
Cell	Profiles	Respondents	Sets	Holdouts
1	Full	321	15	A&B
2	Full	348	15	C&D
3	Partial	330	18	E&F
4	Partial	359	18	G&H

PLANNED COMPARISONS

Aggregate model comparisons

Before comparing results from the individual choice model estimates provided by ICE, the aggregate models will be compared.

The first comparison of the aggregate models involves testing the parameter (utility) vectors for equality using the Swait and Louviere (1993) test. This is a test for parameter equality after correcting for the potential difference in scales of the two models' parameters. Previous studies comparing PPCE and full-profile choice-based conjoint analysis found significant differences in scale, but no significant differences in model parameters.

The relative experimental design efficiencies of the two models will also be interesting. Previous studies have shown PPCE to be 20% to 30% more efficient than full-profile choice-based conjoint models.

Finally, the ability of the two models to predict holdout shares can be compared. Admittedly this is a poor man's measure of predictive validity, but it is all we have available. Vectors of shares predicted by the PPCE and full-profile models can be compared to actual holdout choice shares, and their mean absolute deviations from the holdout shares can be tested via a dependent t-test. No previous study of PPCE has included this metric.

Disaggregate model comparisons

Sawtooth Software's ICE procedure will be run separately for the full and partial profile data sets, so individual level model parameters will be produced for all respondents. The predictive validity of the individual utility data may be tested in two ways: the t-test mentioned above for the aggregate model and the differences in "hit rates" (the percentage of correctly predicted holdout choices).

These comparisons are repeated for disaggregate utilities produced via hierarchical Bayes conjoint analysis kindly run by Rich Johnson.

RESULTS

Aggregate Analyses

MNL analysis *via* SYSTAT Logit produced the model parameters.

The parameter vectors need to be put on the same scale before they can be compared fairly. The full profile model had more unexplained error and so a smaller scale. The scale factor for the full profile model was .54, which means the full profile model contains about twice as much unexplained error as the partial profile model, a finding consistent with past studies.

Table 1 shows parameters of the partial profile model and raw and scale-adjusted parameters for the full profile model.

Table 1
Model Parameters for the Personal Computer Study

		Coefficients	
		Raw	Rescaled*
Parameter	Full Profile	Full Profile	PPCE
PC Brand 1	-.01	-.02	**.11**
PC Brand 2	**-.07**	**-.14**	**-.22**
233 MHz Processor	**-.33**	**-.61**	**-.70**
266 Mhz Processor	.03	-.05	.04
3.2 gigabyte hard drive	**-.32**	**-.59**	**-.62**
6.4 gigabyte hard drive	**.09**	**.16**	**.20**
32 MB RAM	**-.20**	**-.36**	**-.50**
48 MB RAM	**-.04**	**-.08**	.04
14X CD-ROM	**-.15**	**-.27**	**-.36**
20X CD-ROM	-.01	-.01	.06
1.44 MB disk drive	**-.14**	**-.25**	**-.31**
1.44 MB and ZIP drive	.**17**	**.31**	**.38**
15" monitor	**-.25**	**-.47**	**-.46**
17" monitor	**.04**	**.08**	**.19**
$1,299	**.30**	**.56**	**.40**
$1,499	**.06**	**.11**	**.11**
56K modem	.00	.01	**-.07**

Bold: coefficient is significantly different from 0 at p < .05

Underline: full and partial profile coefficients are significantly different from one another, p < .05.

* Rescaled to PPCE scale

The PPCE and scale-adjusted full profile model parameters look similar, and they tell the same relative story. The two sets of coefficients are, however, significantly different (χ^2 = 73.024 with 18 degrees of freedom, p < .001 via the Swait and Louviere test). Five of the coefficients are identified as differing significantly (p = .05) between full and partial profile models using a logit Chow test. This result is different than for previous studies, and may owe to the greater sensitivity that the large sample size affords in this study.

The smaller scale factor of the full profile model harms its efficiency, so that it is only 76% as efficient as the partial profile model. This takes into account and corrects for the fact that there were slightly more partial profile respondents than full profile respondents and that each partial profile respondent answered three more choice questions than did each full profile respondent. The 76% efficiency means that a partial profile model could attain the same level of precision as a full profile model, but with almost a fourth (24%) fewer respondents.

In terms of predictive validity, the full and partial profile models do not differ significantly. The eight holdout questions (four full profile and four partial profile) each have three alternatives, for a total of 24 holdout choice shares. Table 2 shows how the actual holdout shares and the full profile and partial profile predictions of these shares compare. Utility scales were adjusted appropriately when partial profile utilities were used to predict full profile holdouts and vice versa.

Table 2
Predictive Validity – Aggregate Models

Holdout Share	Full Profile Simulation	PPCE Simulation
26.8	28.9	27.7
18.1	18.6	23.6
55.1	52.5	48.7
8.4	10.2	8.2
13.4	16.8	15.8
78.2	73.1	76.0
39.1	39.6	34.3
26.7	30.1	38.3
34.2	30.3	27.4
58.0	52.8	48.7
15.8	19.0	18.7
26.1	28.2	32.6
45.2	41.9	47.4
45.2	41.4	37.5
9.7	16.7	15.2
75.2	74.5	77.2
14.2	20.1	14.0
10.6	5.3	8.8
31.2	37.1	36.2
59.1	52.4	54.5
9.7	10.5	9.3
38.4	33.9	32.2
33.1	21.9	26.7
28.4	44.1	41.1
MAE	4.36	4.76

Test for difference in MAE: t = .99, p = .33

The mean absolute deviations from the actual holdout shares are not significantly different: 4.36 percentage points for the full profile model and 4.76 percentage points for the PPCE model.

(t = .56 with 23 degrees of freedom, p = .58).

Disaggregate Analyses

Individual choice utilities were produced by ICE analysis. First, latent class analysis solutions were produced via the CBC Latent Class module for two through six latent classes. These were used to start ICE analyses with two to six "basis vectors." The latent class and ICE analyses were performed separately for the full profile and partial profile samples. For both the full and partial profile samples, the ICE solution with five basis vectors produced simulated shares which best matched holdout shares, so the five basis vector solutions are used for subsequent analyses and comparisons.

The utilities produced by ICE allow predictions of individual respondents' holdout choices. ICE utilities with a first choice model were used to predict each respondent's holdout responses. Assumed by the first choice model is that each respondent chooses the alternative that has the highest utility for her. The first choice model is appropriate because the study involved home personal computer purchase, not a fast moving packaged goods product wherein variety seeking might occur. Scale transformations and use of the ICE logit market simulator did not appreciably improve predictive accuracy.

"Hits" are holdout responses correctly predicted. Each respondent answered two holdout choices. With three alternatives each, random predictions would be correct 33% of the time. The hit rate was 50.7% for full profile respondents and full profile holdouts. For partial profile respondents predicting partial profile holdouts, the hit rate was 56.2%. Both the full and partial profile ICE models significantly outperform chance predictions. The partial profile ICE hit rate is significantly higher than that for the full profile ICE model (Z = 2.86, p =.01).

Share predictions for the two ICE models appear in Table 3.

Table 3
Predictive Validity – ICE Models

Holdout Share	Full Profile Simulation	PPCE Simulation
26.8	30.9	28.3
18.1	18.1	14.7
55.1	51.0	57.0
8.4	13.2	1.7
13.4	12.4	21.6
78.2	74.4	76.6
39.1	32.6	43.7
26.7	41.7	34.7
34.2	25.7	21.6
58.0	47.5	57.5
15.8	21.1	7.3
26.1	31.4	35.3
45.2	41.9	51.7
45.2	45.6	37.2
9.7	12.6	11.2
75.2	69.8	85.6
14.2	19.7	10.9
10.6	10.5	3.5
31.2	42.0	9.0
59.1	45.6	72.3
9.7	12.4	18.7
38.4	9.7	35.6
33.1	30.5	24.2
28.4	59.8	40.2
MAE	7.34	7.14

Test for difference in MAE: t = .11, p = .91

Again mean absolute deviations from the actual holdout shares are not significantly different for the full and partial profile models: 7.34 percentage points for the full profile model and 7.14 percentage points for the PPCE model (t = .11 with 23 degrees of freedom, p = .91).

The parity of full and partial profile models appears again in disaggregate analyses using the hierarchical Bayesian conjoint utilities: MAEs are 5.18 for full profile predictions and 4.49 for predictions made from partial profile utilities (t = .82, p = .42 with 23 degrees of freedom). Finally, no significant difference between full and partial profile utilities occurs in terms of hits: 60.4% for partial profile compared to 59.7% for full profile (Z = .35. p = .73).

DISCUSSION

Analysis Summary

At the aggregate level, PPCE and full-profile choice-based conjoint analysis have comparable predictive validity and similar, albeit significantly different, coefficients. The big difference between the two is the great (nearly 50%) reduction in unexplained error that results from the use of partial profiles. This makes the PPCE experiment so much more efficient that one could field it with 24% fewer respondents than a full profile model would require for the same precision. This is a substantial practical advantage.

The significant differences in model parameters does not lead to a significant difference in predictive validity, so neither model is demonstrably nearer to the “truth” than is the other.

As for the individual level comparisons allowed by ICE, the full and partial profile models have mixed results on predictive validity. Mean absolute errors of prediction are not significantly different for the two models; the partial profile model has a significantly greater hit rate than does the full profile model using ICE utilities, but not so using utilities generated from hierarchical Bayesian conjoint analysis.

Implications

Consistent with previous research, PPCE matches the quality of full profile choice-based conjoint analysis. It greatly reduces unexplained error variance, however, due to the more respondent-friendly choice task that trades off only five attributes at a time. This means that researchers can cut their sample sizes by about 20% - 25% and still have the same precision afforded by full profile choice-based conjoint analysis. More importantly, PPCE offers researchers a way to do choice-based conjoint analysis with many more attributes than if they used full profiles.

The parity performance of PPCE and full profile choice-based conjoint when analyzed via ICE (and hierarchical Bayesian conjoint analysis) was a surprise. The relative sparseness of the PPCE data set could have been expected to hamper the operation of ICE. Since PPCE and full-profile conjoint performed equally well in the predictive validity of their disaggregate ICE utilities, however, it appears that the marriage of disaggregate analysis and partial profile conjoint analysis may allow researchers to construct conjoint models with large numbers of attributes and still get individual level utility estimates.

Another surprising result is the superior performance of the aggregate MNL model in predicting holdout choice shares. For both full and partial profile samples, mean absolute error of prediction is greater for ICE than for the aggregate MNL model (MAE from MNL and from hierarchical Bayesian utilities are not significantly different). This finding is contrary to earlier findings (Huber *et al.* 1998) and probably reflects ICE’s instability when too few observations are collected from each individual respondent.

REFERENCES

Allenby, Greg M. and James L. Ginter (1995) "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, 392-403.

Ben-Akiva, Moshe and Steve Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Bunch, David S., Jordan J. Louviere and Don Anderson (1994) "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, Keith (1998) "Design Efficiency of Partial Profile Choice Experiments," paper presented at the 1998 INFORMS Marketing Science Conference, Paris.

Chrzan, Keith and Mike Skrapits (1997) "Testing for IIA Violations in Partial Profile Conjoint Models," paper presented at the INFORMS Marketing Science Conference, Berkeley, CA.

Chrzan, Keith, David Bunch and Daniel C. Lockhart (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, Keith and Terry Elrod (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes," *1995 Advanced Research Techniques Conference Proceedings*. Chicago: American Marketing Association (in press).

Green, Paul E. and V. Srinivasan (1978) "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September) 103-23.

Green, Paul E. and V. Srinivasan (1990) "Conjoint Analysis in Marketing Research: New Developments and Directions," *Journal of Marketing*, 54 (October) pp. 3-19.

Hosmer, David W. and Stanley Lemeshow (1989) *Applied Logistic Regression*. New York: Wiley.

Huber, Joel and Klaus B. Zwerina (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33 (August) 307-17.

Huber, Joel, Rich Johnson and Neeraj Arora (1998) "Capturing Heterogeneity in Consumer Choices," paper presented at the Ninth Annual A/R/T Forum, Keystone, Colorado.

Kuhfeld, Warren, Randal D. Tobias and Mark Garratt (1995) "Efficient Experimental Designs with Marketing Research Applications," *Journal of Marketing Research*, 31 (November) 545-57.

Louviere, Jordan J. (1988) "Analyzing Decision Making: Metric Conjoint Analysis" Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-067. Beverly Hills: Sage.

Louviere, Jordan J. and George Woodworth (1983) "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20 (November) pp. 350-67.

Sawtooth Software Inc. (1993) *The CBC System of Choice-Based Conjoint Analysis,* Sun Valley: Sawtooth Software.

Sawtooth Software Inc. (1997) *ICE Module: An add-on to the The CBC System of Choice-Based Conjoint Analysis,* Sequim: Sawtooth Software.

SPSS (1997) *Trial Run*. Chicago: SPSS.

Steinberg, Dan and Phillip Colla (1991) *LOGIT: A Supplementary Module for SYSTAT*. Evanston, IL: SYSTAT Inc.

Swait, Joffre and Jordan Louviere (1993) “The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models,” *Journal of Marketing Research,* 30 (August) 305-14.

Wittink, Dick R. and Philippe Cattin (1989) “Commercial Use of Conjoint Analysis: An Update,” *Journal of Marketing,* 53 (July) 91-6.

APPENDIX 1

DESIGN RECIPE FOR PPCE WITH FIVE ATTRIBUTES/PROFILE

The Design of PPCEs

The following section details the design strategy for choice questions (sets of profiles) differing on five attributes. A design recipe for PPCE differing on three attribute appears in Appendix 2.

1. Compute the number of blocks (B) to be included in the design.
   B=1+[(A-5)/(4)], rounding down any fraction.
   Assign n/B respondents to each block.

2. Starting row of Cth block:

   $$(1\ \ 0^{C-1}\ -1\ 0^{C-1})^{2}\ \ 0^{(A-4C-1)}\ \ X$$

   if B is even, X = 1
   if B is odd, X = -1

   Superscripts refer to the number of times a term appears. For example, if A = 9, B = 1 + (9-5)/4 = 2 and the starting rows of the two blocks are

   1 -1 1 -1 0 0 0 0 1

   and

   1 0 -1 0 1 0 -1 0 -1

3. If all attributes are binary, go to Step 4. If not, duplicate these starting rows but reverse the signs of the Xs on the duplicates.

4. For each block, take row 1 and shift cells one place to the right, wrapping rightmost cell around back to the left. Perform this step A – 1 times for each starting row. In this example, each block has nine rows and nine columns (each block is square).

At this point we have B blocks of A rows each. Each row translates into a choice question. Attributes coded “1” or “-1” in a row are the attributes that will appear in that question’s partial profiles. Attributes coded “0” will not appear. This cyclical design does a pretty good job of keeping attribute presence correlations low. Next we need a method for assigning attribute levels to partial profiles.

Bunch *et al*. (1994) have shown that multinomial logit models are most efficient when a) they are based on designs that orthogonalize attributes and b) the strategy for assigning profiles to choice sets maximizes attribute differences among profiles within choice sets. Bunch *et al.* (1994) recommend “shifting” such that a three level attribute at level 2 in one profile of a choice triple will be at level 3 in the next profile within the choice set and at level 1 in the third.

For PPCE, modify the Bunch *et al.* (1994) shifting strategy:

First, create the first profile in each choice set. For binary attributes, let 1 be the first level of the attribute and let the second level of the attribute occur when the design calls for a –1. For attributes with 3+ levels, randomly assign levels to the attributes that appear in the choice set.

To create the second (third, etc.) profile in each choice set, increment the level of each attribute that appears in the first profile by one to create the second profile in each choice set, by two to create the third (fourth, etc.), wrapping around to level one when an attribute's number of levels is exceeded.

This "recipe" approach has been found to produce good designs (Chrzan 1998). Ways to increase design efficiency further are

a. increase "utility balance" among profiles, and adjust the design to increase it (Huber and Zwerina 1995);

b. create "randomized" designs wherein levels (perhaps even attributes) are randomized into choice sets (Chrzan and Skrapits 1997);

c. use computer search algorithms to find efficient designs (Kuhfeld *et al.* 1995).

APPENDIX 2

DESIGN RECIPE FOR THREE ATTRIBUTES/PROFILE

This strategy is the same as that in the previous Appendix, with these changes:

1. Compute the number of blocks (B) to be included in the design.

   If A is odd, B = (A-1)/2.
   If A is even, B = A/2.
   Assign n/B respondents to each block (n=sample size).

2. Design the first row of each of the B blocks.
   The first row of Cth block:

   $1 \ 0^{C-1} \ 1 \ 0^{A-C-2} \ X$

COMMENT ON CHRZAN

Bryan K. Orme
Sawtooth Software, Inc.

Keith's findings offer some positive news for partial profile advocates. He confirms earlier findings that unexplained error can be lower in Partial Profile (PP) than Full Profile (FP), despite the fact that each PP task contributes less information in the independent variable matrix. This finding is not unlike Huber and Hansen's discovery in the mid 80s that respondents to ACA surveys provided better overall data if they evaluated only two or three, rather than more attributes at a time (Huber and Hansen 1986).

Keith's partial profile design rotated the presentation of attributes within respondent. In the first task, a respondent might see attributes 1, 2, 3, 4 and 5, and in the second task, that same respondent might see attributes 5, 6, 7, 8 and 9. For each task, respondents must orient themselves to a new subset of attributes and their positions. If one intends to analyze the data using an aggregate method such as logit or Lclass, it might not make sense to cycle all attributes through each respondent's interview. Respondents could answer more quickly and reliably if the same subset of attributes were presented in the same order. The subsets of attributes and order of presentation probably should be randomized across respondents, to control context and order bias. I'd be interested in seeing a split-sample study conducted to test whether attribute selection and presentation order should be randomized across respondents or within respondents for partial profile.

Most of my experience with ICE has been positive, so I naturally questioned why it performed so poorly in Keith's study. My first concern regarding the FP cells was that the effects were (as Keith notes) quite flat, reflecting a great deal of noise in the choices. Secondly, the suggested number of choice tasks for analysis with ICE is 20 or more (Johnson 1997). This study used 15 for FP. Finally, the published evidence I'm aware of for ICE has been based on three or more alternatives per task, whereas Keith used only pairs in the calibration tasks. At this forum in 1997, Pinnell suggested that pairs were less efficient and potentially biased relative to more alternatives per task (Pinnell 1997). Respondents, he argued, could deal efficiently and well with more alternatives per task.

Keith was kind enough to share his data with me, permitting an investigating of why ICE performed so poorly. For the analysis below, I considered only the full-profile cells (n=669).

I independently computed ICE utilities also based on a five-group latent class solution. The first thing I noticed is that my ICE predictions of holdout shares were also very bad. I also noticed that the share predictions of the 24 product concepts were quite different from Keith's. Both Lclass and ICE are subject to local minima, so one can get a different answer from a different random starting point.

Given that ICE is an extension of underlying Lclass basis vectors, I decided to set ICE aside and try only the Lclass vectors for predicting shares. Lclass performed slightly better than ICE, but unacceptably worse than logit. Believing that the problem might have to do as much with model reliability as validity, I generated four separate five-group latent class solution from different random starting points. I computed shares for the eight holdout tasks, and (ignoring the actual holdout choices) compared the results across the four replications. The average MAE (mean average error) across replications was 7.51, and the largest difference between computed shares was 29 share points! This means that a product share predicted by a Lclass solution using one random starting point might be 30%, and from another starting point 59%. Next, I computed the across replicate reliability for a simpler three-group Lclass solution, and achieved an average MAE of 2.8, with the largest difference in a predicted share between replicates of 12 share points. This was better, but in my opinion still represented unacceptable instability.

For comparison, I performed a similar exercise with the data set that Huber, Orme and Miller used for their presentation at this same conference. My experience with that data set suggested it to be quite clean. This data set involved choices among television sets, collected at a mall intercept with 352 respondents. Respondents completed 18 choice tasks, with five concepts per task. As with Keith's data, I generated share predictions for eight tasks with three concepts each. The average across-replicate MAE for a five-group Lclass solution for the TV data set was 0.73 (and the largest share difference was 2.61)—significantly (ten times) less variable in terms of MAE than for Keith's data.

I believe two elements might have made it difficult to get good Lclass or ICE predictions with Keith's data. The degree of noise in the data combined with relatively low information content per respondent interview may be causing instability and overfitting, given the number of parameters in a five-group Lclass solution (part worths plus individual probability weights). Given that Lclass could not generate stable share estimates in repeated trials with a five-group solution, it comes as no surprise that ICE, which is an extension to Lclass requiring the estimation of many more parameters, also failed.

Keith's findings suggest that HB may produce more stable individual-level utility estimates than ICE when the information content per respondent is relatively low. However, applying HB does not appreciably improve predictions of holdout choices for this study relative to aggregate analysis. With this data set, the simpler logit model is hard to beat.

Keith's study makes a compelling argument for using holdout choices in both methodological and commercial studies. Without holdout choices, it is difficult to compare the performance of different methods. Keith would have erred to trust either the ICE or a high-order Lclass solution, and thanks to the holdout choices he avoided that mistake.

The common argument I've heard from practitioners in the trenches is that they can't afford to include fixed holdout choice tasks in commercial studies that aren't used in utility estimation. This case study makes me wonder whether one can afford not to include fixed tasks. I strongly advocate including a few well-designed fixed choice tasks in any conjoint study that will be used to develop a predictive market simulator. Huber, Orme and Miller present strong evidence at this conference that holdout choice sets designed with minimal level overlap are not as useful as they could be (Huber, Orme and Miller 1999). Holdout choice sets should reflect a healthy degree of differential similarity (imbalanced level overlap) in the alternatives so that substitution behavior can be modeled for more accurate market share predictions. I’d also suggest that fixed tasks don’t necessarily need to be, in the end, “held out.” One can hold them out when comparing and tuning predictive models. Afterward, they might be added back and used in part worth estimation.

REFERENCES

Huber, Joel and David Hansen (1986), “Testing the Impact of Dimensional Complexity and Affective Difference of Paired Concepts in Adaptive Conjoint Analysis,” Association of Consumer Research Conference.

Huber, Joel, Bryan K. Orme, and Richard Miller (1999), “Dealing with Similarity in Conjoint Simulations,” Sawtooth Software Conference Proceedings.

Johnson, Richard M. (1997), “ICE Technical Paper,” Sawtooth Software Technical Paper.

Pinnell, Jonathan (1997), “The Number of Choice Alternatives in Discrete Choice Modeling,” Sawtooth Software Proceedings.

DEALING WITH PRODUCT SIMILARITY IN CONJOINT SIMULATIONS

Joel Huber
Duke University
Bryan K. Orme
Sawtooth Software, Inc.
Richard Miller
Consumer Pulse

ABSTRACT

Choice simulators have provided very useful ways to allow researchers to project from individual or group preference models to predictions of market share. We propose that effective choice simulators need three properties to effectively mirror market behavior. First, they need to display *differential impact* so that a marketing action at the individual or homogeneous segment level has maximal impact near a threshold but has minimal impact otherwise. Second, simulators need to reflect *differential substitution,* assuring that alternatives take proportionately more share from similar than dissimilar competitors. Finally, they need to exhibit *differential enhancement*, a property whereby a small value difference has a big impact on highly similar competitors but almost none on dissimilar ones. A new method we call Randomized First Choice enables simulators to closely match these properties in market behavior. The method begins with a first choice model in which the item with the highest utility is chosen, but modifies that by adding two kinds of variability. The first, product value variability, adds variability to the alternatives, while the second, attribute value variability, adds variability to attribute part worths. While product variability is mathematically equivalent to the commonly used adjustment for scale, the less known attribute variability is critical if one desires to approximate substitution and enhancement properties among similar alternatives.

We test the value of adding both kinds of error in a choice-based conjoint study of 352 consumers. Respondents make choices among specially designed holdout sets that include perfect duplicates, where one alternative is identical to another in a set, and near duplicates, where the pair only differ on the two minor attributes. Comparing predictions against actual choice shares for these stimuli provides a strong test of the predictive ability of various choice simulators. Additionally, we illustrate the way the methods differ in their representation of differential impact, substitution and enhancement.

We assess the impact of Randomized First Choice applied to four commonly used models: an aggregate logit model, a latent class model, and individual-level models, Sawtooth's ICE and hierarchical Bayes. The pooled models, aggregate logit and then latent class, gain the most. Additionally, we show that the desired properties of differential enhancement and differential substitution require either individual-level data or the addition of attribute variability.

We propose that Randomized First Choice provides a general way to adjust conjoint shares to the marketplace. While the addition of product variability has been commonly used as a way to adjust errors in an experimental choice task, the addition of attribute variability provides a new way to account for heterogeneity that preserves the expected impact of item similarity on choice probabilities. The combination of both kinds of variability permits a great deal of flexibility in matching results from choice experiments to market conditions.

THE VALUE OF CHOICE SIMULATORS

One of the reasons conjoint analysis has been so popular as a management decision tool has been the availability of a choice simulator. These simulators often arrive in the form of a software or spreadsheet program accompanying the output of a conjoint study. These simulators enable managers to perform ‘what if’ questions about their market—estimating market shares under various assumptions about competition and their own offerings. As examples, simulators can predict the market share of a new offering; they can estimate the direct and cross elasticity of price changes within a market, or they can form the logical guide to strategic simulations that anticipate short- and long-term competitive responses (Green and Krieger 1988).

The choice simulators on which we focus have four stages. The first stage requires a preference model that can be applied to each individual or homogeneous segment in the survey. The second stage defines the characteristics of the competitors whose shares need to be estimated. The third stage applies the model to the competitive set to arrive at choice probabilities for each alternative and each segment or respondent. The final stage aggregates these probabilities across segments or individuals to predict choice shares for the market.

We pay the most attention to the third stage—estimating choice probabilities for each individual or segment. We explore the value of adjusting individual choice probabilities with two kinds of variability, each of which has a simple intuitive meaning. The first kind, product variability, occurs when a consumer simply chooses a different alternative on different choice occasions, typically through inconsistency in evaluating the alternatives. The second kind, attribute variability, occurs when a consumer is inconsistent in the relative weights or part worths applied to the attributes. For example, a consumer might notice the nutrition label on breads in one shopping trip, while being price or taste sensitive in other trips. While most simulators do not distinguish between these two forms of variability, we will show that they differ strongly in their treatment of similarity. Attribute variability preserves appropriate similarity relationships among alternatives while product variability clouds them. However, attribute variability by itself allows for no residual error in choice once the part worth values have been simulated. Thus, to appropriately model individual choice it is necessary to include both sources of variability.

We present Randomized First Choice as a general way to “tune” conjoint simulators to market behavior. Conceptually, Randomized First Choice begins with the assumption of no variability—the highest utility alternative in the set is chosen all the time. Then it adds back levels of attribute and alternative variablity that best match choice shares in the environment. This process allows sufficient flexibility to approximate quite complex market behavior.

Mathematically, Randomized First Choice adds variation in the attribute values in addition to variation in the final product valuation. It begins with a random utility model with variability components on both the coefficients and the residual error:

$$U_i = X_i (\beta + E_A) + E_P \qquad (1)$$

where:

U_i = Utility of product i for an individual or homogeneous segment at a moment in time

X_i = Row vector of attribute scores for alternative i

β = Vector of part worths

E_A = Variability added to the part worths (same for all alternatives)

E_P = Variability added to product *i* (unique for each alternative)

In the simulator, the probability of choosing alternative *i* in choice set S is the probability that its randomized utility is the greatest in the set, or:

$$\Pr(i|S) = \Pr(U_i \geq U_j \text{ all } j \varepsilon S). \qquad (2)$$

Equation 2 is estimated by using a simulator to draw U_i's from equation 1 and then simply enumerating the probabilities. To stabilize shares, group or individual choices are simulated numerous times.

Those familiar with logit will recognize that E_P is simply the error level in the logit model. The typical adjustment for scale in the logit model is mathematically equivalent to adjusting the variance of a Gumbel-distributed E_P in RFC simulations. The E_A term then reflects taste variation as has been found in models by Hausman and Wise (1978) and in current work in mixed logit by Revelt and Train (1998).

The purpose of this paper is to provide an understanding of why including attribute variability is superior to just including product variability. The quick answer is that attribute variability is needed to account for expected similarity relationships whereas adding product variability clouds those relationships. The next section begins by detailing the desirable properties of any choice simulator. Then follows an experiment that demonstrates the effectiveness of adding attribute and product variability, particularly when applied to aggregate and latent class segments, but also for individual choice models generated by hierarchical Bayes and Sawtooth's ICE (Individual Choice Estimation).

THREE CRITICAL PROPERTIES OF MARKET SIMULATORS

Market simulators need three properties if they are to reflect the complexity of market behavior. First, the individual- or segment-level model must display differential impact—where the impact of a marketing action occurs as an alternative in a competitive set reaches the threshold for choice. Second, the model needs to exhibit differential substitution, a property where new alternatives take disproportionate share from similar competitors. Finally, the simulator must display differential enhancement, the idea that very similar pairs can produce disproportionately severe choice probabilities. Each of these is detailed below.

Differential Impact is the first requirement of an effective choice simulator. It reflects the property that the impact of a marketing action depends on the extent that the alternative is near the purchase threshold. This point of maximum sensitivity occurs when the value of an alternative is close to that of the most valued alternatives in the set—when the customer is on the cusp with respect to choosing the company's offering. At that time, an incremental feature or benefit is most likely to win the business.

The differential impact implicit in a threshold model can best be understood by examining three cases reflecting different kinds of thresholds. First we present the linear probability model which importantly defines the case of no threshold. Then we examine the other extreme, that of a first choice model, which has the most extreme step-like threshold. Finally we consider the standard choice models (logit, probit) whose threshold has been softened by the addition of variability.

If probability were a linear function of utility, then improving an attribute would have the same effect on choice share regardless of how well it is liked. But such a simple approach could result in share estimates of less than zero or greater than one, so this approach has been rejected. There are other problems with a linear probability model, the worst of which is a lack of differential impact. Under a linear probability model adding, say, an internal fax modem has the same share impact regardless of whether it is added to a high- or low-end computer. By contrast, a threshold choice model would specify that the benefit from adding the modem mainly affects those consumers who are likely to change their behavior. This makes good sense—it does not affect a person who would already have bought it, nor does it affect customers who would never consider the brand. Managerially, the differential impact brought about by a threshold model has the benefit of focusing managerial attention on the critical marginal customer, and thereby avoids expensive actions that are unlikely to alter market behavior.

The first-choice model offers an extreme contrast to the linear model. The first choice model is mathematically equivalent to Equation 1 with no variability ($\text{var}(E_P) = \text{var}(E_A) = 0$). In the first choice simulation, share of an alternative is zero until its value is greater than others in the set. Once its value exceeds that threshold, however, it receives 100%. The problem with the first choice model is that it is patently false. We know that people do not make choices without variability. In studies of experimental choices, given the same choice set (3-4 alternatives, 4-5 attributes) respondents choose a different alternative about 20% of the time. In our study, respondents chose a different alternative in the repeated task 19% of the time. One of the paradoxes we hope to resolve in this paper is why the first choice model operating on individual-level part worths works so well despite its counter-factual premise.

Standard logit and probit models stand between the first-choice and linear model. Instead of the severe step function of the first choice model, the variablity implicit in these models has the impact of moderating the step into a smooth s-shape or sigmoid function. As shown in Equations 1 and 2, we view these models as first-choice models with variability added. For logit, E_P has a Gumbel, while for Probit, it has a Normal distribution. It is important to note, however, that these models are, to use a technical phrase, linear-in-the-parameters. Thus the *utility* of an item generally increases the same amount with a given improvement, however, the *probability of purchase* follows a threshold model.

A little-understood benefit of a threshold model is that it can reflect complex patterns of interactions between, say, a feature and a particular brand simply through the simulation process. An interaction term specifies that a particular feature has a differential impact on particular brands. While these interaction terms can be reflected in the utility function, we propose that many interactions can be better represented as arising from the aggregation of heterogeneous customers each following a threshold model. For example, consider a warranty x price interaction term that might be used to show that a warranty is more valuable for low- over high-priced appliances. The same effect could also emerge in a simulation of respondents under a threshold rule. Suppose there are two segments, one valuing low price and the other desiring high quality. Adding a warranty to the low-priced brand might not be sufficient to raise it past the purchase threshold of those desiring high quality. By contrast, the warranty pushes the alternative past the threshold of those desiring low prices. When these two segments are aggregated it appears that the warranty mainly helps the low priced brand and thus appears to justify an interaction term in the utility function. However, as this example illustrates, the same behavior can be reflected in a simulator with a threshold model. Further, the heterogeneity account is more managerial actionable than the curve-fitting exercise of the cross term.

The greatest difficulty with interaction terms is that their numbers can grow uncontrollably large. Above we illustrated an example of price tiers, but there can be many others. Consider combinations of brand tiers where customers are simply not interested in certain brands; size tiers where a large size never passes the threshold for certain segments, or feature tiers, where certain groups are only interested in certain features. Modeling these with interaction terms in the utility function is both complicated and can lead to problems with overfitting or misspecification. The beauty of a simulator operating on segmented or individual models is that it can approximate this behavior in the context of a simple main-effects additive model (e.g., see as Orme and Heft, this volume).

To summarize, differential impact is critical if we believe that impact on choice of, say, a new feature of a brand depends on values of the brands against which it competes. The threshold model within a random utility formulation focuses managerial attention on those alternatives that are on the cusp, and in that way places less emphasis on alternatives that are already chosen, or would never be. Further, applying the threshold model at the level of the individual or homogeneous segment confers the additional benefit of isolating the differential impact appropriately within each.

Differential Substitution is the second property critical to an effective choice simulator. Its intuition follows from the idea that a new offering takes share disproportionately from similar ones. Differential substitution is particularly important because the major choice model, aggregate logit displays *no* differential substitution. The logit assumption of proportionality implies that a new offering that gets, say, 20% of a market will take from each competitor in proportion to its initial share. Thus a brand with an initial 40% share loses 8 percentage points (40% x .2) and one with 10% share loses 2 percentage points (10% x .2). Proportionality provides a naive estimate of substitution effects and can result in managerially distorted projections where there are large differences in the degree of similarity among brands in the market. For example, a product line extension can be expected to take proportionately most share from its sibling brands. Managers recognize this problem. Successful companies manage their portfolios with new brands that are strategically designed to maximize share taken from competitors and minimize internal share losses. The important point is that proportionality glosses over such strategically important distinctions. In particular, ignoring differential substitution could lead to the managerial nightmare of numerous line extensions whose cost to current brands is regularly underestimated.

An extreme, if instructive, example of differential substitution is the presence of a duplicate offering in the choice set. Economic theory often posits that a duplicate offering should take half the share of its twin, but none from its competitor. However, in practice this expectation is rarely met. If some consumers randomly pick a brand without deleting duplicates, then having a duplicate could increase total choice share. Indeed, the fight for shelf space is directed at capturing that random choice in the marketplace. To the extent that a duplicate brand increases the total share for that brand we label the increase in total share from a duplicate *share inflation*. Clearly some share inflation is needed, but it is unclear how much. In the empirical test we measure the extent to which simulators reflect differential enhancement by how well they correctly predict the combined share of near substitutes in the holdout choice sets.

Differential enhancement is the third property needed by choice simulators. It specifies a second, but less commonly recognized way product similarity affects choices. Under differential enhancement, pairs of highly similar alternatives display more severe choice differences. Psychologically, this phenomenon derives from the idea that similar alternatives are often easier to compare than dissimilar ones. Consider the choice between French Roast coffee, Jamaican Blend coffee and English Breakfast tea. A change in the relative freshness of the coffees can be expected to enhance the relative share of the fresher coffee, while have relatively little impact on the proportion choosing tea.

In its extreme form, differential enhancement arises where one offering *dominates* another in the choice set. Rational economic theory typically posits that the dominated alternative receives no share, while the shares of the other brands are unaffected. Market behavior is not as neat. There are relatively few purely dominated alternatives in the market. Even finding two otherwise identical cans of peas in the supermarket can lead to suspicion that the lower priced one is older. Determining dominance requires work that consumers may be unwilling or unable to perform. For that reason, manufacturers intentionally create differences between offerings (new line, different price, channel), so that dominance, or near dominance is less apparent. From a modeling perspective, the important point is that any choice simulator needs to allow for dominance to produce cases of extreme probability differences and to allow consumers to be fallible in their ability to recognize that dominance.

The modeling implications of differential enhancement parallel those for differential substitution. The standard logit or probit models assume that the relative shares of any pair of alternatives only depend on their values, not on their relative similarity. Referring to a classic example, if trips to Paris and to London are equally valued, then a logit model predicts that adding a second trip to Paris with a one-dollar discount will result in one-third shares for the three alternatives. There are numerous ways researchers have attempted to solve this problem, from nested logit to correlated error terms within probit. Within the Sawtooth family Model 3 penalizes items that share attribute levels with other alternatives in the choice set. We will show that a simple first choice simulation with suitable variability added to both attributes and alternatives provides a robust way to mirror these complex market realities.

A Market Study to Validate Choice Simulators

As we approached the task of comparing the ability of different choice simulators to deal with varying degrees of alternative similarity, it became apparent that choice sets typically used for choice experiments would not work. For the sake of efficiency, most choice experiments feature alternatives where the numbers of levels differing among pairs of alternatives are relatively constant. For example, it would not typically make sense to include an alternative twice since its inclusion adds no additional information. In this study we deliberately add alternatives which are duplicates or near duplicates to be able to test the ability of various simulators to appropriately handle these difficult choices.

Three hundred ninety-eight respondents completed computerized surveys in a mall intercept conducted by Consumer Pulse, Inc. The survey involved preference for mid-sized televisions and was programmed using Sawtooth Software's Ci3 and CBC systems. Respondents over 18 who owned a television or were considering purchasing a mid-sized television set in the next 12 months qualified for the survey. The first part of the interview focused on attribute definitions (described in terms of benefits) for the six attributes included in the design. The main part of the survey involved 27 choices among televisions they might purchase. Each choice involved five televisions described with six attributes: brand name (3 levels), screen size (3 levels), picture-in-picture (available, not), channel blockout (available, not) and price (4 levels). Table 1 gives an example of a choice set that illustrates the levels. We gave respondents a $4.00 incentive to complete the survey, and urged them to respond carefully.

Preliminary data from a small pre-test suggested that respondents were not giving sufficient effort to answer consistently. In an attempt to improve the quality of the data, we revised the survey. We told them that the computer would "learn" from their previous answers and know if they were answering carefully or not. The "computer" would reward them with an extra $1.00 at the end of the survey if they had "taken their time and done their task well." (We displayed a password for them to tell the attendant.) In terms of programming the survey logic, we rewarded them based on a combination of elapsed time for a particular section of the survey and test-retest reliability for a repeated holdout task. Though it is difficult to prove (given the small sample size of the pretest), we believe the revision resulted in cleaner data. Nearly two-thirds of the 398 respondents received the extra dollar. We discarded 46 respondents based on response times to choice tasks that were unusually low, leaving 352 for analysis.

The first 18 choice tasks were CBC randomized choice sets that did not include a "None" option. After completing the CBC tasks, respondents were shown an additional nine holdout choice tasks, again including five alternatives. The holdout tasks were different in two respects. First, to test the market share predictions of the different simulators, it was critical to have target sets for which market shares could be estimated. Respondents were randomly divided into four groups with approximately 90 in each group that would receive the same nine holdout choice tasks. Additionally, we designed the holdout choices to have some extremely similar alternatives. Four of the five alternatives in the holdout tasks were carefully designed to have approximate utility and level balance (Huber and Zwerina 1996). However, the fifth alternative

duplicated another alternative in the set, or duplicated all attributes except the two judged least important in a pretest. To provide an estimate of test-retest reliability, two choice sets were perfect replicates. Across respondents, the computer randomized both choice set and product concept order.

The Contenders

We analyzed the CBC data using four base methods for estimating respondent part worth utilities: Aggregate Logit, Latent Class, Sawtooth's ICE (Individual Choice Estimation) and Hierarchical Bayes (courtesy of Neeraj Arora, Virginia Tech). There is logic behind picking these four methods. Aggregate logit is important in that it reflects what happens when all respondents are pooled into one choice model. By contrast, latent class analysis seeks sets of latent segments (we used an eight-group solution) whose part worths best reflect the heterogeneity underlying the choices (Kamakura and Russell 1989; Chintagunta, Jain and Vilcassim 1991; DeSarbo, Ramaswamy and Cohen 1995). ICE then takes these segments and builds a logit model that predicts each individual's choices as a function of these segments (Johnson 1997). It thereby is able to estimate a utility function for each person. Hierarchical Bayes assumes respondents are random draws from a distribution of part worth utilities with a specific mean and variance. It produces a posterior estimate of each individual's part worths reflecting the heterogeneous prior conditioned by the particular choices each individual makes (Lenk, DeSarbo, Green and Young 1996; Arora, Allenby and Ginter 1998). Both ICE and hierarchical Bayes reflect current attempts to generate each individual's utility functions from choice data, while latent class and aggregate logit typify ways to deal with markets as groups.

For each of these base models we examine the impact of adding three levels of variability within the Randomized First Choice framework. The initial condition is the first choice rule that assumes respondents choose the highest valued alternative in a choice set with certainty. The second condition adds the level of product variability that best predicts holdout choice shares. This latter condition is identical to adjusting the scale under the logit rule to best predict these shares. The third condition tunes both product and attribute variability to best predict the holdout choice shares. The mechanism of the tuning process is simple but tedious: we use a grid search of different levels of each type of variability until we find those that minimize the mean absolute error in predicting holdout choice shares.

Results

We examine the ability of different simulators to handle product similarity from different perspectives. First, we measure deviations from predicted and actual share for the duplicates and near-duplicates that were included in the holdout choice sets. This focus enables us to uncover ways the different models appropriately account for approximate differential substitution and differential enhancement. Then we broaden our perspective to consider the overall fit of the models—how well the models predict choice shares for all items in the choice set.

Differential substitution requires that similar items take disproportionate share from each other. Thus, our near and perfect substitutes should cannibalize share from each other. For example, if an alternative would receive 20% share individually, the joint share of the two alternatives should be only marginally more than 20%, since the new one takes most of its share from its twin. A first choice simulator, with its assumption of zero variability puts the joint share at exactly 20%, but in the marketplace this combined share is likely to be somewhat higher. Put differently, due to fundamental noise in the consumer choice processes we can expect some share inflation.

Table 2 gives predicted combined share of the near and perfect substitutes divided by the actual share. Thus, a value of 100% means that the degree of differential substitution reflected in the holdout choices was estimated perfectly. (The first two rows of the first column are not computed, since the first choice rule is generally never applied to aggregate logit or latent class results.) Notice that the first choice rule underestimates the joint share of the near substitutes by about 20%, indicating that the first choice rule of no variability is too severe. The next column shows the result of adding the level of product variability that best predicts the holdouts. In this case, adding that variability seriously overestimates the share inflation for the near substitutes, in effect, assuming too much variability. The third column then adjusts both product and attribute variability to optimally predict choice shares. By allowing some attribute variability to substitute for product variability, we are able to more closely track actual differential substitution in this data set for all models except ICE.

It is also instructive to compare the rows representing the four core models. The two aggregate models, logit and latent class, suffer most from overestimation of share inflation for similar products under product variability. However, when both forms of variability are combined, they do remarkably well. The two individual models appear both less sensitive to the addition of variation and less in need of it. We will discuss the implications of this phenomenon after the other results from the study are presented.

Differential enhancement occurs when a given quality difference results in a greater share difference between highly similar pairs. We examine the share difference between the alternative with higher expected share and its near duplicate. Table 3 gives the model's prediction of this difference as a percent of the actual difference. Once again a score of 100% indicates perfect level of differential enhancement relative to the actual choices.

The two aggregate models with product variability only are least able to accurately represent the differential enhancement reflected in the holdout choices. The first choice rule applied to the individual level models performs very well in this case. That false assumption appears to cause little harm when applied to ICE and hierarchical Bayes. In all cases, adding the optimal level of product variability tends to understate desired levels of differential enhancement. Optimizing both kinds of variability has a small incremental benefit but results in predictions that still underestimate the appropriate level of differential enhancement.

It needs to be emphasized that these measures of differential substitution and enhancement only relate to the shares of near substitutes. By contrast, the optimization to choice shares counts all five alternatives, not just the two most similar ones. The overestimation of differential substitution shown in the last column of Table 2 and the underestimation of differential enhancement in the last column of Table 3 could have been improved by decreasing the level of product variability, but overall fit would have suffered. An interesting implication of this result is that

the actual variability around judgments relating to the share sums and share differences of these near substitutes may be smaller than for alternatives generally. An interesting path for future research involves allowing variability to change as a function of similarity of an alternative within each set.

Relative error measures the degree that the different simulators predict the market shares across all alternatives in the holdout tasks for the study. Our variance measure is the mean absolute error (MAE) predicting the market shares for the holdout stimuli. The test-retest MAE of the identical choice sets was 0.035, indicating that the shares between these identical holdout choice sets differed by 3.5 percentage points, on average.

Table 4 shows MAE as a percent of this test-retest MAE. For example, the 151% for aggregate logit indicates that adding product variability only results in an error that is about one and one-half times as great as for the choice replication. Adding attribute variability helps all the models, but the greatest gains occur for the aggregate models.

Table 4 offers several surprises. The first surprise is that Randomized First Choice applied to latent class does as well as any of the models. The positive impact of both kinds of variability on latent class makes sense because the original latent class model assumes that there is no heterogeneity within each latent class. By optimizing both product and attribute variability we are able to transform latent class from an elegant but counterfactual model into one that tracks choice shares remarkably well.

The second surprise is that the addition of attribute variability has very little impact on either of the individual level models. For both hierarchical Bayes and ICE the addition of product variability is the major benefit. There is a simple reason for this result. The individual level models are not estimated with perfect accuracy, but have significant variation due to the noise in individual choices and the fact that many parameters are being estimated from few observations. Thus, when estimates from these models are put in a simulator they act as if variability has been added randomly to the part worths. However, in this case instead of attribute variability coming from the RFC process, it comes from the inherent variability in the estimation model. This insight then leads to an important conclusion: where variability in the estimation technique is greater than in the market, then the optimal variability to add to the first choice model will be zero (see also Elrod and Kumar, 1989).

The final surprise is that Randomized First Choice predictions are quite good regardless of the core estimation method used (except aggregate logit). That is, using RFC produces accuracy that is within 10% of what one would get asking the same question again. Clearly few techniques are going to do much better than that. There simply is not much room for further improvement.

Before concluding, it is important to mention Sawtooth's Model 3, a long-available method for accounting for item similarity in a simulation. Model 3 operates by penalizing alternatives that have high numbers of levels in common with other attributes in a choice set. It does so in such a way that adding a perfect duplicate perfectly splits share with its twin when these duplicates share no levels in common with the other alternatives. Model 3 acts like the first choice model in underestimating the joint share of the two identical alternatives for the holdout choices in our study. (Some share inflation is justified to model actual choices). Model 3 reflects a simple (and inflexible) rule regarding differential substitution. It also does not address differential enhancement. Thus, on both fronts, Model 3 is not a theoretically complete model of similarity effects. It doesn't surprise us that for our study Model 3 was consistently outperformed by RFC on all dimensions. Thus, in our view, Sawtooth users should replace Model 3 with RFC.

SUMMARY AND CONCLUSIONS

The purpose of this paper has been to examine ways to build choice simulators that correctly reflect similarity effects. We began with the introduction of three principles needed for sound conjoint simulations, and in the light of those principles developed Randomized First Choice. RFC provides better choice share predictions by determining the optimal levels of attribute and product variability when generating simulated choices.

The first requirement of effective simulators is that they reflect differential impact. This property permits the simulator to focus managerial attention on those actions that are most likely to impact their customers. In addition, a little-known implication of the threshold model at the level of a segmented (such as by lclass) or individual model is that it automatically allows for various kinds of price and offering tiers without the necessity of interaction terms. The value of that benefit is best seen in the poor effectiveness of the aggregate logit simulation, even with variability added. In simple, main-effects aggregate logit, there is no way the threshold effect can display the action of different segments. Either the homogeneous segments from latent class or individual models are necessary for that benefit.

Effective simulators also need to reflect differential substitution. Our analysis of the combined share of near and perfect substitutes indicates that the first choice model underestimates, while adding product variablity overestimates their combined share. The joint optimizations of both product and attribute variability then permit the estimates of combined share to closely approximate the actual choices. One can tune the appropriate balance of differential substitution/share inflation.

The third requirement of effective simulators is that they demonstrate differential enhancement. We illustrated this requirement by examining the share difference of nearly identical alternatives. The first choice rule overestimates differential enhancement in aggregate models by giving all share to the preferred alternative. By contrast, adding product variability underestimates the predicted share differences. Adjusting both kinds of variability improved this underestimation but did not solve it completely. Since differential enhancement comes in part from a psychological mechanism whereby decisions between similar alternatives are easier, a full solution to this problem may await models that adjust item variability to the difficulty in making the choice.

We demonstrated the benefits of RFC on a study in which the holdout choices included "difficult" alternatives that included near and true duplicates. However, a greater benefit for Sawtooth users may come in contexts where it is possible to project to actual market shares. Most markets will have far more complicated similarity structures than our simple problem, resulting from competition among family brands, different sizes, price tiers and subbrands. We believe that RFC with its two kinds of variability will be very useful in tuning the simulator to successfully account for market behavior in such cases.

TABLE 1
Example of a Holdout Choice Set

25” JVC, Stereo, Picture in Picture, No Blockout, $350	26” RCA, Surround Sound, Picture in Picture, Blockout, $400	25” JVC, Monaural, No Picture in Picture, No Blockout $300
27” Sony, Surround Sound, No Picture in Picture, No Blockout $450	25” JVC, Stereo, Picture in Picture, No Blockout, $350	

TABLE 2
Differential Substitution:
Predicted Combined Share of Near Substitutes
As Percent of Actual Share

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	139%	108%
Latent Class	N/A	119%	105%
Hierarchical Bayes	91%	117%	104%
ICE	89%	101%	94%

TABLE 3
Differential Enhancement:
Predicted Difference between Similar Alternatives
As Percent of Actual Differences

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	63%	73%
Latent Class	N/A	71%	74%
Hierarchical Bayes	100%	73%	77%
ICE	90%	77%	79%

TABLE 4
Relative Error:
Mean Absolute Error Predicting Market Share
As Percent of Test-Retest

	First Choice Rule	Product Variability	+Attribute Variability
Aggregate Logit	N/A	151%	112%
Latent Class	N/A	117%	105%
Hierarchical Bayes	125%	110%	107%
ICE	112%	106%	106%

REFERENCES

Arora, Neeraj; Greg Allenby, and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science, 17*, 29-44.

Chintagunta, Pradeep; Dipak C. Jain and Naufel J. Vilcassim (1991), "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data," *Journal of Marketing Research, 28,* 417-428.

DeSarbo, Wayne S., Venkat Ramaswamy and Steve H. Cohen (1995) "Market Segmentation with Choice-Based Conjoint Analysis," *Marketing Letters, 6,*137-148.

Elrod, Terry and S. Krishna Kumar (1989), "Bias in the First Choice Rule for Predicting Share," *Sawtooth Software Conference Proceedings*.

Green, Paul E and Abba M. Krieger (1988) "Choice Rules and Sensitivity Analysis in Conjoint Simulators," *Journal of the Academy of Marketing Science 16.1*, 114-127.

Hausman and Wise (1978), "A Conditional Probit Model for Quantitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 43, 403-426.

Huber, Joel and Klaus Zwerina (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research, 23*, 307-317.

Johnson, Richard M. (1997), "ICE: Individual Choice Estimation," Sawtooth Software Technical Paper.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research, 26,* 339-390.

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, 15.2,* 173-191.

Orme, Bryan K. and Michael Heft (1999), "Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results," *Sawtooth Software Conference Proceedings*.

Revelt, David and David Train (1998), "Mixed Logit with Repeated Choices: Household's Choices of Appliance Efficiency Level," *Review of Economics and Statistics.* (Forthcoming).

Rossi, Peter and Greg Allenby (1993), "A Bayesian Approach to Estimating Household Parameters," *Journal of Marketing Research, 30,* p. 171-182.

COMMENT ON HUBER, ORME & MILLER

Richard M. Johnson
Sawtooth Software, Inc.

Several papers from previous Sawtooth Software conferences have turned out to be genuinely important, but I think this paper may be the most important yet. These authors should be congratulated for an idea that seems likely to revolutionize the way we do conjoint simulations. I'd like to place RFC in the context of simulation methods offered in the past by Sawtooth Software.

The "first choice" model assumes that a respondent will choose the product that has highest utility. It has several advantages: it's easy to compute, easy to explain, and doesn't give too much share to product families with many similar entries. Unfortunately, though, it does a poor job of predicting shares in the market place. It's usually too extreme, typically over-predicting large shares and under-predicting smaller ones. And there's no good way to tune it.

The logit or "share of preference" model is more flexible. To predict each respondent's share of preference you add up his/her part worths to get the utility for each product, exponentiate those utilities, and then percentage the results. This model has a valuable property —it can be tuned. By scaling individual utilities you can make it as extreme as the first choice model, or as flat as you like. Unfortunately, the share of preference model does over-predict shares for product families containing multiple entries.

The "share of preference model with correction for product similarity" tries to overcome those problems by decreasing estimated shares for similar products. It does this in a somewhat arbitrary way, but until recently has been almost the only game in town.

It seems likely that all of these approaches should be replaced by RFC. The RFC model assumes that an individual's part worths vary randomly around their mean. The simulation consists of drawing many sets of part-worths from each individual's assumed distribution, adding them up to get product utilities, adding additional "product variability" if that is indicated, and then using a first choice model for the resulting utility sums. This is done over and over for each individual, and the results of many independent draws are accumulated to get that individual's share of preference.

Since it is a first choice model, RFC does not suffer from IIA problems. And, unlike an ordinary first choice model, it can be tuned by adjusting two constants: the size of the attribute variance and the size of the product variance.

This is an exciting development, and I thank the authors for an impressive paper.

A Comparison of Alternative Solutions to the Number-of-Levels Effect

Dick R. Wittink
Yale University
P. B. Seetharaman[1]
Washington University

Abstract

We define the number-of-levels (NOL) effect, and show its magnitude in full-profile data. In a separate experimental study we determine its existence in choice data, (and in ACA). We show the effect on derived attribute importances and on predicted choice shares. We then test two competing explanations for the effect in ACA data. We find that an algorithmic solution, in which the number of levels is customized based on self-explicated importances, obtains superior aggregate-level predictions. Individual-level holdout choices are best predicted when the number of levels is the same for all attributes. We also obtain statistical support for each solution within a given experimental group by relating covariates to the (incremental) predictive power of the preference intensity judgments in ACA.

Introduction

In conjoint analysis, the number-of-levels (NOL) effect is that the distance between the largest and smallest part worths of an attribute is a positive function of the number of interior levels, holding other things constant. This effect is usually examined with derived attribute importances, and it has been shown to exist for virtually all data collection and analysis methods. Currim et al. (1981) stumbled upon this systematic effect in a non-experimental study. Wittink et al. (1982) verified its existence experimentally. In the experiment, an artificial effect for both full profile and tradeoff matrix rank order data, analyzed by metric (least squares) and nonmetric (LINMAP) methods, was obtained.

These results might simply mean that rank order preferences suffer from inherent weaknesses. However, Reibstein et al. (1989) obtained approximately the same magnitude of a NOL effect for ratings as for ranks. Steenkamp and Wittink (1994) found the NOL effect to be about the same for ratings and magnitude estimation data. One possible interpretation of these results is that the ratings, which are the hallmark of much survey research, do not have better properties than rank order data. In that case, most survey results are based on data with much weaker properties than is commonly thought. The question then is whether conjoint study designs can be adapted to minimize the consequences of ratings having properties that are closer to ordinal- than interval- scale measurement.

[1] Dick R. Wittink is the General George Rogers Clark Professor of Management and Marketing at the Yale School of Management (dick.wittink@yale.edu); P. B. Seetharaman is Assistant Professor of Marketing at the Olin School of Business, Washington University, St. Louis (seethu@mail.olin.wustl.edu). We thank Richard Johnson for many useful discussions and helpful comments on an earlier version of this paper.

An alternative perspective on the NOL effect is that it is due to a behavioral or psychological sensitivity. That is, respondents may pay more attention to attributes with a greater number of levels. For example, as the number of levels for an attribute increases, objects will differ more often on that attribute, and this higher frequency of differences may draw heightened attention. Green and Srinivasan (1990) offer such an explanation, and Johnson (1991) has obtained a result that seems consistent with a behavioral interpretation of the NOL effect. However, it is worth pointing out that the NOL effect in rank order data occurs independent of a possible psychological sensitivity.

In this paper, we first show the magnitude of the NOL effect and contrast it with the effect due to an increase in the range of attribute variation, based on full-profile ratings. We then discuss the experimental design for a personal computer study in which both ACA and CBC data were collected. We mention the NOL effect for those data in terms of predicted shares in holdout choices. The holdout data are also used to compare the predictive accuracy of three conjoint designs. We find that both a customized ACA design (which solves the NOL effect if the explanation is “algorithmic”) and a design with equal numbers of levels (which solves the problem if the source of the NOL effect is “psychological”) outperform a design in which the attributes differ in the number of levels (the differences being the same for all respondents).

Specifically, we obtain the best predictions for the “customized” version on an aggregate predictive measure. However, the “equal number of levels” version is best on a disaggregate predictive measure. When we focus on the gain in disaggregate predictive power due to the preference intensity judgments in ACA, we also find that deviations from equal numbers of levels in the customized version reduce this gain considerably. Similarly, we find a reduction in the gain in predictive accuracy in the “equal number of levels” version that is attributable to deviations from the levels that would have been assigned in the customized design. This reduction is larger in the former case, suggesting that the NOL effect may be reduced, if not eliminated, by employing equal numbers of levels across the attributes. Thus, our results suggest that conjoint studies may benefit from the use of equal numbers of levels across the attributes, where feasible. However, we cannot claim that it is necessarily the ‘psychological’ explanation which is responsible for this result. For example, it is possible that statistical arguments favor the use of equal numbers of levels for the maximization of percent hits in holdout choices, everything else being equal. In addition, the preference intensity responses may have an excessive amount of noise in the customized version due to the constraint that all paired objects have equal predicted utilities (based on the self-explicated data).

The Magnitude of a NOL Effect

To illustrate how large the effect is, we show results (Schifferstein et al. 1998) from a full-profile experiment. In this experiment, the two manipulations are: (1) the number of attribute levels, either 2 or 4; (2) the range of attribute variation differing by a ratio of 3 to 1. Respondents judged the attractiveness of color TV’s on a 9-point rating scale. We show both the manipulations and average derived attribute importances in Table 1.

Table 1
DERIVED IMPORTANCES OF COLOR TV ATTRIBUTES

	Experimental Cell					
	A		B		C	
Attribute						
* Warranty (number of years)	0.5 1 3 5	**0.21**	0.5 5	**0.13**	0.5 1 3 5	**0.20**
* Price (dollars)	325 550	**0.17**	325 400 475 550	**0.26**	400 475	**0.08**
* Screen size (inches)	20 24	**0.07**	20 24	**0.07**	16 28	**0.14**
Subtotal Derived Attribute Importances		**0.45**		**0.47**		**0.42**

The experimental design has three cells, and respondents were randomly allocated to one of these cells. We show the attribute levels only for three attributes manipulated experimentally. For the attribute ‘Warranty’, cells A and C have 4 levels, and cell B has 2 levels, while the range of variation is held constant. For ‘Price’ cell B has 4 levels, and cell A has 2 levels, holding the range of variation constant. Cell C also has 2 levels but these levels differ by only $75 versus $225 in cell A, corresponding to a 1:3 ratio. Finally, for ‘Screen size’, the range of variation is 12 inches in cell C but 4 inches in cells A and B, all with two levels, again according to a 3:1 ratio.

Starting with ‘Screen size’, we see that both cells A and B have an average derived importance of 0.07. This importance equals 0.14 in cell C. Thus, an increase in the range by a factor of three, increases the importance by 0.07, holding the number of levels constant. For ‘Price’, a similar difference in the range by a factor of three produces a difference in importance of 0.09 (cell A has 0.17, cell C has 0.08). Although it is impossible to say what is a correct change in importance for this manipulation, it is clear that we should see a dramatic change in an attribute’s importance when the range of variation is expanded by a factor of three.

By contrast, the NOL effect should be zero. That is, for conjoint results to have external validity, we would expect that the derived attribute importances do not depend on the number of interior levels. Yet we find that for ‘Price’ the difference in importance is 0.09 between cells A and B. Also, for ‘Warranty’ it is 0.07 (C versus B) or 0.08 (A versus B).

The results in Table 1 suggest that, on average, the full-profile ratings produce derived attribute importances that are about equally sensitive to a 2:1 increase in the number of levels, holding the range constant, as a 3:1 increase in the range, holding the number of levels constant! Thus, the NOL effect magnitude is extremely large in full-profile data. We note that Wittink et al. (1991) compared the magnitude of the effect for full-profile and ACA data, and found the NOL effect to be about twice as large in full profile as in ACA.

It is interesting to compare the results for ‘Price’ across the three experimental cells. Starting with cell C, we conclude that a difference of $75 represents 0.08 in relative importance. We see that the result in cell B is quite consistent with this representation: in cell B a difference of $225 produces a relative importance of 0.26. Thus, an increase in the range by a factor of 3 ($225/$75) corresponds to an increase in relative importance of about 3 (0.26/0.08), when the number of levels is also increased from two to four. Cell A, on the other hand, generates for the same increase in the range, compared with cell C, a ratio of relative importances equal to about 2 (0.17/0.08). Thus, an expansion of the range when the number of levels is held constant produces inconsistent results.

TWO COMPETING THEORIES

In a separate study, we use ACA version 3.0 (and CBC) to: (1) determine the NOL effect for choice data (CBC) and compare it with the effect for ACA; (2) demonstrate that it affects predicted choices; and (3) compare the predictive accuracy of ACA for three designs. One design (VARIED) has variation in the number of levels across attributes as follows: 2 x 2 levels, 1 x 3 level, 2 x 4 levels. A second design (EQUAL) has the same number of levels for all attributes: 5 x 3 levels. The third design (CUSTOM) customizes the number of levels based on each respondent’s self-explicated importances. If the NOL effect is primarily “psychological” then the EQUAL design should have the highest predictive accuracy. If it is primarily “algorithmic” (see Wittink et al. 1997 for an explanation), then the CUSTOM design should be best. Either way, the VARIED design is expected to be the worst, because its results may suffer from either or both of the possible causes of the NOL effect.

The design for the experimental study of personal computer preferences is reported in Wittink et al. (1997). Briefly, each attribute has a maximum of 5 levels defined as follows:

Attributes	**Levels**
Brand name	Compaq; Dell; Gateway; IBM; Packard Bell
Speed	200; 175; 166; 150; 133 mhz
Hard drive	2.0; 1.8; 1.6; 1.4; 1.2 GB
RAM	64; 48; 32; 24; 16 MB
Price	$1,200; $1,400; $1,600; $1,800; $2,000

Only for 'Brand name' is each respondent's preference order for the levels elicited. Now imagine that for each attribute the levels are ordered according to preference from 1 through 5. Customization in the CUSTOM design then occurs as follows:

Self-explicated importance	**Number of levels**	**Specific levels**
4	5	1, 2, 3, 4, 5
3	4	1, 2, 4, 5
2	3	1, 3, 5
1	2	1, 5

In EQUAL, levels 1, 3 and 5 are used for all five attributes. For 'Brand name' the specific brands depend on each respondent's stated preference order. For 'Price' the levels are always $1200, $1600 and $2000, etc. In VARIED, the same procedure is used for 'Brand name' as in EQUAL. For 'Speed' and 'RAM' levels 1 and 5 are used in VARIED, whereas for 'Hard drive' and 'Price', levels 1, 2, 4 and 5 are used.

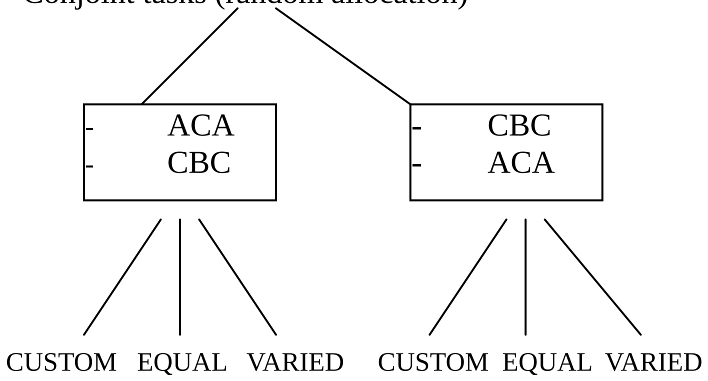
We show the experimental design in Figure 1. All respondents provided self-explicated data and judged 5 holdout choice sets before they completed two conjoint tasks, either ACA followed by CBC, or the same in reverse order. Following the conjoint tasks, all respondents completed 10 holdout choice sets (5 replicates to determine holdout choice consistency), and provided personal data.

Figure 1
Experimental Design

- Self-explicated data
  (preference order for brand names, importances for all attributes)

- 5 Holdout choice sets

- Conjoint tasks (random allocation)



- 10 Holdout choice sets (5 replicates)

- Personal data (PC ownership, PC attributes if known)

Results

The NOL effect on derived attribute importances is slightly greater for CBC than for ACA. Based on individually calculated part worths, the average derived importances across respondents are on average (across attributes) 0.05 greater for an attribute with more levels in ACA, and 0.07 greater in CBC (comparing EQUAL with VARIED). If we use average part worths for each experimental cell instead, the average NOL effect across four attributes is 0.04 for ACA and 0.06 for CBC. See Wittink et al. (1997) for more detail.

In the holdout sets all objects are defined in terms of extreme levels only (since these levels always occurred for every respondent in each design). Predicted shares differ systematically between EQUAL and VARIED, consistent with the NOL effect on derived importances. In ACA the average difference is 0.11 (11 absolute percentage points), while it is 0.14 for CBC. Importantly, in four out of ten holdout choice sets the majority of respondents is predicted to choose one object in EQUAL but the other object in VARIED, based on ACA (see Wittink et al. 1997, Table 4).

PREDICTIVE ACCURACY

To examine how well the ACA results predict holdout choices we use an aggregate- and an individual- level measure. At the aggregate level, we compare the predicted shares with actual shares, and use the differences as indicators of prediction error. We note that the use of shares is attractive for several reasons. For example, in many commercial applications conjoint results are used especially to predict preference shares for a new or modified product under various scenarios.

From an analytical perspective, an aggregate measure is attractive because it captures primarily the bias in predictions. We know that the NOL effect causes biases in part worths, and the issue we should focus on is which experimental cell, EQUAL or CUSTOM, has less bias. Nevertheless, we also use percent hits as an individual-level measure of predictive accuracy. Because the percent hits does not allow a prediction error for one respondent to be compensated for by an opposite prediction error for another respondent, this measure will reflect both variance (unreliability) and bias in part worths.

For both aggregate- and individual- level measures we want to take into account two other aspects. Specifically, the respondents who were randomly allocated to the three versions may differ in the consistency with which they make holdout choices (5 replicate sets). Consistency is the percent of times a respondent chooses the same object in the replicated holdout choice tasks. Everything else being equal, the greater this consistency the greater the maximum possible predictive accuracy (see Wittink and Johnson, 1997). The other relevant aspect is that the three versions may have different predictive accuracies from the self-explicated data. Thus, we desire measures that determine the increase in accuracy (reduction in error in predicted share) from the self-explicated data to the final ACA solution (i.e., due to the preference intensity judgments from which the NOL effect emanates), relative to the maximum possible increase.

We define:

Y_A = error in share from self-explicated data minus error in share in holdout choices

X_A = error in share from self-explicated data minus error in share from final ACA

% Gain $_A$ = percent gain in accuracy, due to the preference intensity judgments, at the aggregate level ($100 * X_A/Y_A$)

On this measure, we obtain the following percentage gains:

	CUSTOM	**EQUAL**	**VARIED**
% Gain_A →	71.6%	64.6%	42.2%

Thus, the CUSTOM version provides the strongest gain relative to the maximum possible. It captures almost 3/4 of the possible gain. The EQUAL version is second, and covers almost 2/3 of the maximum possible. VARIED does worst, and improves less than 1/2 relative to the upper bound.

As mass customization of products and services increases in popularity, individual-level predictive accuracy becomes more relevant. We use a measure that, similar to the one above, determines the improvement in percent hits from self-explicated data to final ACA relative to the maximum possible improvement (which we define as the difference between the percent agreement in the five replicated holdout sets, corrected for attenuation, and the percent hits for the self-explicated data), i.e.:

Y_I = attenuated percent agreement in replicated holdouts minus percent hits for self-explicated data

X_I = percent hits for final ACA minus percent hits for self-explicated data

%Gain_I = percent gain in accuracy, due to the preference intensity judgments, at the <u>individual</u> level ($100 * X_I/Y_I$)

Wittink and Johnson (1997) show why and how the maximum possible percent hits is greater than the percent agreement in replicated holdouts. If p_{cc} is the percent agreement in replicated choices, and Π_c is the probability that an observed choice is the "true" choice, then agreement occurs if the observed choices either both agree or both disagree with the true choice:

$$p_{cc} = \Pi_c^2 + (1 - \Pi_c)^2$$

It follows that Π_c can be estimated:

$$\hat{\Pi}_c = \frac{1 + \sqrt{2p_{cc} - 1}}{2} \qquad \text{if } p_{cc} \geq 0.5$$

For completeness we provide the details on the components of Y_I and X_I, including the attenuated percent agreement from the formula for $\hat{\Pi}_c$:

Table 2
COMPONENTS OF THE INDIVIDUAL-LEVEL PERCENT GAIN MEASURE

	Experimental Group		
	CUSTOM	**EQUAL**	**VARIED**
percent hits final ACA	63.9	63.6	58.8
percent hits self-explicated data	50.7	49.0	46.1
percent agreement in replicated holdouts (p_{cc})	62.8	62.1	62.1
attenuated percent agreement ($\hat{\Pi}_c$)	75.3	74.6	74.6

By using the appropriate component values, we obtain the following percentage gains:

	CUSTOM	**EQUAL**	**VARIED**
% Gain $_I$ →	53.7%	57.0%	44.6%

These individual-level results show that the EQUAL version is best, followed by CUSTOM. As at the aggregate level, VARIED clearly is worst. The reversal of relative performance for EQUAL and CUSTOM suggests that EQUAL has lower variance (and CUSTOM has smaller-bias). We discuss briefly why this might occur.

To consider differences in variance in individual-level part worths between EQUAL and CUSTOM, we need to mention the principle underlying the customized ACA version. According to the algorithmic explanation, the NOL effect occurs because of an imbalance in predicted utilities for the paired objects. By setting the number of levels for each attribute equal to its self-explicated importance (on a 4-point scale) plus one, we can create all paired objects to have perfectly equal predicted utilities (from the self-explicated data). Wittink et al. (1997) argue that this condition makes it impossible for the NOL effect to occur in ACA (assuming no 'psychological' or other reasons apply). But the creation of utility balance in the paired objects has the consequence that respondents may often be truly indifferent between the paired objects (the more accurate the initial ACA solution, the more likely this is). As a result, the preference intensity judgments may have more noise in CUSTOM than in EQUAL, especially if respondents do not want to express "indifference" or "equal preference" all the time. And, the task may discourage respondents in the sense that the preference intensity judgment is a difficult one in CUSTOM, if the true utilities of paired objects are also very close. This would further increase the variance of the judgments and hence the variance of the part worths in CUSTOM.

Finally, we pursue an individual- level test of the experimental results. To do this, we focus on the percent hits for each respondent in the CUSTOM and EQUAL versions. As before, we examine the improvement in hits from the self-explicated data. However, since the consistency in replicated holdout choices can be zero, we do not use the percent gain relative to the maximum possible. Instead, we use consistency as a covariate.

The argument is that the respondents in CUSTOM differ in the extent to which the number of levels varies across the attributes. For example, if a respondent gave all attributes a self-explicated importance of 3, then all attributes had 4 levels for that respondent. Being assigned to the CUSTOM version, the respondent will not be subject to the algorithmic source of a NOL effect. And the psychological source is necessarily absent as well due to the equal self-explicated importances. However, a respondent in CUSTOM who provided widely different self-explicated importances for the five attributes would be exposed to different numbers of levels. Assuming that this customization does not neutralize the psychological effect (since “more important” attributes have more levels), then the improvement in percent hits will be less for respondents with much variation in the number of levels. Thus, we propose that variation in the number of levels for respondents in CUSTOM decreases the incremental predictive accuracy of the preference judgments, with the decrease being a function of the amount of variation (or the deviation from equal numbers of levels).

Similarly, consider a respondent in the EQUAL version who also happens to have equal self-explicated importances for all attributes. For this respondent there is no opportunity for the paired objects to produce distortions in the part worths. That is, this respondent would also have had equal numbers of levels in the CUSTOM version. By contrast, a person with widely different self-explicated importances would have had very different numbers of levels in CUSTOM. Given the use of equal numbers of levels for all respondents in EQUAL, we cannot predict the direction of distortions. But we expect the greater a respondent’s deviation from equal self-explicated importances, the greater the distortion in the final ACA part worths will be.

For the respondents in CUSTOM we estimate the following equation:

$$\mathrm{CHANGE}_i = f(\mathrm{DEV}_i, \mathrm{CON}_i)$$

where:

CHANGE_i is the change in percent hits from the self-explicated data to the final ACA solution for respondent i;

DEV_i is the deviation from equal numbers of levels for respondents;

CON_i is the consistency or percent agreement in the five replicated holdout choices for respondent i.

The estimated equation is: $0.16 - 0.14\ \mathrm{DEV} + 0.09\ \mathrm{CON}$

$$(-4.82) \qquad (2.39) \quad t - \text{ratios}$$

Thus, in the CUSTOM version there is a significant reduction in the gain due to the preference intensity judgments as the deviation from equal numbers increases. The maximum gain is 0.25, i.e. if DEV = 0, and CON = 1, the predicted change in hits equals 0.25. Note also that for a respondent whose holdout choices are perfectly consistent (CON = 1), the gain in accuracy is only 0.02 if the (average) deviation from equality is 1 unit. On average, this gain is 0.13 (see Table 2 where it can be found to be 13.2).

In a similar vein, we examine for EQUAL how deviations from the numbers of levels a respondent would have seen in the customized version influence the change in hits between final ACA and self-explicated data:

$$CHANGE_i = g(DEV_i, CON_i)$$

where:

$CHANGE_i$ is the change in percent hits from the self-explicated data to the final ACA solution for respondent i;

DEV_i is the deviation from the number of levels that would have been customized for individual i;

CON_i is the consistency or percent agreement in the five replicated holdout choices for respondent i.

For EQUAL the estimated equation is: 0.17 – 0.08 DEV + 0.04 CON

(-2.85) (1.08) t-ratios

In this case the effect of deviations appears to be not as strong as it is for CUSTOM. Still, if DEV = 0 and REL = 1, the maximum predicted increase in hits is 0.21, versus on average a gain of 0.15 (see Table 2 where it is 14.6). And the significance of DEV is consistent with the idea that there is an algorithmic explanation for the NOL effect. Thus, the incremental predictive accuracy for ACA due to the preference judgments at the individual level goes down as the deviation from the optimal customized numbers of levels increases (if DEV = 1, and CON = 1, the predicted gain is 0.13). Importantly, while we can imagine alternative explanations for the effect of deviations from equal numbers of levels, we are unable to conjure any for the effect of DEV in the latter equation.

CONCLUSIONS

The NOL effect is large. Specifically this artificial effect, when the number of levels is increased by a factor of 2, while the range of attribute variation is held constant, is slightly greater than the effect of an increase in the range of variation by a factor of 3, holding the number of levels constant. The NOL effect is also pervasive. It occurs in preference ranks and ratings, and in metric- and nonmetric estimation methods. In addition, choice data are sensitive to the effect, and CBC shows a larger effect than ACA does. However, the effect in ACA is much smaller than it is in full profile. And the effect in ACA 4.0 is much smaller than it is in ACA 3.0 (see Orme 1998). Importantly, we used ACA 3.0.

We have demonstrated that predicted holdout choice shares are systematically affected, to such an extent that the NOL effect in ACA reverses which object is predicted to be the most frequently chosen alternative in 4 out of 10 holdout sets. We compare two alternative solutions to the NOL effect. One is based on a psychological explanation of the effect which mandates the use of equal numbers of levels across the attributes. The other is based on an algorithmic explanation which requires that the numbers of levels be customized in ACA based on each respondent's self-explicated importances.

We find that for both aggregate- and individual- level predictive accuracy measures, a version with differences in the numbers of levels (the same differences for all respondents in VARIED) is the worst. On the aggregate measure, the customized version is somewhat better than the equal levels version. On the individual measure the equal levels version is slightly better than the customized version. Thus, which solution is best depends on the accuracy measure. Our results do not provide unequivocal support for one explanation of the phenomenon over the other. The reasons for this ambiguity are the following. In the EQUAL version there is the potential for distortions (if paired objects have unequal predicted utilities), but it is impossible to say in which direction the part worths are affected. Thus, there is no discernable bias even though the part worths become noisier. And in the CUSTOM version any psychological effect may be decreased since the number of levels is based on each respondent's self-explicated importances. Nevertheless we can take comfort that either of the two solutions is likely to improve the substantive value of conjoint results.

REFERENCES

Currim, Imran S., Charles B. Weinberg, and Dick R. Wittink (1981), "Design of Subscription Programs for a Performing Arts Series," *Journal of Consumer Research*, 8 (June), 67-75.

Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54 (October), 3-19.

Johnson, Richard M. (1991), "Comment on "Attribute Level Effects Revisited…," *Advanced Research Techniques Forum, Proceedings of Second Conference*, Rene Mora (ed.), Chicago AMA, 62-4.

Orme, Bryan (1998), "Reducing the Number-of-Levels Effect in ACA with Optimal Weighting," Working Paper, Sawtooth Software.

Schifferstein, Hendrik N.J., Peter W. J. Verlegh and Dick R. Wittink (1998), "Range and Number-of-Levels Effects in Derived and Stated Attribute Importances," Working Paper.

Steenkamp, Jan-Benedict E.M. and Dick R. Wittink (1994), "The Metric Quality of Full-Profile Judgments and the Number-of-Attribute-Levels Effect in Conjoint Analysis," *International Journal of Research in Marketing*, 11 (June), 275-86.

Wittink, Dick R. and Richard M. Johnson (1997), "Estimating the Agreement between Choices Among Discrete Object and Conjoint-Ratings-Based Predictions After Correcting for Attenuation," Working Paper.

Wittink, Dick R., Lakshman Krishnamurthi, and Julia B. Nutter (1982), "Comparing Derived Importance Weights Across Attributes," *Journal of Consumer Research*, 8 (March), 471-4.

Wittink, Dick R., Lakshman Krishnamurthi, and David J. Reibstein (1989), "The Effect of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters*, 1 (Number 2), 113-23.

Wittink, Dick R., William G. McLauchlan and P. B. Seetharaman (1997), "Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis," *Sawtooth Software Conference Proceedings*, 227-240.

Wittink, Dick R., Joel Huber, John A. Fiedler, and Richard L. Miller (1991), "Attribute Level Effects in Conjoint Revisited: ACA versus Full Profile," *Advanced Research Techniques Forum, Proceedings of Second Conference*, Rene Mora (ed.) Chicago: AMA, 51-61.

COMMENT ON WITTINK & SEETHARAMAN

Joel Huber
Duke University

We owe a debt of gratitude for Dick Wittink and P. B. Seetharaman for their paper, and to Dick McCullough for his earlier paper. Both papers join a guerilla campaign that for the last 15 years has sought to demonstrate that the number-of-levels effect is both large and pervasive.

Still, we hope these authors will forgive us if we do not express our gratitude too fervently. The problem is that the number-of-levels effect is a message we do not want to hear. We do not want to hear that adding an intervening level can change the results of a conjoint analysis. We may be intrigued by the different analytic and psychological explanations, but we would all prefer our cherished measurement techniques not depend on what appear to be trivial design differences.

There are two common reactions to such unwelcome news. Given we cannot shoot the messenger, we may react by criticizing the message. Such criticism is pretty easy to do in this case, particularly for those with academic inclinations. We could criticize the method by which the analysis is done, the respondents on which the tests were performed, or the relevance of product category. The idea is to distance the number-of-levels effect from any work we might do.

I must go on record saying that such sophistic attack on the number-of-levels effect is just that—sophistry. In simple terms, there have been too many studies demonstrating the effect with different methodologies, product classes and kinds of respondents to lead one to believe that the distortion is localized to particular methods or decision contexts. The current study has the most devastating finding: adding 50% more interior levels has the same impact on attribute importance as tripling the range. The effect is neither localized nor minor. Like global warming, it is pervasive. The number-of-levels effect won't go away.

Given we cannot argue it away, the second reaction to unwanted news is much more simple—denial. Denial takes many shapes. We can sleep during the presentation, ignore the paper, or simply do not feature it among the key lessons that we bring back to share with our fellow workers.

Denial, while perhaps useful to our emotional health, is dishonest both to our clients and ourselves. Further, it is not needed. In the short time remaining, I would like to put the number-of-levels effect into a broader context. I will try to persuade you that we need to ride rather than fear the elephant in the room. The elephant in this case reflects the many ways that conjoint design affects its outcome. Consider as examples:

- A range effect whereby the impact of a subrange such as 5 miles per gallon has less impact as the total range of miles per gallon increases from 10 to 20.
- A task complexity effect whereby adding attributes or increasing their detail can lead to fewer attributes being considered as respondents try to cope with the task.

- An attribute order effect whereby attributes at the beginning and the end of a list have greater weight than those lost in the middle.

In simple terms, we as a field need to acknowledge that the results from conjoint depend importantly on how we define the exercise. Rather than bemoan the fact that our results depend on design parameters, we need to establish ways that the design parameters are themselves not arbitrary. Let me suggest the following three guidelines:

- Conjoint studies should mimic market behavior. For example, if price is considered last, put it there. Find out which attributes are used and how they are described. If many levels are used, include those in the conjoint design.
- Be intentional when adding attributes or levels that do not reflect current market behavior. It is fine to add novel attributes so long as it is understood that conjoint simulates what would happen if these attributes are introduced and sufficiently promoted.
- Develop standard methods for setting design parameters for conjoint studies. Then the outcome for a given research provider is not arbitrary but reflects an explicit philosophy of survey design. Over time experience with different designs develops skill in how to interpret the outcomes of the studies.

For a well-run market research provider following these guidelines, the number-of-levels effect is good, not bad news. If conjoint were as immutable as some of its protagonists claim, then anyone could do it. It is precisely because the design of a conjoint study takes wisdom, experience and thought that we researchers have added value in the process.

USING LISREL AND PLS TO MEASURE CUSTOMER SATISFACTION

Lynd D. Bacon, Ph.D.
Lynd Bacon & Associates

There's a consensus that customer satisfaction is important. There have been numerous books, articles, and conferences devoted to it and its determinants. Firms have invested huge amounts in collecting satisfaction data, analyzing them, and reporting on the results. Satisfaction has been argued to be a critical component of brand equity (e.g. Aaker, 1991). Firms and managers perceive it to be related to customer retention, which is related to profitability. Governments have funded national studies to monitor it. Although other constructs have been proposed to be more important correlates of loyalty (e.g. customer value, see Gale, 1994), it is definitely the case that customer satisfaction is a central construct in the relationship between firms and customers.

Researchers and managers have used a variety of definitions for customer satisfaction. Most refer to a psychological process involving prior knowledge, beliefs, or expectations; perceived performance; and the evaluation of this information, or an affective response to it. Oliver (1997, p. 13) provides the following general definition:

> *Satisfaction* is the consumer's fulfillment response. It is a judgment that a product or service feature, or the product or service itself, provided (or is providing) a *pleasurable* level of consumption-related fulfillment, including levels of under- or over overfulfillment.

It's important to note that the fulfillment response Oliver refers to is *internal* to the customer.

SATISFACTION IS UNOBSERVABLE

Unlike physical quantities like temperature or weight, the extent of a customer's satisfaction must be inferred. It is assessed by what Torgerson (1958) called *measurement by fiat.* Since we can't measure it directly, we instead measure other variables that are observable. These are sometimes called indicator, or manifest, variables. Based on a priori grounds, or perhaps on more sophisticated procedures, we ascribe meaning to what we observe based on the presumed relationship between satisfaction and our indicator variables.

Unobserved, or latent, variables are very common in marketing research. They include real income, socioeconomic status, perceived quality, utility, brand attitude, purchase intention, and loyalty. To measure them we rely on observable indicators that are (hopefully) correlated with them.

It's important not to treat indicators and the latent variables that may underlie them as being the same thing. We should observe the Chinese proverb (Bagozzi, 1994a):

"Do not confuse the finger pointing at the moon with the moon."

Now it's probably unlikely that many managers or researchers truly believe that the numeric responses elicited by satisfaction measurement scales really *are* satisfaction per se. Yet the practice of reporting and modeling the observed responses while ignoring any errors in measurement that may be present, can mislead. One reason is that estimates of the correlation between satisfaction and other variables will be biased. Some form of correlational analysis is often used to do what's called "key driver analysis" (Oliver, 1997) or "revealed preference analysis" (Hauser, 1991) in satisfaction research, so the effects of measurement error have practical implications.

EFFECTS OF UNRELIABILITY

The reliability of a measure is the extent to which it provides consistent results from one application to the next, or the degree to which it is free of random error (Vogt, 1993, p. 195). When a measure's unreliability is not taken into account, estimates of its correlation with other variables will be biased, and differences in the measure across groups or over time may be obscured.

An instance of this problem that is particularly relevant to the analysis of satisfaction data has to do with the effects of unmodeled measurement error in variables used to predict satisfaction. A common form of key driver analysis consists of deriving attribute importance by regressing a satisfaction measure on a set of predictor variables that consist of ratings of perceived performance. Unmodeled measurement error in the predictors will cause bias in the estimated regression coefficients, even when the expected value for the errors is equal to zero.

This problem is well known in econometrics, and is referred to as the "errors in variables" problem (Maddala 1977). It turns out that if only one of the predictors has measurement error, there is downward bias in the coefficient for the less-than-completely-reliable predictor. The coefficients of the *other* coefficients are also biased either upwards or downwards, but the direction can be calculated if the reliability of the predictor with error can be estimated. Thus, the effect of having only one unreliable predictor is that *all* coefficients are biased, even those for perfectly reliable predictors. When more than one predictor is unreliable, it is possible to approximate the extent and direction of the bias in each coefficient, but doing so is rather cumbersome (Maddala, 1977, p. 294), and an estimate of each variable’s reliability is required.

In sum, unmodeled measurement error results in biased estimates. This is even though the expected value of the measurement error may be zero. A simple simulated example of the effects of unmeasured can be found in Bacon (1997a,b). See also Rigdon (1994).

MODELING LATENT VARIABLES

One way of dealing with measurement error is to use measurement models that separate error from what you want to estimate. A measurement model is a device for connecting observed, or indicator, variables to one or more latent variables (LVs) such that "true" values on the latter can be separated from error. A familiar example is the classic psychometric measurement model:

$$Y_{observed} = Y_{true} + \zeta$$

Where $Y_{observed}$ is the observed value on a continuous variable for a single observation, Y_{true} is the true value on the unobserved continuous variable (i.e. an LV) for that observation, and ζ is a measurement error that is also unobserved. It's often assumed that $E[\zeta] = 0$ and $\mathrm{cov}(Y_{true}, \zeta) = 0$. A definition for the reliability of $Y_{observed}$ is the ratio of the variances of Y_{true} and $Y_{observed}$.[1]

Latent variable models are models that include measurement models for LVs. They may also estimate relationships between the LVs. These relationships can include multiple criterion, or endogenous, variables, as well as multiple predictor, or exogenous, variables, and are expressed as multiple equations.

In LV models, the relationships between variables are often represented as a set of directed or undirected "paths." The paths to each variable describe that variable's dependencies. Each directed path represents an equation, and a picture of the paths is called a "path diagram." Path diagrams provide a concise way of describing complex models. Figures 1 and 2 provide examples of path diagrams. These will be discussed below.

There exists a variety of LV models, and two of the ways they differ is in terms of whether their observed and latent variables are continuous or discrete. Examples of LV models with discrete variables include latent class models with nominal observed and LVs, and item response theory (IRT) models with nominal or ordinal observed variables and interval or ordinal LVs. Heinen (1996) provides a summary of LV models for discrete latent or discrete observed variables. Hagenaars (1993) describes latent class path analysis models based on the work of Goodman, Haberman, and others. Vermunt describes models for event history analysis (1997) and path analysis models (1996) for data with missing values.

Another type of LV model describes the variances and covariances of observed variables, or their correlations, and these are the focus of this paper. The predominant form is called a covariance structure model (CSM), or a structural equation model (SEM).[2] The LVs in them are continuous. The observed variables are typically treated as interval-level continuous variables, although the CSM approach has been extended to accommodate nominal and ordinal observed, or "indicator," variables (Muthén 1984, 1987). The most commonly used software for fitting CSM models is the LISREL (linear structural relations) program developed by Jöreskog (1973). The LISREL specification extended maximum likelihood (ML) factor analysis by combining it with path analysis. General specifications of CSMs combine measurement models for LVs, and a path model for relationships between LVs.

At about the same time that Jöreskog was developing his ML procedures, Wold (1973, 1980; see also Jöreskog and Wold 1982) described a set of procedures for estimating path models with LVs that used methods based on ordinary least squares (OLS). This approach has come to be

[1] The unique error in the classic measurement model, ζ , may have more than one source (e.g. Bollen, 1989). In some applications this may result in unreliability not being properly accounted for, which will in turn result in biased estimators (DeShon, 1998)

[2] Some authors (e.g. Bollen, 1989) call CSMs SEMs. I use CSM here because both CSMs and PLS can be used to describe SEMs.

known as the PLS approach for structural equation modeling. It is distinct from the PLS regression method commonly used in chemometrics to develop predictive models using intercorrelated inputs.[3] It is also distinct from CSM methods in several respects. For simplicity I'll call it PLS here. Bear in mind that both kinds of models can be used to estimate parameters for a system of equations in LVs.

In the following two sections I describe some of the basic concepts of the CSM and PLS approaches to structural equation modeling, and provide a simple example of each.

CSM AND PLS

Both CSM and PLS offer unique features for modeling customer satisfaction. They provide measurement models for continuous LVs, and the ability to estimate models comprised of systems of equations. As we've indicated, measurement models provide a means of taking unreliability into account. Being able to model systems of equations is important when mediating variables are present, when endogenous variables predict other endogenous variables, or when there is "reciprocal causality" between variables. Path diagrams illustrating these features are in Figures 1 and 2.

In Figure 1, all variables are observed, or "manifest:" there are no measurement models. There are regression residuals, however. This model has three equations for the three dependent variables labeled **purc_intent**, **satisfied**, and **sup_brand**. It is a non-recursive since the variables **sup_brand** and **satisfied** predict each other. Figure 2 shows a recursive model with LVs perceived **value**, perceived **support** quality, and perceived **hardware** quality. Each of LV is modeled using observed indicator variables. These models are each like a confirmatory factor model with one factor. Combined they are often referred to as the measurement model of a CSM.

CSM or PLS can be used to estimate models like those in Figures 1 and 2. They can both accommodate measurement models, and both can estimate systems of equations. In the case of both, the modeler must specify the form of the model. The estimation software does not decide what paths to put in, or what indicator variables to use for different LVs. The basic equations and other specifications for CSM and for PLS are given in Table 1.

PLS was developed to estimate recursive models, like the one in Figure 2. It wasn't developed for non-recursive models. Non-recursive models have one or more loops in them, like the loop between **sup_brand** and **satisfied** in Figure 1. Both PLS and CSM were developed for estimating linear relationships, and so estimating interactions between or nonlinearities in LVs has been problematic. There have been some recent progress on this issue for CSMs (Schumacker and Marcoulides, 1998; Arminger & Muthén, 1998) and also PLS (Chin, Marcolin & Newsted, 1996).

CSM and PLS differ in many ways. The differences are due to what the two methods were designed for, and the kinds of estimation procedures they use. Table 2 lists a number of differences that are relevant to modeling satisfaction. Here we will elaborate on some of the more important of them.

[3] See Frank & Friedman (1993) for a critical comparison of PLS regression to other methods, including ridge methods, and Bacon (1997) for a simple example of how PLS regression algorithms work.

The CSM approach emphasizes estimating and testing model parameters. It was developed out of a tradition of modeling as a way of developing and evaluating theories. PLS, on the other hand, was developed to maximize predictive accuracy (Jöreskog and Wold, 1982; Wold, 1982), while providing flexibility for exploratory modeling. PLS doesn't require the distributional assumptions that CSM does, and hence has been called "soft modeling" (Wold, 1980, 1982). It was originally viewed as a complement for the LISREL approach to fitting SEMs (Ibid.).

Fitting a CSM involves minimizing the difference between observed and predicted variance-covariance (VCV) matrices of the observed variables. An iterative fitting algorithm is used to simultaneously estimate all parameters. The algorithm starts from a set of initial values for the parameters being estimated, and then adjusts them over successive iterations until a scalar measure of the discrepancy between observed and predicted has been minimized. The most widely used procedure provides maximum likelihood (ML) estimates for parameters. It requires that the observed data are distributed as multivariate normal, and that the observations be independent. Some other estimation methods are less restrictive. Browne's (1982) asymptotically distribution-free (ADF) fitting criterion, for example, only requires that the observed data be continuous.

In general, CSMs will be unidentified, so some of their parameters must be constrained. A model is unidentified when unique solutions cannot be obtained for all of its parameters. In terms of CSMs this may mean that there are more parameters to be estimated than the number of unique elements of the VCV matrix. Or it may mean something less obvious. Heuristics or empirical tests are typically used to determine whether a CSM is identified. The constraints serve to reduce the number of parameters to be estimated.

LVs in CSMs are estimated basically like common factors in confirmatory factor analysis are. Constraints must be used to set their measurement scales. The scores on CSM LVs are not estimated directly. They may be estimated after fitting a SEM by using on of several different multiple regression approaches, but the values obtained depend on the method used. Therefore, they should be used with some care.

PLS estimation involves estimating the parameters of a model by iterating over a sequence of parts of the model with the goal of minimizing the residual variance associated with all endogenous variables in the model. The model parts consist of measurement models for the LVs, which are called "blocks," and a set of relations that connect the LVs. These are called the "outer" and "inner" models, respectively, and they are analogous to the measurement and structural models of CSMs. The data analyzed with PLS is typically a correlation matrix. The algorithms used are different kinds of OLS procedures, depending on what kind of parameter is being estimated.

As is the case with CSMs, observed variables in PLS models can be related to LVs in one of two ways. They may be reflective indicators, meaning that in a path diagram an arrow points to them from an LV, as in Figure 2, for example. These are like the effects indicators in CSMs. The other kind of indicator variable in PLS models is a formative indicator. These have an arrow pointing *towards* a LV, and are like causal indicators in CSMs.[4]

The LVs in PLS are more like principal components than they are like the factor-like LVs in CSM. The are estimated as exact weighted linear combinations of observed variables. Parame-

[4] Causal indicators in SEM can be difficult to use. See McCallum & Browne (1993).

ter identification is not a problem when using PLS, and assumptions about how the observed variables are distributed are not needed. PLS does not require that the residuals for different observations be distributed the same way, or that observations be independent.

PARAMETER ESTIMATION AND INFERENCE

An important difference between CSM and PLS is in terms of their parameter estimates. The maximum likelihood estimates provided by CSM are just that: they are M.V.U.E. (minimum variance unbiased estimates) under certain conditions of regularity, assuming that the model and its assumptions is correct, and given that the sample is large. ML estimates have several desirable characteristics, and they are relatively robust against violations of assumptions.

PLS estimates of the scores on LVs are "consistent at large." That is, they become consistent as the sample size and the number of indicators per LV both become large. Under conditions of finite sample size and number of indicators, the lack of complete consistency in the scores produces biased estimates of component loadings, which relate reflective indicators to PLS LVs, and in path coefficients. There is no closed-form solution for estimating the size of the bias in PLS estimators (Lohmöller, 1989; Chin, 1998a).

Lohmöller (1989) has worked out solutions for cases involving one or two LVs and given equal correlations between indicators. They express the bias in PLS estimates as the ratio of each PLS estimate to its corresponding ML estimate. This is only useful if the model used to obtain the ML estimates is in fact correct to begin with. Chin (1998a) suggests that in general the effects of inconsistency will be that loading estimates will tend to be biased upwards, and path coefficients will tend to be biased downward. It is theoretically possible that biases in PLS estimates will make different coefficients appear to be the same, and coefficients that are truly the same appear different. As Chin (1998b) suggests, there is a need for research to more broadly understand the implications of the consistency at large assumption in PLS.

Ryan & Rayner (1998) provided a good example of bias in PLS estimates. They compared PLS and CSM results across simulated data sets. They used non-recursive models to generate their data, varying sample size and the values of model parameters. Their models had four exogenous LVs, and one endogenous LV. Their results indicate that the parameter estimates produced by PLS were on average further from the true values than those from CSM models, while the root mean squared error for the PLS models was on average smaller.

The flexibility of using PLS comes from not having to make assumptions about how variables are distributed. This provides the freedom to use any kind of indicator variables. Part of the price for this is that direct statistical tests are not available. PLS provides no statistical tests of parameter significance, of model fit, or of differences between models. Inference is possible by using jackknife or bootstrapping procedures, however.

CSMs, on the other hand, provide estimates of standard errors for estimated parameters, various tests of model fit, statistical comparisons of nested models, and the ability to test very general linear and nonlinear constraints on parameter values. These require distribution assumptions, of course. When the distributional assumptions are not tenable, inferences about parameters can still be entertained by bootstrapping.

CSMs can be fit to data from multiple groups simultaneously so that differences between models for the groups can be tested. They can be used to estimate means and intercepts for LVs. Recent work on modeling multi-level (hierarchical, or clustered) data with CSMs includes estimating individual-level coefficients (e.g. Kaplan & Elliott, 1997; McArdle, 1998; Muthén, 1994). Muthén's (1987) LISCOMP specification provides a means of estimating CSMs with discrete observed variables. His recent work further extends the LISCOMP framework to accommodate finite mixture models (Muthén, 1998).

A SIMPLE EXAMPLE

To illustrate the use of CSM and PLS models we'll consider a very simple example: a bivariate linear regression with LVs. The data are from a survey of 200 customers of an automotive service organization. They include ratings on four performance attributes, and four satisfaction ratings. The four performance attributes were courtesy, accessibility, speed with which routine service is completed, and cleanliness of the service location. A seven point (or six interval), unipolar rating scale was used for each one.

The satisfaction measures included a bipolar scale with end points of very dissatisfied and very satisfied, and an agreement/ disagreement scale for the statement "Overall I am very satisfied with my visit to WeLubeM.[5]" Each of these two scales had eight points. The third measure was a six-point rating of satisfaction relative to other automotive service providers that ranged from must less satisfied to much more satisfied. The fourth satisfaction measure was a judgment of confidence that service would be satisfactory on the next visit. The arcsin transform of this last variable was used for the following analyses.

ML estimation was used to fit CSM models to these data since some preliminary analysis indicated that the observed variables didn't depart too far from multinormality. The simplest useful model that could be specified involved a single LV underlying the four performance measures, and a single satisfaction dimension. This model was estimated using a widely available CSM program (AMOS 3.61). The data consisted of the variance-covariance matrix of the eight observed variables. Judging from the fit indices obtained, and from the estimated residuals, the fit of this model to the variance-covariance matrix is adequate. The ML χ^2 for the model is 23.09 with 19 degrees of freedom. A commonly used fit index, the root mean square error of approximation (RSMEA), is 0.064. A heuristic for RSMEA is that values between 0.05 and 0.08 indicate moderate fits. The residuals look approximately normally distributed based on a normal quantile-quantile plot. There are a number of other fit measures for CSMs.

Standardized estimates of the model's coefficients are shown on the path diagram in Figure 3. In this diagram, the observed performance indicators are labeled **p1** through **p4**, and the satisfaction indicators **s1** through **s4**. When this model was estimated, the measurement scales for the two LVs **performance** and **satisfaction** were set by constraining the factor loadings for p1 and s1 to be equal to 1.0. What this does is to make the unit of measurement for each LV the same as the unit for the observed variable with the fixed loading. An analogous contraint was used to fix the measurement scale of zeta, the regression residual.

[5] Not the real name of the business.

You'll note from the path diagram that the estimated correlation between the LVs is 0.85, and that the standardized loadings vary from 0.51 to 0.80. All are reliably different from zero based on their estimated standard errors. Each standardized loading can be interpreted as the square root of the reliability coefficient for the observed variable it is associated with. So, for example, only about 26% of the variation in p4 is associated with the performance LV, assuming that the model is correct.

Figure 4 shows PLS coefficients obtained by using Lohmöller's LVPLS program. Here again, the summary statistics indicate a reasonably good fit. The average $R^2 = 0.58$. The root-mean-square covariance (RMS COV) between the residuals of the latent and manifest variables is 0.04. Like RMSEA, smaller values of RMS COV are better. Figure 4 also shows the standardized CSM estimates from the model in Figure 3.

You can see from Figure 4 that the point estimates of the CSM loadings are each smaller than their PLS counterparts. Thus in this example PLS makes it look like the LVs are somewhat better measured by their indicators than the CSM results suggest. Another difference between results is that the point estimate of the path coefficient between the two LVs is smaller in the PLS model (0.68) than in the CSM model (0.85). Both kinds of difference are commonly observed (Chin, 1994) when CSMs and PLS models are compared. In this case it's not clear which set of estimates is closer to the true parameter values. The same pattern of differences can emerge due to bias in the PLS estimates (Ibid.).

USING CSM AND PLS FOR SATISFACTION MEASUREMENT

The differences between the two methods indicate the circumstances under which one might be preferred over the other. If you need to estimate non-recursive models, then CSM is the method of choice. If you are in particular interested in accurately predicting LV scores, then you may want to consider PLS, since it provides a direct method for doing so.

If your primary interest is in comparing path coefficients or factor loadings, CSM could be the better choice, given either that the necessary assumptions are tenable or that it's feasible to obtain bootstrap estimates. PLS estimates of loadings and coefficients will always be biased. The important question for any particular application of PLS is, by how much?

Some other circumstances under which you might plan on using CSM include:

- You want to do traditional statistical inference;
- You have theory or strong domain knowledge about the causal relationships you want to quantify, and it's important to compare models;
- Construct validity is of high importance;
- You want to test whether models for different groups are the same.

A comparative strength of PLS is its ability to accommodate variables regardless of their type of measurement. This feature of PLS can make it more convenient to use on satisfaction data that have already been collected, or that will be collected without consideration of what would be must suitable for the purposes of fitting CSMs.

In the experience of this author, satisfaction data collected without regard to plausible models or scale construction are unlikely to provide useful CSM results. As Rigdon (1998) has pointed out, there can be a substantial risk of failure in using such a demanding analytical method when its use was not originally considered. Regardless of recent progress towards making CSMs more flexible, they are still a confirmatory method, and don't lend themselves well to ad hoc use. In the case of using either CSM or PLS, much judgment is required. They are both relatively complicated modeling procedures, and there is often little consensus amongst experts on how they are best applied to particular non-trivial problems.

Neither CSM or PLS will have great utility if the data they are used on do not include measures of important variables, or if models are specified in a grossly incorrect manner. Oliver (1997) notes that there seems to be a general disregard of the role and importance of process variables in most customer satisfaction research, and points to the common practice of only collecting data about perceived performance on product attributes. It's clear that ignoring such variables can be very misleading. For example, there is ample empirical evidence that the disconfirmation of expectations either partially or fully mediates[6] the effects of perceived performance on satisfaction in many circumstances (Ibid.). Therefore, omitting this variable from either type of model has the potential to produce parameter estimates that don't fully reflect the true impact of variations in perceived performance. It's obvious that theory and prior knowledge must be used to guide both data collection and model specification.

Both CSM and PLS can accommodate multicollinearity[7] in predictors of satisfaction. Multicollinearity is often a problem for satisfaction researchers who want to do key driver analyses, since the variables they would like to use to explain variations in satisfaction often evidence moderate to severe amounts of it. The basic approach when using either method is to model multicollinearity in some fashion. Using either method, you can use an LV to represent a set of highly interrelated observed variables, assuming it makes substantive sense to do so. Using CSMs you can also estimate covariances between LVs or between the uniquenesses of indicator variables, or you can use generic "method" factors. The method used should depend on what makes the most sense given what you believe your data measure. It certainly should not be based just on what makes a model fit the data better.

When using PLS to model multicollinear variables you can summarize interdependent predictors with one or more LVs by using the predictors as reflective indicators. Using them as formative indicators can prove problematic, as there is no way of taking the interdependence between the variables into account, and the result can be instability in the estimates obtained. More generally, standard PLS does not provide ways of modeling undirected correlation, i.e. association between variables that is not assumed to have a direction.[8] Such relationships are assumed to not exist when using PLS.

[6] A mediating variable is one that partially or fully intervenes in the path from one variable to another (Baron & Kenny, 1986). In Figure 1, **satisfied** *fully* mediates the effect of **perform** on **purc_intent,** since there is no direct path from **perform** to **purc_intent.**

[7] Multicollinearity is linear dependency between two or more variables. Substantial multicollinearity makes it difficult to separate the predictive effects of variables, and results in unstable estimates. Maddala (1977) attributes the term to Ragnar Frisch (1934).

[8] It's also the case that some of the bias PLS's OLS estimates can be due to covariation between LVs and errors in equations that can't be modeled in standard PLS (Wothke, 1998).

An important consideration for research practitioners is the availability of software and background literature. There are at least five commercially available software packages for fitting CSMs, and there are a few more in the public domain. The available PLS software consists mainly of a public domain version of a program written by Lohmöller in the 1980's, although a more contemporary application may soon become available (Chin, 1998b). The research literature on CSMs is large and growing rapidly, providing a rich knowledge base for using these models. Literature about PLS is scarce, and there have been few developments in it over the last decade. The Appendix provides sources for software and other resources for using CSM and PLS.

In sum, CSM and PLS offer distinct advantages as methods for analyzing satisfaction data. They both provide a means of recognizing measurement error and controlling its effects on the estimation of other quantities. Both methods can estimate systems of relationships, which provides a way of expressing the multiple dependencies noted in the research literature on customer satisfaction. Each provides a means of modeling multicollinearity so that its deleterious effects on estimation can be reduced. They also differ from each other in important ways. These include the kinds of distributional assumptions needed in order to use them, the degree of bias in the estimates they provide, and the ease with which inferences about models and parameters can be made.

Both methods can be demanding to use, and each requires specific knowledge and experience on the part of the analyst. PLS was developed for applications in which little theory is available and predictive accuracy is of paramount importance. CSM, on the other hand, was developed for theory-driven modeling. In keeping with the philosophy that no model is correct, it may be useful (or at least instructive) to apply both procedures when possible so that discrepancies between their results can be examined, and perhaps even be reconciled.

REFERENCES

Aaker, D. (1991), *Managing Brand Equity*, New York: MacMillian Publishers.

Arminger, G. and B. Muthén (1998), "A Bayesian Approach to Nonlinear Latent Variable Models Using the Gibbs Sampler," *Psychometrika*, 63, 271-300.

Bacon, L. (lynd.bacon@lba.com) (1997a), "1997 ARTF SEM Tutorial Web Page," http://www.lba.com/art97.html, June.

------ (1997b), "Introduction to Structural Equation Modeling," Tutorial, AMA Advanced Research Techniques Forum, Monterey CA, June.

Bagozzi, R. (1994a), "Panel Discussion on Teaching CSMs," American Management Association Research Division Conference on Causal Modeling, Purdue University, March.

------ (Editor) (1994b), *Advanced Methods of Marketing Research*, Malden MA, Blackwell.

Baron, R. and D. Kenny (1986), "The Moderator-Mediator Variable Distinction in Social Psychological Research," *Journal of Personality and Social Psychology*, 51, 1173-1182.

Bollen, K. and S. Long (1993), *Testing Structural Equation Models*, Thousand Oaks CA: Sage.

Browne, M. (1982), "Covariance Structures," in *Topics in Applied Multivariate Analysis*, ed. D. Hawkins, Cambridge: Cambridge University Press, 72-141.

Chin, W., B. Marcolin and P. Newsted (1996), "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Study and Voice Mail Emotion/Adoption Study," in *Proceedings of the Seventeenth International Conference on Information Systems*, J. Degross, S. Jarvenpaa and A. Srinivasan, editors. Cleveland OH, December.

Chin, Wynne (1998a), "The Partial Least Squares Approach for Structural Equation Modeling," in *Modern Methods for Business Research*, G. Marcoulides, editor. Mahwah NJ: Erlbaum, 275-337.

------ (1998b), Personal Communication, Impending availability of the PLS-Graph software program, December.

DeShon, R. (1998), "A Cautionary Note on Measurement Error Corrections in Structural Equation Models," *Psychological Methods*, 3, 412-423.

Dillon, W., J. White, V. Rao and D. Filak (1997), "Good Science," *Marketing Research: A Magazine of Management and Applications*, 9, 22-31.

Falk, F. and N. Miller (1992), *A Primer for Soft Modeling*, Akron OH: University of Akron Press.

Frank, I. and J. Friedman (1993), "A Statistical View of Some Chemometrics Regression Tools (With Discussion)," *Technometrics*, 35, 119-148.

Frisch, R. (1934), *Statistical Confluence Analysis by Means of Complete Regression Systems*, Olso, Norway: University Institute of Economics.

Gale, B. (1994), *Managing Customer Value*, New York: Free Press.

Hagganaars, J. (1993), *Loglinear Models with Latent Variables*, Quantitative Applications in the Social Sciences, vol. 94, Thousand Oaks CA: Sage.

Hauser, J. (1991), *Comparison of Importance Measurement Methodologies and Their Relationship to Consumer Satisfaction*, Cambridge MA: Sloan School of Management, MIT.

Heinen, T. (1996), *Latent Class and Discrete Latent Trait Models*, Advanced Quantitative Techniques in the Social Sciences, Thousand Oaks CA: Sage.

Jöreskog, K. (1973), "A General Method for Estimating a Linear Structural Equation System," in *Structural Equation Models in the Social Sciences,* A. Goldberger and O. Duncan, editors. New York: Academic Press, 85-112.

------ and H. Wold (1982), "The ML and PLS Techniques for Modeling with Latent Variables," in *Systems Under Indirect Observation (I),* K. Joreskog and H. Wold, editors. Amsterdam: North-Holland.

Kaplan, D. and P. Elliott (1997), "A Didactic Example of Multilevel Structural Equation Modeling Applicable to the Study of Organizations," *Structural Equation Modeling,* 4, 1-24.

Kenny, D. and C. Judd (1984), "Estimating the Nonlinear and Interactive Effects of Latent Variables," *Psychological Bulletin,* 96, 201-210.

Lohmöller, J. (1989), *Latent Variable Path Modeling with Partial Least Squares*, Heidelberg: Physica-Verlag.

MacCallum, R. and M. Browne (1993), "The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues," *Psychological Bulletin,* 114(3), Nov, 533-541.

Maddala, G.S. (1977), *Econometrics,* New York: McGraw-Hill.

Marcoulides, G. (Editor) (1998), *Modern Methods for Business Research,* Quantitative Methods Series, Mahwah NJ: Erlbaum.

------ and R. Schumacker (1996), *Advanced Structural Equation Modeling: Issues and Techniques,* Mahwah NJ: Erlbaum.

McArdle, J. (1998), "Modeling Longitudinal Data by Latent Growth Curve Methods," in *Modern Methods for Business Research*, G. Marcoulides, editor. Mahwah NJ: Erlbaum, 359-406.

Muthén, B. (1984), "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators," *Psychometrika*, 49, 115-132.

------ (1987), *LISCOMP. Analysis of Linear Structural Equations Using a Comprehensive Measurement Model*, Chicago IL: Scientific Software International.

------ (1994), "Multilevel Covariance Structure Analysis," *Sociological Methods & Research*, 22, 376-398.

------ (1998), Personal Communication, Extensions of the LISCOMP method in recently available MPlus software, November.

Oliver, R. (1997), *Satisfaction: A Behavioral Perspective on the Consumer,* New York: McGraw-Hill.

------ and W. Desarbo (1988), "Response Determinants in Satisfaction Judgements," *Journal of Consumer Research*, 14, 495-507.

Rigdon, E. (1994), "Demonstrating the Effects of Unmodeled Measurement Error," *Structural Equation Modeling*, 1, 375-380.

------ (1998), "Structural Equation Modeling," in *Modern Methods for Business Research*, G. Marcoulides, editor. Mahwah NJ: Erlbaum, 251-294.

Ryan, M. and B. Rayner (1998), "Estimating Structural Equation Models with PLS," AMA Advanced Research Techniques Forum, Keystone CO, June.

Schumacker, R. and G. Marcoulides (1998), *Interaction and Nonlinear Effects in Structural Equation Modeling*, Mahwah NJ: Erlbaum.

Torgerson, W. (1958), *Theory and Methods of Scaling*, New York: Wiley & Sons.

Vermunt, J. (1996), "Causal Log-Linear Modeling with Latent Variables and Missing Data," U. Engel and J. Reinecke, editor. New York: de Gruyter, 35-60.

------ (1997), *Log-Linear Models for Event Histories*, Advanced Quantitative Techniques in the Social Sciences, Thousand Oaks CA: Sage.

Vogt, W. (1993), *Dictionary of Statistics and Methodology*, Newbury Park CA: Sage.

Wold, H. (1973), "Nonlinear Iterative Partial Least Squares (NIPALS) Modeling: Some Current Developments," in *Multivariate Analysis- III*, P.R. Krishnaiah, editor. New York: Academic Press, 383-487.

------ (1980), "Soft Modeling: Intermediate Between Traditional Model Building and Data Analysis," *Mathematical Statistics*, 6, 333-346.

------ (1982), "Soft Modeling- The Basic Design and Extensions," in *Systems Under Indirect Observation (II)*, K. Jöreskog and H Wold, editors. Amsterdam: North-Holland, 1-53.

Wothke, W. (1998), Personal Communication, Bias in OLS estimators of PLS, June.

APPENDIX: SOFTWARE AND OTHER RESOURCES

A variety of software programs are available for fitting CSMs. Commercial programs include LISREL from Scientific Software Inc., AMOS from Smallwaters Corp. and SPSS Inc.; EQS from Multivariate Software, LISCOMP from Scientific Software as well as other distributors, Mplus from Muthén and Muthén, and the CALIS procedure in SAS. Non-commercial programs include MX and GENBLIS.

Software for PLS is quite scarce. Lohmöller's FORTRAN program LVPLS is available in compiled form for MS-DOS. As of this writing it can be obtained from John ("Jack") McArdle at the University of Virginia, and from two sources in Germany. The URL is:

ftp://kiptron.psyc.virginia.edu/pub/lvpls/

If you have trouble accessing this site, you might try e-mailing Fumiaki Hamagami at Virginia, fh3s@kiptron.psyc.virginia.edu.

Falk and Miller (1992) indicate that the FORTRAN source for LVPLS is still available from a source in Germany.

Wynne Chin at the University of Houston indicates he is developing a Windows-based program called PLS-GRAPH for fitting PLS-SEMs. He has a web page at

http://disc-nt.cba.uh.edu/chin/indx.html

that includes some PLS background and resources.

Lohmöller's seminal 1989 PLS book "Latent Variable Path Modeling with Partial Least Squares" is no longer in print. Falk & Miller (1992) provide a gentle and practical introduction to Lohmöller's LVPLS program, as well as a conceptual overview of soft modeling.

A number of good general references on CSMs exist. Bollen's 1989 text is seminal, although somewhat dated. Bollen and Long's "Testing Structural Equation Models" (1993) is an edited book with several useful chapters. Marcoulides & Schumacker (1996) and Schumacker and Marcoulides (1998) cover key issues in modeling nonlinearities, interactions, and models for change and growth. Marcoulide's 1998 book includes useful chapters on PLS-SEM by Chin (Chapter 10) and on CSMs by Rigdon (Chapter 9), Kaplan (Chapter 11), and McArdle (Chapter 12). The Chapters by Bagozzi & Yi, and by Fornell & Cha in Bagozzi's 1994(b) book "Advanced Methods of Marketing Research" are good introductions to CSMs and PLS. Dillon et al. (1997) provide an accessible overview to structural equation modeling. Finally, the November 1982 issue of Journal of Marketing Research was dedicated to SEMs (both CSMs and PLS), and is of at least historical interest.

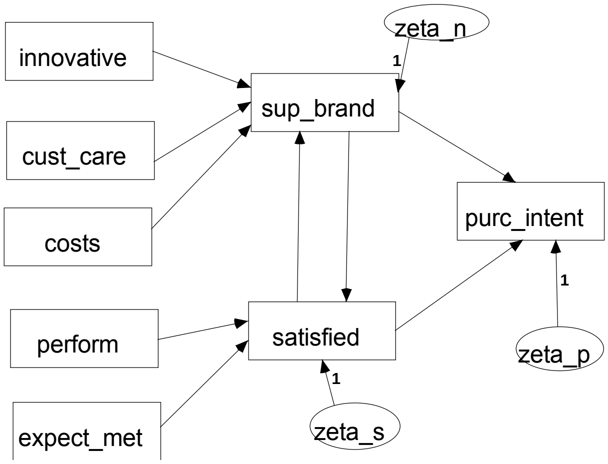


Figure 1. Example path diagram. Boxes are observed variables. Directed arrows show dependencies, e.g. **satisfied** depends on the three variables **perform**, **expect_met**, and **sup_brand**. Ellipses labeled **zeta_s**, **-_p**, and **-_n** are regression errors. They aren't in boxes because their values are inferred, i.e. they are latent variables. The "**1**" next to the arrows from the zetas represent a constraint that makes the measurement scales of the zetas match that of the dependent variable in their corresponding regression equation. This is a non-recursive model because of the loop between **satisfied** and **sup_brand**.

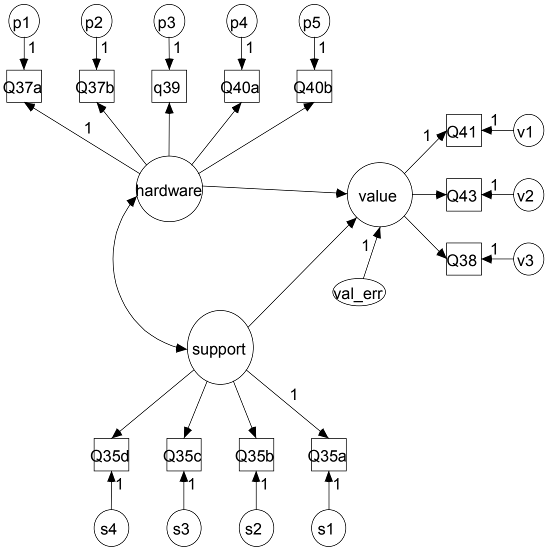


Figure 2. A recursive model for how **value** depends on **support** and **hardware**. All three of these variables are latent, and are estimated using measurement models having three, four, and five indicator variables, respectively. The indicators are in squares and have arrows point to their respective LVs. Each indicator has associated with it a measurement error named as a single letter and digits, e.g. **s1**. Each of these is drawn in a circle since they are unobserved. The curved, two headed arrow between **support** and **hardware** is a covariance. It represents an association that isn't directional, unlike the paths from **hardware** and **support** to **value**. This model is essentially a regression model in LVs that has two predictors that are correlated.

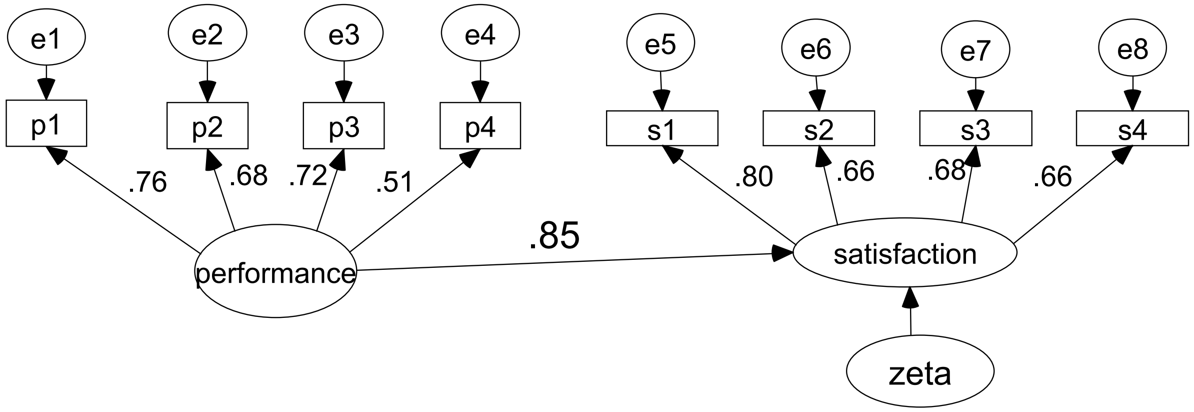


Figure 3. Path diagram showing standardized coefficient estimates for a CSM fit using AMOS 3.61. The variance-covariance matrix of the observed variables p1…s4 was analyzed. The variables p1 through p4 are the indicators for the LV performance. s1 through s4 are the indicators for the satisfaction LV. Zeta is a regression error. The coefficients labeling paths from the LVs to the observed variables are estimates of standardized factor loadings. The unobserved variables e1 through e8 are sources of variation unique to the observed variables they point to. The coefficient on the path between performance and satisfaction is the standardized regression coefficient for performance. N=200.

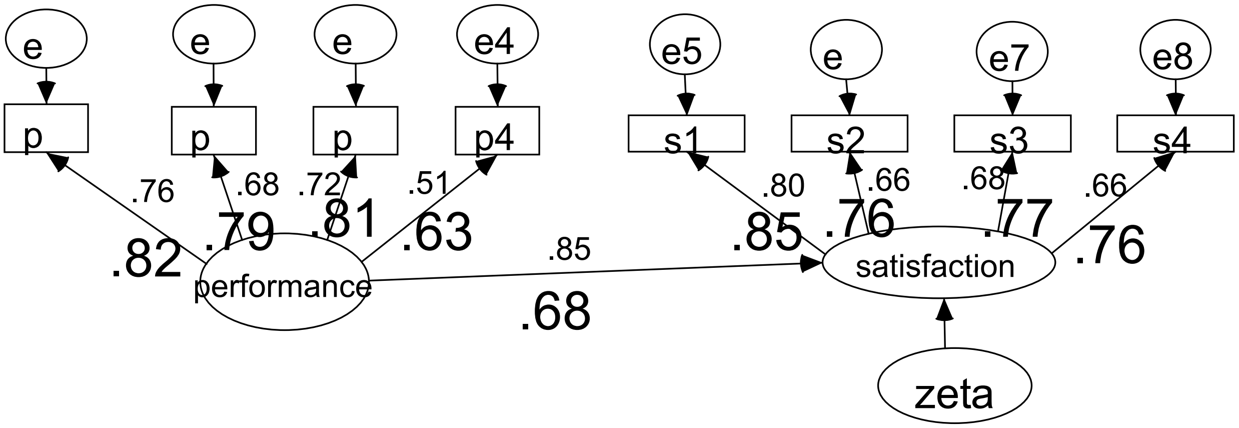


Figure 4. PLS coefficients shown in larger fonts from analysis of the correlation matrix of the observed variables p1…s4. The quantities in smaller font are standardized estimates from the CSM model shown in Figure 3. The diagram is otherwise labeled like Figure 3.

Table 1.
Basic specifications for covariance structure and PLS models.

The basic **CSM structural model** expresses the relationship between endogenous LVs η , intercepts α , regression coefficients B and Γ , exogenous LVs ξ, and errors in equations ζ .

$$\eta = \alpha + \mathrm{B}\eta + \Gamma\xi + \zeta$$

Where η , ξ, α , and ζ are vectors, and B and Γ are matrices.

CSM measurement models connect observed indicator variables Y for endogenous LVs, and observed indicators X for exogenous LVs:

$$Y = \Lambda_Y \eta + \varepsilon$$
$$X = \Lambda_X \xi + \delta$$

The matrices Λ contain factor loadings for the observed variables on the LVs, and $\varepsilon,\ \delta$ are "uniquenesses," i.e. sources of variance other than the LVs. Their expected values are assumed to be zero. Intercepts may also be included in the measurement models. A complete CSM specification includes covariance matrices for the various kinds of errors and for the exogenous LVs.

PLS has **inner** and **outer** models. The **PLS inner model** can be expressed as:

$$\eta = \alpha + \mathrm{B}\eta + \Gamma\xi + \zeta$$

Where the LVs in η can be ordered such that B is lower diagonal. Assumptions include $E[\zeta = 0], \mathrm{cov}(\zeta, \eta) = \mathrm{cov}(\zeta, \xi) = 0$. The **outer model** for *reflective* indicators is:

$$Y = \Lambda_Y \eta + \varepsilon_Y$$
$$X = \Lambda_X \xi + \varepsilon_X$$

And for *formative* indicators:

$$\eta = \Pi_Y Y + \delta_\eta$$
$$\xi = \Pi_X X + \delta_\xi$$

The Λ are loadings and the Π are regression coefficients. The expected values of the ε, δ are zero, and they are assumed to not covary with $\eta,\ \xi$.

The PLS specification also includes *weights* that are used to estimate the scores on $\eta, \; \xi$. The predicted values are obtained as, e.g. $\hat{\xi} = \sum wx$, and how the weights w are determined depends on the kind of outer model used for each LV as well as the LVs role in the inner model.

Table 2. Short comparison of CSM and PLS Characteristics

Feature	CSM	PLS
Distribution assumptions	Observed variables are MVN for ML, GLS estimation. Independent observations. Continuously distributed observed variables for ADF/WLS.	None
Types of models that can be fit	Recursive, non-recursive	Recursive
Types of observed variables	Continuous Ordered discrete (using polychoric correlations as input, assuming robustness, or using the LISCOMP specification.	Continuous Ordered and unordered discrete
Type of latent variables that can be modeled	Continuous	continuous
Types of indicators for latent variables	Effect- arrow from LV to indicator Causal- arrow from indicator	reflective, formative (analogous to effect and causal indicators, respectively)
Identification of parameters	Must be considered. Heuristics are available for common model forms. Otherwise, empirical procedures are required	Not an issue for standard PLS models.
Effects indicators per factor	Can be as few as one if the indicator's error is constrained. Otherwise, the number depends on parameter identification requirements.	One or more, but see "Consistency of estimators"
Factors per indicator	An observed variable can indicate more than one LV	Observed variables can only indicate one LV
Correlation between LVs can be estimated as undirected	Yes	No
Correlation between measurement errors can be modeled	Yes	no
Estimation of means and intercepts on LVs	Yes	No
Type of Fitting algorithm	Simultaneous estimation of parameters by minimizing discrepancies between observed and predicted VCV matrix (or correlation matrix). Full information methods.	Multi-stage iterative procedure using OLS. Model is divided into blocks whose parameters are estimated separately. A limited information method.

Table 2. (continued)

Feature	CSM	PLS
Consistency of estimators	Consistent, given correctness of model and appropriateness of assumptions	"Consistency at Large:" Estimates become consistent when the sample size gets large, and the number of indicators/LV gets large.
Availability of statistical tests for estimates	Available and valid given key model assumptions are tenable. Inference by bootstrapping, otherwise.	Inference requires jacknifing or bootstrapping.
Measures of fit	A great variety of them. Distributional theory worked out for some, not for others. Some allow inferences about nested models.	Coefficient of determination for each equation, Q^2 predictive relevance measures. Can be used for nested model comparisons with bootstrapping.
Assessing measurement model quality	Measures for assessing reliability and validity are available that permit observed variables to indicate more than one LV	Measures of reliability and validity are available.
Sample size requirements	Larger than for multiple regression. Procedures for estimating required N and for power analysis are available.	Small to moderate. A heuristic is to use 10-20 observations per parameter in the largest model block. See "consistency of estimators."
Factor indeterminancy	Latent variable scores are not estimated directly. Scale of measurement for each factor must be set using a constraint.	None. Latent variable scores are estimated as exact linear combinations of observed variables.
2nd-order factors can be modeled	Yes	Yes
Estimation of random coefficients	yes, for some model types	No
Latent class/finite mixture modeling	Yes, for some model types	No
Missing data	Algorithms assume complete data, but imputation can be done w/in some available SEM software packages.	Assumes complete data. Imputation using other software.

COMMENT ON BACON

P. B. Seetharaman
Washington University

OVERVIEW

This paper's point is that when one uses customer satisfaction data, measurement error in the variables used to predict satisfaction produces biased estimates of their effects ("errors in variables"). The paper emphasizes that such errors can be explicitly handled using structural equation models (SEM), and compares two SEM approaches – LISREL and PLS. It is an elegant review for practitioners especially since satisfaction studies are on the rise these days.

SPECIFIC COMMENTS

1. It will be useful to know which of the two approaches – LISREL or PLS – is more meaningful with the types of customer satisfaction data that are commonly used by marketing practitioners today. By explaining the two approaches using typical marketing data, the paper can demonstrate pertinence to marketers.
2. Satisfaction is not useful in and of itself as much as its effect on purchase intent in the product category under question. Therefore, it will be useful to use a model that has purchase intent as the dependent variable in order to illustrate the managerially relevant consequences of using one methodology versus the other.
3. The paper says that while LISREL may be better for theory-testing, PLS may be better for making predictions. How do the two approaches fare from a *policy-making* standpoint? Specifically, which variables should one control in order to boost repeat-purchases? To the extent that the estimated standardized coefficients depend on the variability of the predictor variables across respondents, they may be "masking" true importances of variables. Can one correct for this?
4. The paper makes a strong point in favor of LISREL. Are there compelling scenarios where one may prefer PLS to LISREL? For example, while performing *mixed-mode* estimation using an SEM that uses both reflective and formative indicators simultaneously, is it feasible to employ LISREL?
5. Expectations are an important mediating variable in customer satisfaction models. Such expectations are, by their very nature, dynamic. How does one incorporate such time-variations in SEMs? Can PLS handle this better than LISREL (or factor-analytic approaches in general)?

6. The paper says that there is a natural way to address collinearity issues in SEMs. However, collinearity can be due to two effects: *low discriminant validity* i.e. multiple constructs capturing the same effect, or *scale defects* i.e. respondents being overburdened with the task and giving similar responses on the scale for many questions. The solution to the first problem may not be the same as that for the second. It will be useful to understand the differences between the two cases.

PRODUCT MAPPING WITH PERCEPTIONS AND PREFERENCES

Richard M. Johnson
Sawtooth Software, Inc.

BACKGROUND

In the ‘50s and ‘60s, mathematical psychologists developed theories about how perceptions and preferences might be related. They considered objects to be arranged in some kind of perceptual space, determined either with respect to perceived similarities, or with respect to ratings on descriptive attributes. Each individual was also thought to have an ideal direction in the space and to prefer objects that were farther in that direction, or to have an ideal point in the space and to prefer objects closer to that point.

Market researchers have found these ideas very fruitful. Use of product maps became wide-spread in marketing research in the ‘60s and ‘70s, and they have proved to be useful aids for thinking about differences among products, customer desires, and ways in which products might be modified to become more successful.

Perceptual data have been used most often to create product spaces. In early years judgements about overall similarity of pairs of products were used with multidimensional scaling techniques. However, in later years attribute ratings have been used more widely, analyzed with factor analysis, discriminant analysis, or correspondence analysis.

Preference data have also been used to develop product spaces in marketing research. The first techniques for making maps based on preferences were developed in the early ‘60s: Coombs’ Unfolding method (which assumed each individual had an ideal point) and Tucker’s Points of View approach (which assumed each individual had a preferred direction in space). In an important contribution in 1970, Carroll and Chang showed that vector models can be regarded as special cases of generalized ideal point models, and they also provided the first practical method for estimating ideal points.

Most methods for making product maps have used either perceptual data or preference data, but seldom both. And there have been problems with maps of both types.

Maps based on **perceptions** are easy to interpret and good at conveying insights, but they are often less good at predicting individual preferences. One reason is that they may focus on differences that are easy to see but less important in determining preferences.

Maps based on **preferences** are better at accounting for preferences, but their dimensions are sometimes hard to interpret. For example, consider cups coffee with differences in temperature ranging from boiling to tepid. Most of us would probably prefer some middle temperature and reject both extremes. But if no perceptual information is available to establish their differences on the underlying temperature scale, the most extreme cups may be close together in a preference-based map, because their only recognized property is that they are both rejected by nearly everyone.

There is still another problem with **aggregate** maps of both types: a product's position on a map is based on the average of many individuals' perceptions or preferences. Because individuals differ, a single map can seldom describe different individuals' perceptions or preferences very precisely.

At the previous (1997) Sawtooth Software Conference, John Fiedler and Terry Elrod presented papers analyzing the same data but using different methods. John used a discriminant-based method which considered only **perceptual** data in the form of attribute ratings, and Terry used a technique he had developed which considered only **preference** data. Their maps were surprisingly similar. This reinforced the underlying theory relating perceptions and preferences, and suggested that even better maps might be produced if based on *both* perceptions and preferences. In the few instances where both types of data have been used, the usual practice has been first to use perceptual data to make the map, and then to fit preference data to it "after-the-fact." By contrast, because the methods described here use *both* perceptual and preference data simultaneously, we call them "composite" methods.

COMPOSITE METHODS

We have developed both "vector" and "ideal point" models. Each model uses both **perceptual** data, consisting of product ratings on attributes, and **preference** data, consisting of paired-comparison preference ratings for products. These are the same types of data as provided by APM, a perceptual mapping product released by Sawtooth Software in 1985. In fact, both composite models use APM data files, although they make no use of product familiarities or explicit ideal point ratings.

These two new models share several characteristics:

- Every respondent has a unique perceptual space, determined by his/her own ratings of products on attributes.
- Each dimension in the individual's space is a weighted combination of his/her ratings of products on attributes.
- However, the attribute weights defining the dimensions are required to be identical for all respondents.
- Those attribute weights are determined by optimizing the fit (over all individuals) between actual preferences and the preferences inferred from the individual perceptual spaces.

The overall product map is just the average of the individual maps. It is also a weighted combination of average product ratings, so it is truly a ***perceptual*** map, although with dimensions chosen so as to account best for preferences. In this way we produce maps which are firmly grounded in descriptive attributes, but which better account for individual preferences.

Both models require the estimation of two sets of parameters. One set consists of **attribute weights**, identical for all individuals, to be applied to attribute ratings to obtain the dimensions of individuals' perceptual spaces. The other parameters are unique for each individual: either **individual importance weights** in the case of the vector model, or **individual ideal point**

coordinates in the case of the ideal point model. For both models, estimation is done using alternating least squares.

The ***vector model*** assumes that each individual has some preferred direction in space, and prefers products that are "farther out" in that direction. It is appropriate for product spaces that have dimensions where "more (or less) is always better." An initial guess is made at the attribute weights. Based on the implied perceptual spaces, the best-fitting importance weights are estimated for each individual. Then, given those importance weights for all individuals, an improved set of attribute weights is estimated. The procedure alternates between estimation of individual importance weights and common attribute weights, continually improving the goodness of fit, as measured by an r-square value. To aid interpretability, the overall map is constrained to have orthogonal dimensions, and each dimension is scaled so that the sum of squared product coordinates is equal to the square of the number of products.

The ***ideal point model*** assumes that each individual's liking for products depends on products' perceived distances from an ideal point. Such models are more appropriate for product categories in which respondents may prefer combinations of attributes corresponding to interior regions of the space. Our approach is somewhat simpler than that of Carroll and Chang in PREFMAP. To keep things simple, we assume that individual preference contours are circular; in other words, we do not permit individuals to weight dimensions differently. Also, our implementation of the ideal point model assumes that each individual's ideal point is interior to the convex hull of his/her perceived product locations. As a result, our vector model is not a special case of our ideal point model.

Estimation of the ideal point model is done by minimizing a weighted sum of squared distances. From each respondent's preference data we obtain a set of positive weights that sum to unity, similar to shares of preference in conjoint analysis. We also compute the squared distances from each respondent's ideal point to each product in his/her perceptual space. Finally, we sum the products of those squared distances times the corresponding preference weights. We minimize this sum over all respondents, producing a solution in which more preferred products have smaller distances from ideal points. (This approach has some similarity to that of Desarbo and Carroll (1985), although they use weights based on preferences to scale discrepancies between observed and predicted distances, whereas we use the weights to scale the distances themselves.)

If the preference weights were all equal, then the algorithm would minimize the sum of squared distances from respondents' ideal points to all of the products, and all ideal points would be estimated to be at the center of the space. At the other extreme, if the preferred product had a weight of unity and all others had weights of zero, each individual's ideal point would be estimated as coincident with his or her preferred product. As with conjoint simulators, one can choose an scale factor to apply to the preference weights that produces any behavior between those two extremes. For our examples we use an scale factor of 10, which requires each individual's ideal point to be quite close to his or her preferred product. However, there can be a lot of heterogeneity in the way individuals perceive brands. Even if each person's ideal point were made to coincide perfectly with his or her preferred brand, ideal points could still be dispersed all over the aggregate map.

As with the vector model, we constrain the overall map to be orthogonal, with the sum of squared product coordinates for each dimension equal the square of the number of products. This avoids degenerate solutions in which all products and ideal points are coincident.

In what follows, we compare composite mapping results to those from discriminant analysis, using two data sets. The first data set is an artificial one, deliberately constructed to show the potential superiority of composite methods. The second data set consists of real data for motorcycles.

An Artificial Example

The first data set consists of perceptual and preference data for 300 artificial respondents, with three attributes and eight products. Imagine the product category to be busses used in rush hour commuting in a large city. The first attribute is Color of the bus, with levels red or blue. Color is deliberately chosen as an attribute on which people would agree which color a bus actually had, but which would have little impact on preference. The second attribute is Speed with levels fast and slow. The third attribute is Roominess with levels of roomy and cramped.

The eight busses to be rated have all combinations of levels of the three attributes. The perceptual data were made heterogeneous by adding an independent random variable to each product's design value for each individual. Heterogeneity for Color was only half as great as for Speed and Roominess.

Respondent ideal points were also random, with mean near the center of the scale for each attribute, and with a large amount of random preference heterogeneity. Respondents' preference data were generated by constructing constant sum paired comparison answers for 12 pairs of bus concepts. Preferences were deliberately constructed so as not to be affected by Color. The sum of squared differences was first computed between each respondent's ideal point and his/her perception of each bus, considering only Speed and Roominess. Reciprocals of those sums were then exponentiated and percentaged, to simulate paired comparison preference values between 0 and 100.

Perceptual and preference data are like those used by Sawtooth Software's APM system, which uses discriminant analysis to construct perceptual maps. Discriminant analysis has optimal mathematical properties, guaranteeing that its maps will contain the greatest amount of information about how products are seen to differ from one another for a given number of dimensions.

Although we have data on three attributes, we desire a map using only two dimensions.

The results for APM's perceptual map are given in Table 1 and Figure 1. The first dimension consists almost entirely of Color. This is expected, since discriminant analysis accounts for as much variance as possible with its first dimension, and our variables are orthogonal, with Color having the least amount of disagreement about which color each bus has. The second dimension consists mostly of a combination of Speed and Roominess. The third dimension, if we had included it, would have consisted of another combination of Speed and Roominess, which would account for the balance of the systematic differences among products.

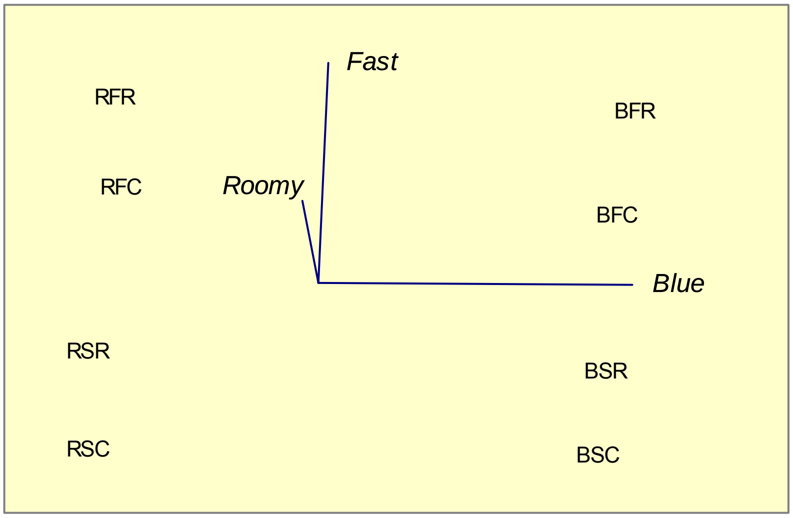
Table 1

```
        Discriminant Map for Synthetic Data

                          1      2
                        -----  -----
            Blue/Red     1.00  -0.01
           Fast/Slow     0.03   0.95
       Roomy/Cramped    -0.05   0.35

                 RSR    -8.61  -3.12
                 RFR    -7.73   7.90
                 RSC    -8.62  -7.30
                 RFC    -7.48   3.99
                 BSR     7.74  -3.91
                 BFR     8.94   7.34
                 BSC     7.46  -7.66
                 BFC     8.29   2.76
```

The second panel of Table 1 shows the coordinates of the 8 products in the discriminant space. Each product is identified by a three-letter string. The first letter is R or B indicating whether the bus is red or blue. The second letter is S or F, indicating whether it is slow or fast. The final letter is R or C, indicating whether it is roomy or cramped. As expected, the four red busses are all at one end of the first dimension and the four blue busses are all at the other end. The second dimension is characterized mostly by differences in Speed, with a small amount of information about differences in Roominess.

Since we constructed preferences to depend on Speed and Roominess but not Color, the first dimension of this space is useless for explaining preferences, and we have failed to capture important information about Roominess and Speed that would be required to account well for preferences.

Figure 1
Discriminant Map

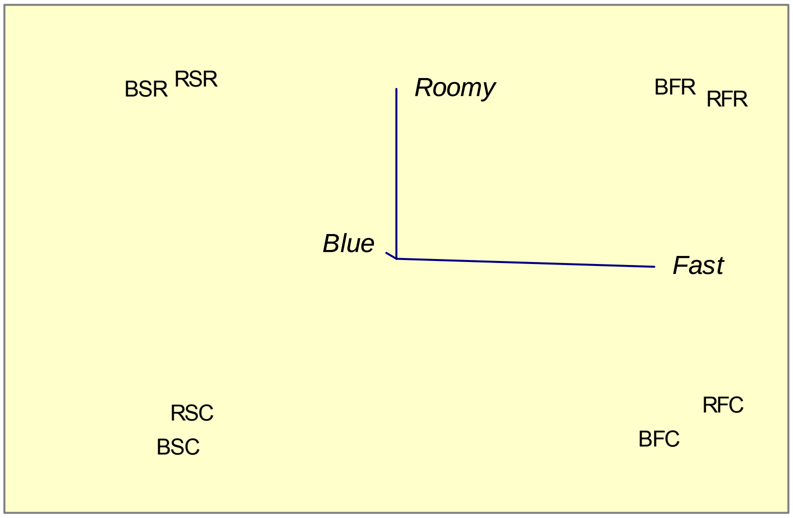


We next consider maps of the same data produced using both perceptual and preference data. First, here are results for the vector map:

Table 2
Composite Vector Map for Synthetic Data

	1	2
Blue/Red	-0.04	0.04
Fast/Slow	0.99	-0.05
Roomy/Cramped	0.00	1.00
RSR	-0.92	1.05
RFR	1.11	0.97
RSC	-0.93	-0.92
RFC	1.10	-0.87
BSR	-1.13	0.99
BFR	0.92	0.98
BSC	-1.01	-1.12
BFC	0.86	-1.07

Figure 2
Composite Vector Map



The first panel of Table 2 shows that Color scarcely enters into either dimension, and that both dimensions are concerned almost solely with Speed and Roominess, which we know are required to account for preferences in this example. The second panel shows that the corresponding red and blue products occupy similar positions in the space.

Composite maps may be subjected to any orthogonal rotation, so we have chosen to make the vectors for Speed and Roominess nearly horizontal and vertical. It is clear that the corresponding red and blue products occupy similar positions in the space. The small differences in location between corresponding products are due to the random heterogeneity of perception that was used in construction of the data file.

More important, this map accounts for preference more successfully than the discriminant map. If we find the direction in each map which best accounts for each respondent's preferences, we get an average r-squared between predicted and actual preferences of .59 for the discriminant map, vs. .74 for this map. We may also count the average number of correct orders for pairs of products when comparing actual rank orders of preference vs. predicted rank orders. For the discriminant map, 74% of the pairwise comparisons are correct, vs. 80% for this map. We should not expect either map to work perfectly because the vector model assumes that individual ideal points are infinitely far from the center of the space, and we constructed this data set so that a large proportion of the ideal points were near the center of the space.

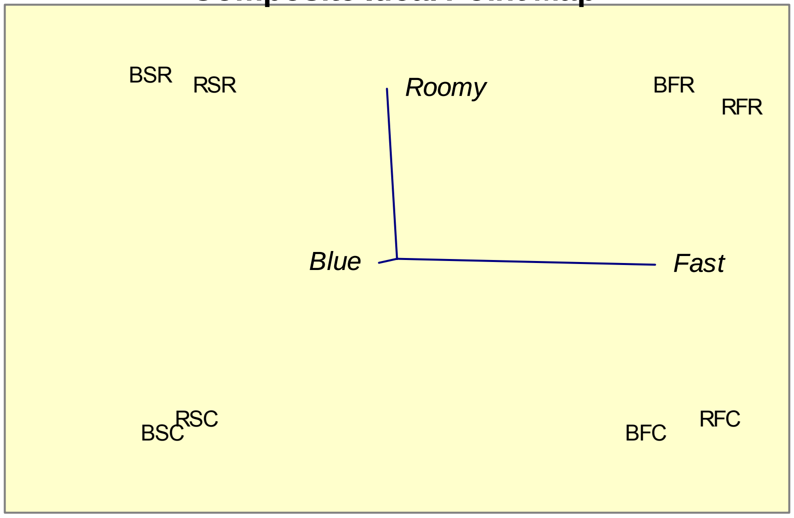
We next consider a map of the same data produced using a composite ideal point map.

Table 3
Composite Ideal Point Map for Synthetic Data

	1	2
Blue/Red	-0.07	-0.02
Fast/Slow	0.99	-0.03
Roomy/Cramped	-0.04	1.00
RSR	-0.85	1.01
RFR	1.17	0.89
RSC	-0.92	-0.96
RFC	1.09	-0.96
BSR	-1.12	1.08
BFR	0.91	1.01
BSC	-1.07	-1.04
BFC	0.80	-1.04

The ideal point map, like the preceding vector map, ignores Color almost entirely, concentrating on Speed and Roominess. As with the vector map, the products that are identical except for Color are nearly superimposed, and the small differences between them are due to the random heterogeneity built into the perceptual data.

Figure 3
Composite Ideal Point Map



This map also accounts for preferences far more successfully than the discriminant map. We can use the preference data to estimate individual ideal points for both maps, and then compute distances from each product to each respondent's ideal. We can evaluate the performance of each map in accounting for preference by counting the number of product pairs for which the preferred product is closer to the respondent's ideal point. For the discriminant map, 79% of the pairs are correct, and for this composite ideal point map 92% of the product pairs are correct. This map does not provide perfect prediction because it restricts ideal points to lie within the convex hull of the product points, and the data set was constructed so that many of them actually lie outside that region.

In addition to the results shown, the composite vector mapping approach also estimates an ideal direction for each respondent, and the composite ideal point mapping approach estimates an ideal point for each respondent. We don't show those because the ideal points were constructed so as to comprise an undifferentiated "blob" of little interest. However, we shall consider individual preference information for the next data set.

MOTORCYCLES

The second data set concerns road-going motorcycles, and was contributed by Tom Wittenschlaeger of Hughes Aircraft Company and John Fiedler of POPULUS, Inc. The data were collected in the United States in 1994 from a sample of 150 motorcycle riders. The project was methodological rather than substantive, so the data should not be used to develop marketing strategy, but serve nicely to illustrate mapping techniques.

The interview was typical of an APM questionnaire. Each respondent rated the importance of 10 attributes, and his/her familiarity with 11 motorcycle brands, and then rated the most familiar 5 motorcycles on the 5 most important attributes. The respondent's "ideal motorcycle" was also rated on the same scales. Finally, eight random pairs of motorcycles were presented and the respondent was asked to allocate 100 points between the members of each pair, indicating the relative likelihood of choosing each one in a purchase situation. Neither attribute importances ratings, product familiarity ratings, nor explicit ideal points were used in this analysis.

The attributes and products rated were as follows. Each has been given a short label to identify it on the maps:

Attributes:

Image	Has the image I prefer
Safe	Meets high standards of safety
Perform	Has high performance
Unique	Has a unique look and feel
Value	Offers good value for the money
Service	Has excellent service and support
Quality	Has high quality
Style	Has beautiful styling
Engin	Has excellent engineering
Fun	Is fun to ride

Products:

HON	Honda
KAW	Kawasaki
SUZ	Suzuki
YAM	Yamaha
DUC	Ducati
GUZ	Moto Guzzi
BIM	Bimota
BMW	BMW
TRI	Triumph
NOR	Norton
HAR	Harley Davidson

We have no market data with which to compare inferences from maps, but we do have preference data from the same respondents. Recall that each respondent selected the five products with which he or she was most familiar, and then answered paired comparison preference questions for eight random pairs of those products. We can accumulate the average preference proportions awarded to each product, which are shown in Table 4.

Table 4
Average Preference Percentages

HAR	71
HON	60
BMW	54
KAW	47
YAM	42
DUC	42
GUZ	41
SUZ	40
TRI	39
BIM	38
NOR	37

We should expect Harley Davidson, Honda, and BMW to have positions on the map indicating relative desirability.

The values in Table 4 do not reflect differences due to familiarity. Harley Davidson and the four Japanese brands were familiar to many respondents, while Bimota, Ducati, Moto Guzzi, and Bimota were familiar only to few. The more familiar brands will have more influence in determining the structure of the maps.

We now compare the maps produced by discriminant analysis with those produced by the two composite methods, using both vector and ideal point models.

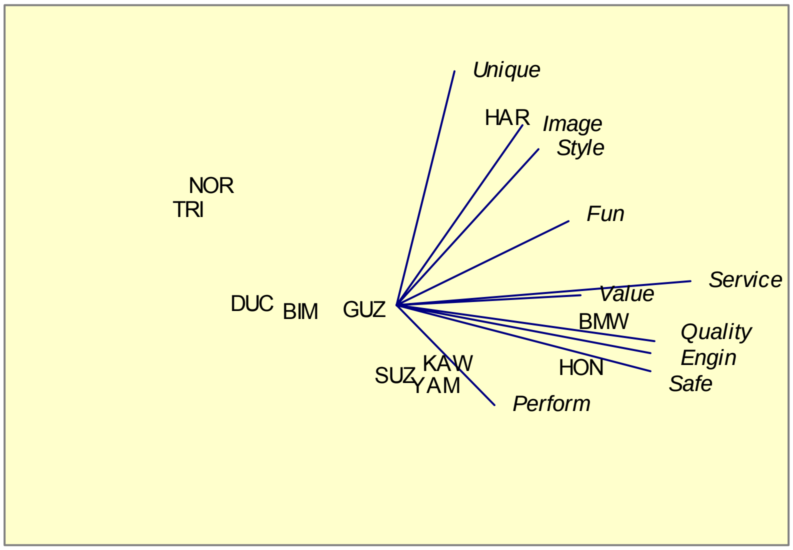
Table 5

Discriminant Map for Motorcycle Data

	1	2
Image	0.32	0.60
Safe	0.65	-0.17
Perform	0.25	-0.33
Unique	0.15	0.78
Value	0.47	0.03
Service	0.75	0.08
Quality	0.66	-0.12
Style	0.36	0.52
Engin	0.65	-0.16
Fun	0.44	0.28
HON	0.37	-0.21
KAW	0.02	-0.20
SUZ	-0.10	-0.24
YAM	-0.01	-0.27
DUC	-0.47	-0.00
GUZ	-0.18	-0.02
BIM	-0.34	-0.05
BMW	0.42	-0.06
TRI	-0.64	0.32
NOR	-0.58	0.40
HAR	0.18	0.62

The correlations of the attributes with the two largest dimensions in the discriminant space are given in Table 5. This map has been rotated so all attribute vectors point toward the right side of the space. These attributes would all be regarded as favorable by most motorcycle riders, so one would expect the preferred brands to be toward the right side of the space. The graphical representation of these data is given in Figure 4. (The product coordinates have been scaled down by a factor of .25 to make their average absolute values approximately equal to those of the correlations.)

Figure 4
Discriminant Map



Harley Davidson is alone in the upper right quadrant, with large projections on Uniqueness, Image, Style, and Fun. BMW and Honda are in the lower right quadrant, with large projections on the Service, Value, Engineering, Quality, Safety, and Performance. BMW has larger projections on the attributes pointing upwards, and Honda is larger on Performance. The British brands Norton and Triumph are in the upper left quadrant, with large projections on Unique, Image, and Style. The three Italian brands are in the lower left quadrant, rather close to the center, and the Japanese brands Kawasaki, Suzuki, and Yamaha are toward the bottom of the map, with high projections on Performance.

In APM questionnaires the respondent is asked to describe his ideal product on the same attributes as he describes existing products. These explicit ideal points can also be incorporated into the map, although there is some question about the reasonableness of this procedure with attributes where the ideal levels might be at infinity. Rather than show all 150 respondents' explicit ideal points individually, we have done a cluster analysis, and show the locations of the centers of three clusters in Table 6.

Table 6
Centroids of Ideal Point Clusters in Discriminant Map

	1	2
Cluster A (39%)	1.16	-0.07
Cluster B (36%)	0.22	0.06
Cluster C (25%)	0.74	1.32

Cluster A, which contains 39% of the respondent sample, has an average position much farther to the right than any product, and slightly below the horizontal axis. Those respondents would be expected to favor BMW and Honda.

Cluster B, which contains 36% of the respondent sample, has an average position slightly to the right of and above the origin. Those respondents might be expected to prefer any of the products.

Cluster C, which contains 25% of the respondent sample, has a position far to the right, and very far above all the products, approximately in the direction of the Image vector. Those respondents would be expected to have very strong preferences for Harley Davidson.

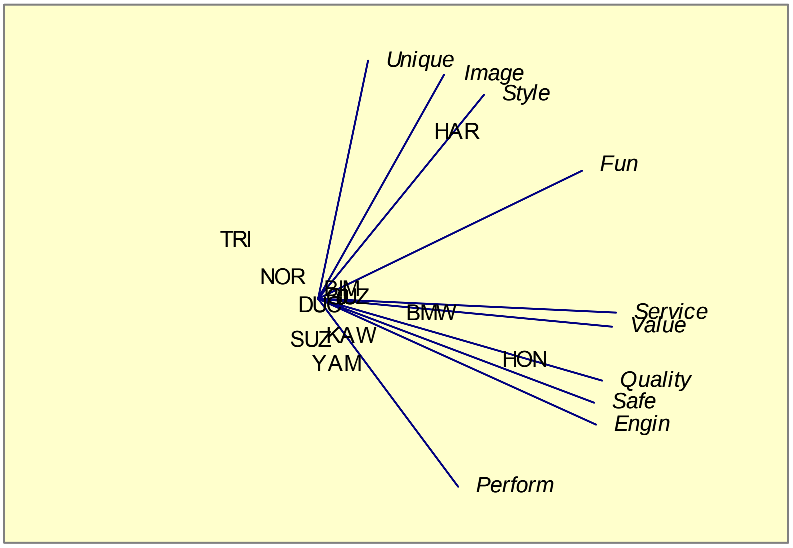
We turn now to the composite vector map, for which data are given in Table 7.

Table 7
Composite Vector Map for Motorcycle Data

	1	2
Image	0.40	0.91
Safe	0.88	-0.43
Perform	0.45	-0.77
Unique	0.16	0.97
Value	0.94	-0.12
Service	0.95	-0.06
Quality	0.91	-0.34
Style	0.53	0.83
Engin	0.89	-0.42
Fun	0.84	0.52
HON	2.11	-1.01
KAW	-0.11	-0.62
SUZ	-0.57	-0.70
YAM	-0.29	-1.01
DUC	-0.49	-0.13
GUZ	-0.13	-0.14
BIM	-0.15	-0.01
BMW	0.90	-0.27
TRI	-1.58	0.93
NOR	-0.96	0.32
HAR	1.26	2.69

This map has also been rotated so all attribute vectors point toward the right side of the space. It has strong similarities to the one produced by discriminant analysis. Harley Davidson is again alone in the first quadrant, with high projections on Unique, Image, Style, and Fun. BMW and Honda are again in the lower right quadrant, with strong projections on the remaining six attributes. The British products are again in the upper left quadrant, but this time much closer to the center of the space. The Italian products are very close to the center, and Suzuki, Kawasaki, and Yamaha are again in the lower left quadrant. The main apparent differences are that BMW is now much closer to the center than Honda, and the relatively low-rated products, which are those on the left side of the space, have moved toward the center.

Figure 5
Composite Vector Map



The fact that the discriminant map looks much like the composite map suggests that none of the attributes captures large but unimportant differences. This map is the average of 150 individual maps, on which respondents have different opinions about the positioning of products. The fact that the less popular products are closer to the center in the aggregate map suggests that they are in fact preferred by some respondents, who view them more favorably in terms of these attributes.

Since this is a vector map, individual preferences are given by “ideal direction” in space. We summarize those data by a count of the proportion of individuals whose ideal direction lies in each of eight “compass directions.”

Table 8

Summary of Ideal Vector Locations

-----Bearing-----	Percent
0 to 45 degrees	31
46 to 90 degrees	26
91 to 135 degrees	20
136 to 180 degrees	7
181 to 225 degrees	6
226 to 270 degrees	0
271 to 325 degrees	2
326 to 360 degrees	6

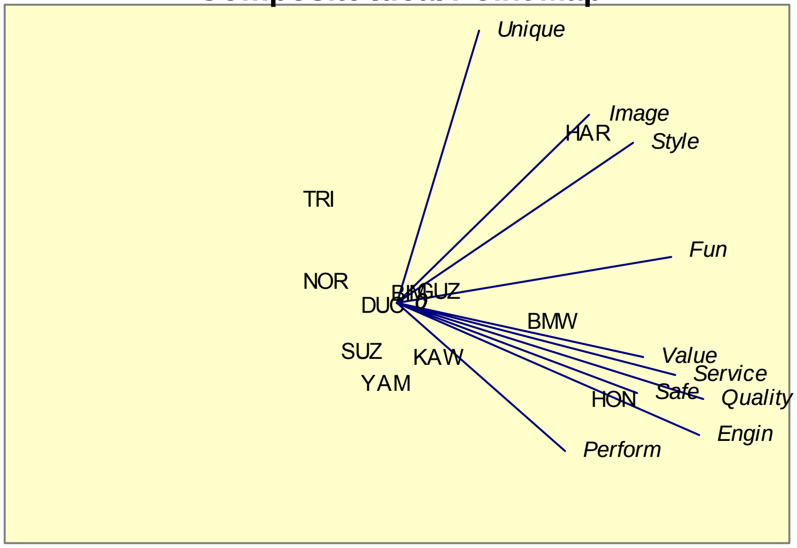
Nearly all respondents' ideal directions are in the right side of the space, with 57 percent in the upper-right quadrant and 27% in the lower right quadrant. These results are also similar to those for the discriminant map, which suggested that the strongest demand was for products in the upper right quadrant. In fact, this map and the discriminant map are very similar in terms of fitting preferences. If we find the direction in each map which best accounts for each respondent's preferences, and then count the percentage of correct orders for pairs of products when comparing actual rank orders of preference vs. predicted rank orders, we get 92% correct for the discriminant map vs. 93% correct for this map.

Finally, we consider the composite ideal-point map, for which data are given in Table 9.

Table 9
Composite Ideal Point Map for Motorcycle Data

	1	2
Image	0.49	0.63
Safe	0.61	-0.30
Perform	0.43	-0.49
Unique	0.21	0.91
Value	0.63	-0.18
Service	0.71	-0.24
Quality	0.78	-0.32
Style	0.60	0.54
Engin	0.77	-0.44
Fun	0.70	0.16
HON	1.80	-1.29
KAW	-0.02	-0.73
SUZ	-0.76	-0.67
YAM	-0.55	-1.09
DUC	-0.56	-0.05
GUZ	0.02	0.15
BIM	-0.24	0.02
BMW	1.15	-0.25
TRI	-1.23	1.37
NOR	-1.15	0.28
HAR	1.54	2.26

Figure 6
Composite Ideal Point Map



This map is very similar to the vector map of Figure 5. Both products and attribute vectors are in similar positions, and it seems doubtful that one would reach different conclusions from the two maps.

The mapping computation creates a file of individual ideal point estimates. We have subjected those to a cluster analysis, and have chosen to report three clusters. The coordinates of their centroids are given in Table 10.

Table 10
Centroids of Ideal Point Clusters in Ideal Point Map

	1	2
Cluster A (62%)	2.61	0.63
Cluster B (27%)	13.49	1.95
Cluster C (11%)	15.72	14.19

It may seem surprising that clusters B and C have locations far to the right of all of the products, when ideal points are required to be close to those individuals' preferred products. The reason for this is that there is considerable variation in individual product perceptions. Some individuals see their preferred product as very far to the right of its average location, and their ideal points are estimated to be near those locations.

As with the discriminant and vector maps, there is clear evidence of preference for products to the right of and above the horizontal axis. The largest cluster (62%) is centered to the right of and slightly above the origin. Since there is a lot of dispersion within each cluster, those individuals might prefer any of the products.

The second largest cluster (27%) is centered very far to the right, and moderately above the horizontal axis. Those respondents seem likely to favor Harley Davidson, Honda, and BMW.

The third largest cluster (11%) are extremely far to the right and extremely high, in the direction of Harley Davidson but much farther. They seem likely to be strong Harley preferrers.

This map also accounts for preferences slightly more successfully than the discriminant map. We can use the preference data to estimate individual ideal points for both this map and the discriminant map, and then compute distances from each product to each respondent's ideal. We can evaluate the performance of each map in accounting for preference by counting the number of product pairs for which the preferred product is closer to the respondent's ideal point. For the discriminant map, 89% of the pairs are correct, and for this composite ideal point map 92% of the product pairs are correct.

Estimating Demand

Since composite maps provide a relatively tight linkage between perceptions and preferences, it is tempting to consider ways of using composite maps to estimate demand for new products. There are at least two ways to do so.

One way might be to construct a simulator to predict preferences for new products, or products modified on one or more attributes. That approach was used in Sawtooth Software's APM product, although the manual listed several reasons why it might not be completely successful. The basic problem is that there is a many-to-one mapping of attributes into dimensions. Products can have quite different levels on several attributes, and yet occupy the same point in space. Multicolinearity among attributes makes it difficult to infer the relative importance of each. Also, if a product is changed on one attribute without corresponding changes on others, then the space itself may change, and with it the capability of making inferences.

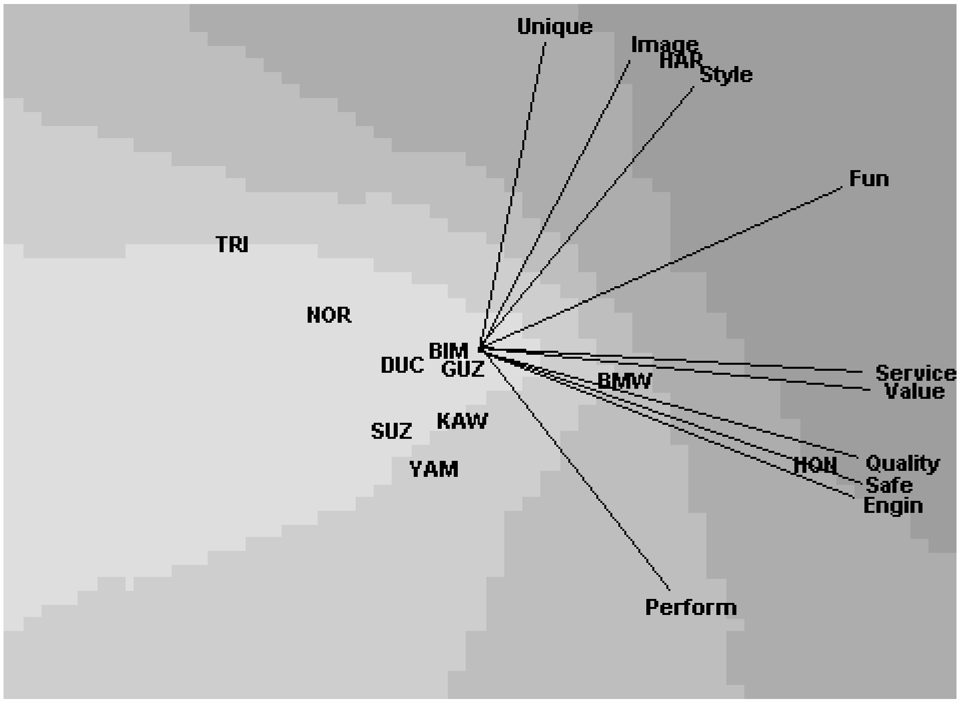
The other way to estimate demand for modified products is to consider products that differ in terms of locations in the existing space, rather than differing on specific attributes. That is the approach used in what follows.

Figures 7 and 8 present that information for the composite vector and ideal point maps, respectively. Relative demand is estimated using a "first choice rule." For the vector map, we see whether a product at each grid point would be the "farthest out" in each respondent's ideal direction (relative to other products), and score a 1 if so and a 0 if not. For the ideal point map we see whether a product at each grid point would be closest to each respondent's ideal point (relative to other products), and score a 1 if so and a 0 if not. The total number of hits is thus computed for a hypothetical new product at each grid point. The grid points are divided into quintiles in terms of their numbers of "hits," and each point is given background shading to indicate its quintile. The darker regions indicate higher demand.

As is evident in Figure 7, which provides a display for the composite vector model, the demand for new products would be greatest if they were positioned far to the right and in the upper half of the space.

Figure 7

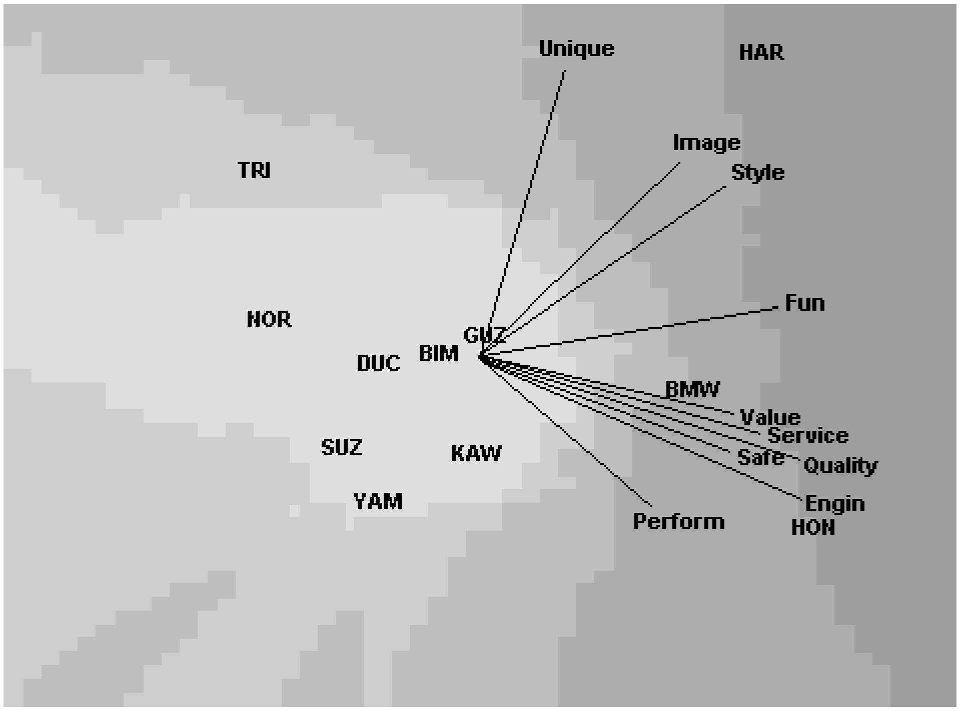
Density of Demand for Vector Model



Similar conclusions would be reached from Figure 8, which provides a similar display for the composite ideal point model, although in this map the region of highest demand seems to extend somewhat further downward, suggesting that there is more opportunity for the attributes pointing in that direction.

Their ability to portray relative demand for new or modified products is a major benefit of composite maps. Because composite maps are based on simultaneous analysis of perceptual and preference data from the same individuals, they have a strong advantage in this regard over other methods, which enhances their usefulness to managers. (However, we should repeat our earlier caveat, that the data for our examples were collected for methodological rather than substantive purposes, and these particular maps should not be used for marketing strategy purposes.)

Figure 8
Density of Demand for Ideal Point Model



Although the two maps produce similar information about likely demand for new products, it is reasonable to inquire which is better. The vector model accounts for preferences slightly more effectively, correctly fitting 93% of the pairwise judgements as opposed to 92% for the ideal point model. However, the product points seem to be more spread out in the ideal point map, suggesting that it supports finer distinctions among products. It would be premature to decide that one method is better thasn the other based just on this data set.

Summary and Conclusions

We have introduced techniques for Composite Mapping, which make use of both perceptual and preference data. The basic idea is that each respondent has an individual map in which product locations are weighted combinations of attribute ratings, and the weights are identical for all individuals. The aggregate map is the average of the individual maps. The weights used for all individuals are determined so as to maximize the correspondence between individuals' stated preferences and the preferences that would be inferred from the resulting maps. There are separate algorithms for vector maps and ideal-point maps.

Our results lead to these conclusions:

- Although mapping based on perceptual data alone can portray product images efficiently in maps of few dimensions, it can err by concentrating on differences among products that are easy to see but not important for preference.
- For an artificial data set in which two attributes were involved in preferences but a third had larger perceived differences among products, a perceptual map using discriminant analysis failed to account for preferences, but both composite mapping methods reproduced the known preference structure of the data.
- Although mapping based on preference data alone may be successful at explaining product preferences, the lack of perceptual information may lead to maps that are difficult to interpret.
- For a real data set, the perceptual map using discriminant analysis predicted preferences quite well, and was visually very similar to both the composite vector map and the composite ideal point map. This should occur when the attributes are approximately equal in importance for predicting preferences.
- The fact that the composite methods appear to produce better results when attributes differ strongly in their importances in affecting preference, but similar results when attributes are well chosen, suggests that composite maps can provide insurance against unfortunate choices of attributes
- Since composite maps provide a relatively tight linkage between perceptions and preferences, they may be used for forecasting relative demand for new or modified products.

All in all, there seems to be no downside to using composite mapping methods, and the benefit of possibly improved interpretation and prediction can be great.

APPENDIX

Estimation of Composite Mapping Models

The vector model and the ideal point model both use the following definitions:

Let there be ***N*** respondents, ***n*** products, ***p*** attributes, ***d*** dimensions.

For the i-th individual, let:

$\mathbf{X_i}$ = an (*n* x *p*) matrix of product ratings, with column sums of zero.

$\mathbf{Y_i}$ = an (*n*) vector of product preference values. APM provides constant-sum preference information obtained by having respondents divide 100 points among members of each of several product pairs. We construct $\mathbf{Y_i}$ by taking logits of those preference percentages, and awarding half the logit value to the winning product in that pair, and penalizing the losing product with half of the logit value. This results in $\mathbf{Y_i}$ values similar to conjoint utilities and that sum to zero for each individual.

T = a (*p* x *d*) matrix of weights used to transform attribute ratings into dimensional coordinates. **T** is common to all respondents. We want to find a **T** which permits the best fit to all respondent's preferences. We start with an approximation obtained from a principal components analysis of attribute ratings, and then improve it iteratively.

$\mathbf{C_i}$ = an (*n* x *d*) matrix giving the configuration of products for the i-th individual.

$$\mathbf{C_i} = \mathbf{X_i}\,\mathbf{T} \qquad (1)$$

The Vector Model

In the vector model each individual is thought of as having an ideal direction in the space, represented by a vector, and should prefer products according to their projections onto that vector.

Let $\mathbf{W_i}$ = a (*d*) vector of importance weights to be applied to columns of the individual's configuration to best predict that individual's preferences. Our basic individual preference equation is

$$\mathbf{C_i}\ \mathbf{W_i} - \mathbf{Y_i} = \mathbf{E_i} \qquad (2)$$

where $\mathbf{E_i}$ is a vector of errors of fit. Equation 2 says that the individual's configuration of products in space $\mathbf{C_i}$ is weighted by the elements of $\mathbf{W_i}$ to get a prediction of $\mathbf{Y_i}$.

Substituting from (1) into (2), we get

$$\mathbf{X_i\ T\ W_i\ =\ Y_i + E_i} \qquad (3)$$

If we knew **T**, we could solve for $\mathbf{W_i}$, using ordinary least squares. We start with an initial approximation of **T** and improve it in subsequent iterations. After estimating a $\mathbf{W_i}$ for each individual, we then combine information from all individuals to find a **T** that fits individuals better on average. By alternating between re-estimation of the **W**'s and **T**, we eventually find estimates for the **W**'s and **T** that best fit the data.

An initial estimate of **T** is obtained either from random numbers or from the principal components of the sum of all individuals' **X** matrices. In each iteration we solve for weights for each individual using ordinary least squares, minimizing the sum of squared errors in $\mathbf{E_i}$:

$$\overset{\wedge}{\mathbf{W}}_{\mathbf{i}} = \mathbf{(T'\,X_i^{'}XT)^{-1}T\ \ X_i^{'}Y_i} \qquad (4)$$

We also record the r-square for each individual as a measure of how well the model fits that individual's preferences.

To produce an improved estimate of **T**, we could use the partial derivatives of the sum of squared errors with respect to **T**. For the i-th individual, those partials are given in equation (5):

$$\frac{\partial\mathbf{(E_i^{'}E_i)}}{\partial\mathbf{T'}} = 2\mathbf{X_i^{'}X_iTW_iW_i'} - 2\mathbf{X_i'\,Y_iW_i'} \qquad (5)$$

The partial derivatives of the total sum of squared errors involves summing equation (5) over all respondents Setting the partial derivatives of that sum to zero yields an expression for **T** which minimizes the sum of squared errors.

$$\sum_{i}^{N}\mathbf{X_i^{'}X_iTW_iW_i'} - \mathbf{X_i'\,Y_iW_i'} = \mathbf{0} \qquad (6)$$

However, this equation appears to be intractable. In each term of the sum, **T** is premultiplied by $\mathbf{X_i'X_i}$ and postmultiplied by $\mathbf{W_iW_i'}$, so it is not clear how to solve for **T**.

One possibility would be separately to sum the products $\mathbf{X_i'X_i}$, $\mathbf{W_iW_i'}$, and $\mathbf{X_i'Y_iW_i'}$ and then to premultiply the sum of $\mathbf{X_i'Y_iW_i'}$ by the inverse of the sum of $\mathbf{X_i'X_i}$ and postmultiply by the inverse of the sum of $\mathbf{W_iW_i'}$. We have tried that, but find that the sum of squared errors obtained with that approximation does not decrease monotonically from iteration to iteration.

However, we have had success with a slightly different procedure. In equation (3) the weights $\mathbf{W_i}$ are applied to the product $\mathbf{X_i\ T}$ to predict $\mathbf{Y_i}$, but it is also true that the weights $\mathbf{TW_i}$ are applied to the matrix of attribute ratings, $\mathbf{X_i}$, to predict $\mathbf{Y_i}$. We may estimate weights ($\mathbf{T\ W_i}$) by premultiplying equation (3) by $(\mathbf{X'_i\ X_i})^{-1}\ \mathbf{X'_i}$:

$$\mathbf{T\,W_i} = (\mathbf{X_i'X_i})^{-1}\,\mathbf{X_i'Y_i} + (\mathbf{X_i'X_i})^{-1}\,\mathbf{X_i'E_i} \qquad (7)$$

The first term on the right hand side of (7) is the estimate of regression weights that would be obtained in trying to predict $\mathbf{Y_i}$ from the individual's ***entire*** set of attribute ratings, $\mathbf{X_i}$. The second term on the right hand side is an error term that we hope is small. If it were zero, then equation (7) would state that the individual's weights for predicting his/her preferences in his/her entire attribute space would be expressible as a weighted combination of the columns of **T**. If that were true, then there would be no loss of predictive ability from using a subspace of small dimensionality common to all respondents. Our method of estimation improves **T** in each iteration so as to minimize the sum of squares of the last term in equation (7).

To show this more clearly we define

$$\mathbf{V_i} = (\mathbf{X_i'X_i})^{-1}\,\mathbf{X_i'Y_i} \qquad (8)$$

$$\mathbf{F_i} = (\mathbf{X_i'X_i})^{-1}\,\mathbf{X_i'E_i} \qquad (9)$$

$\mathbf{V_i}$ is the vector of weights that would best predict $\mathbf{Y_i}$ from $\mathbf{X_i}$. Consider a new regression computation for each individual, fitting $\mathbf{V_i}$ as a weighted combination of the columns of **T.** Equation (10) is obtained by substituting from (8) and (9) into (7). Since the errors are different, (**F** rather than **E**), the estimated coefficients will be different as well. Call these coefficients $\mathbf{U_i}$ (rather than $\mathbf{W_i}$).

$$\mathbf{T\,U_i} = \mathbf{V_i} + \mathbf{F_i} \qquad (10)$$

The OLS solution for $\mathbf{U_i}$ is:

$$\mathbf{U_i} = (\mathbf{T'T})^{-1}\,\mathbf{T'}\,\mathbf{V_i} \qquad (11)$$

Assemble the $\mathbf{U_i}$ = and $\mathbf{V_i}$ vectors as columns of matrices **U** and **V**:

$$\mathbf{U} = (\mathbf{U_1}\,,\,\mathbf{U_2},\ldots\mathbf{U_N})$$

$$\mathbf{V} = (\mathbf{V_1}\,,\,\mathbf{V_2},\ldots\mathbf{V_N})$$

Then the OLS estimate of **T** that best fits equation (10) for all individuals is:

$$\hat{\mathbf{T}} = \mathbf{VU'}\,(\mathbf{UU'})^{-1} \qquad (12)$$

Each iteration consists of two steps. During the first step, individual r-squares are computed using the regression indicated in equation (3) to measure the goodness of fit to each individual's preferences with the current estimate of **T**. A second r-square value is also computed for each individual using equation (11), to indicate how successfully the individual's preference structure is captured by the **T** matrix. The $\mathbf{U_i}$ and $\mathbf{V_i}$ vectors are also saved for each individual, as determined in equations (8) and (11).

In the second step, the $\mathbf{U_i}$ and $\mathbf{V_i}$ vectors are assembled and used as in equation (12) to re-estimate **T**.

Because the same **T** is used for all individuals, the aggregate product configuration is obtained by averaging the individual $\mathbf{C_i}$ matrices. Since the goodness of fit to individual data is not affected by a linear transformation of the columns of **T**, we also adjust **T** in each iteration so that the columns of the aggregate configuration are orthogonal and each has sum of squares of *n*.

As iterations progress, the sum of goodness-of-fit r-squares from the regression of equation (3) tends to increase, but it is not required to do so, and it often fluctuates in later iterations. However, the sum of "preference structure capture" r-squares from the regression of equation (10) increases monotonically, and iterations are terminated when the increases fall to less than a small positive value.

There are some numerical problems that must be overcome. Some respondents produce little or no useful information in their attribute ratings, and there may be more attributes than products. In either case the inverse of $\mathbf{X_i'X_i}$ will not exist. Therefore all matrices to be inverted first have a small positive amount added to their diagonal elements. This "ridge regression" trick remedies both problems.

The individual weights $\mathbf{W_i}$ are saved in a file, together with the goodness-of-fit r-square value for that individual. Coordinates of a map are provided, showing the average positions of products and attribute vectors in a space. The product configuration is just the average of individuals' configurations. The attribute vector positions are indicated by correlations between average product attribute ratings and product coordinates on each dimension.

THE IDEAL POINT MODEL

For convenience, we repeat definitions stated above. For the i-th individual, let:

$\mathbf{X_i}$ = an (*n* x *p*) matrix of product ratings, with column sums of zero, and $\mathbf{X} = {}^{1}/_{N} \sum \mathbf{X_i}$

$\mathbf{Y_i}$ = an (*n*) vector of product preference values. APM provides constant-sum preference information obtained by having respondents divide 100 points among members of each of several product pairs. We construct $\mathbf{Y_i}$ by taking logits of those preference percentages, and awarding half the logit value to the winning product in that pair, and penalizing the losing product with half of the logit value. This results in $\mathbf{Y_i}$ values similar to conjoint utilities and that sum to zero for each individual.

T = a (*p* x *d*) matrix of weights used to transform attribute ratings into dimensional coordinates. **T** is common to all respondents. We want to find a **T** which permits the best fit to all respondent's preferences. We start with an approximation obtained from a principal components analysis of attribute ratings, and then improve it iteratively.

$\mathbf{C_i}$ = an (*n* x *d*) matrix giving the configuration of products for the i-th individual.

$$\mathbf{C_i} = \mathbf{X_i\, T} \qquad (1)$$

We start by exponentiating and then percentaging each individual's $\mathbf{Y_i}$ values to get a set of positive values that sum to unity, similar to "shares of preference" in conjoint analysis.
Call this vector of preference information $\mathbf{R_i}$.

We assume each individual to have an "ideal point" in his/her perceptual space, defined as the row vector $\mathbf{P_i'}$. The squared distance from each product to that ideal point is obtained by subtracting $\mathbf{P_i'}$ from each row of $\mathbf{C_i}$ and then summing the squared differences. Call the vector of squared distances $\Delta_\mathbf{i}^\mathbf{2}$.

We expect that the distances should be small for products most preferred, and larger for products less preferred. We can express this desired relationship between preferences and distances in terms of the sum over products of the individual's preference weights times his/her squared distances, which we wish were small. Call this value for the ith individual $\theta_\mathbf{i}$.

$$\theta_\mathbf{i} = \mathbf{R_i'}\,\Delta_\mathbf{i}^2 \qquad (13)$$

If $\theta_\mathbf{i}$ is small, then for the ith individual the products with large preferences must have small distances from the ideal. Our goal is to find an ideal point for each respondent ($\mathbf{P_i'}$) and a matrix of weights common to all respondent (**T**) that minimize the sum of the θ values for all respondents:

$$\theta = \sum \theta_\mathbf{i} \qquad (14)$$

To estimate the ideal point for the ith respondent, we differentiate $\theta_\mathbf{i}$ with respect to $\mathbf{P_i'}$ and set the result to zero. Observing that the sum of the **R**'s is unity, we get the equation:

$$\mathbf{P_i'} = \mathbf{R_i'\, C_i} \qquad (15)$$

The estimate of the individual's ideal point ($\mathbf{P_i'}$) which minimizes $\theta_\mathbf{i}$ is simply the weighted average of the rows of his/her matrix of perceived product locations, where the weights are the $\mathbf{R_i}$ values. Recall that the $\mathbf{R_i}$ values are positive and sum to unity. If the respondent has such extreme preference for one product that its value is unity and the rest are all zero, then the ideal point will be estimated to be coincident with that product's location. If the respondent is indifferent among products so that all $\mathbf{R_i}$ values are equal, then the ideal point will be estimated to be at the center of the perceptual space. No matter what the respondent's preferences, ideal points estimated this way will always lie within the convex hull of that respondent's perceived product locations.

Given an estimate of the ideal points for each individual, an improved estimate of **T** can be obtained as follows. Let $\mathbf{D_i}$ be a diagonal matrix whose diagonal elements are corresponding elements of $\mathbf{R_i}$. Then, differentiating θ with respect to **T** and setting the partial derivatives to zero and summing over respondents gives the equation:

$$\sum \mathbf{X_i' D_i X_i \; T} = \sum \mathbf{X_i' R_i P_i'} \qquad (16)$$

It would seem that one way to estimate T would be to cumulate the two sums

$$\mathbf{A} = \sum \mathbf{X_i' D_i X_i} \qquad (17)$$

and

$$\mathbf{B} = \sum \mathbf{X_i' R_i P_i'} \qquad (18)$$

so that

$$\mathbf{A\,T = B}$$

and then simply estimate **T** as $\mathbf{A}^{-1}\,\mathbf{B}$.

However, the problem with this approach is that θ is minimized trivially by a **T** of zero, and an iterative process which estimates **T** in this way eventually converges to a **T** of zero. To avoid that, it is necessary to impose constraints on **T**. We choose to make columns of the overall configuration **XT** orthogonal, and for each column to have sum of squares equal to the number of products, ***n***. (Recall that **X** is the average of the $\mathbf{X_i}$ matrices.)

This is done with a symmetric matrix of Lagrange multipliers, following Schonemann (1965). We differentiate the sum of $\theta + \phi$ with respect to **T**, where

$$\phi = \text{trace}(\mathbf{S\,T'\,X'\,X\,T}) \qquad (19)$$

with **S** an unknown symmetric matrix.

Setting the sum of partial derivatives to zero yields:

$$\mathbf{AT = B + X'X\ T\ S} \tag{20}$$

Premultiplying by **T'** and recalling that **T'X'X T** is constrained to equal the identity matrix, we get:

$$\mathbf{S = T'(\ AT - B)} \tag{21}$$

Premultiplying (20) by the inverse of **A**,

$$\mathbf{T = A^{-1}B + A^{-1}\ X'X\ T\ S} \tag{22}$$

Equation (20) cannot be solved explicitly for **T**, but does submit to an iterative solution consisting of the following steps:

1) Use the value of **T** from the previous iteration or some initial value, to compute **S** as in equation (21). Early estimates of **S** will not be symmetric, so force symmetry by averaging corresponding elements above and below the diagonal.

2) Obtain the product $\mathbf{A^{-1}\ X'X\ T\ S}$, using current estimates of **T** and **S**.

3) Determine a scalar α by which to multiply the product obtained in step 2) so that $\mathbf{XA^{-1}B + \alpha XA^{-1}\ X'X\ T\ S}$ has sum of squares equal to ***n* d*** .

4) Use the sum of terms: $\mathbf{A^{-1}B + \alpha A^{-1}\ X'X\ T\ S}$ as an interim estimate of **T** to obtain an interim (not necessarily orthogonal) estimate of the configuration of products in space, **X T**.

5) Find the matrix with orthogonal columns and column sums of squares equal to ***n*** which is closest in the least squares sense to **X T**, using the procedure of Johnson (1966), as well as the right-hand transformation matrix that performs that orthogonalization.

6) Finally, postmultiply the interim estimate of **T** from step 4) by the transformation matrix determined in step 5) to get the estimate of **T** for the current iteration.

Repeat steps 1-6 until estimates of **T** stabilize.

REFERENCES

Carroll, J. D. and Chang, J. J. (1970), “Analysis of Individual Differences in Multidimensional Scaling Via an N-way Generalization of Eckart-Young Decomposition,” *Psychometrika*, 35, 283-319.

Coombs, Clyde H. (1964) *A Theory of Data*. New York: Wiley.

DeSarbo, Wayne S. & J. Douglas Carroll (1985), “Three-Way Metric Unfolding Via Alternating Weighted Least Squares,” *Psychometrika*, 50, September, 275-300.

Elrod, Terry (1997), “Obtaining Product-Market Maps from Preference Data,” Sawtooth Software Conference Proceedings, 273-288.

Johnson, Richard M. (1966) “The Minimal Transformation to Orthonormality,” Psychometrika, 31, March, 61-66.

Schonemann, Peter H. (1965) “On the Formal Differentiation of Traces and Determinants,” Research Memorandum No. 27, The Psychometric Laboratory, University of North Carolina.

Tucker, Ledyard R and Samuel Messick (1964) “An Individual Differences Model of Multi-Dimensional Scaling,” *Psychometrika*, 333-367.

Wittenschlaeger, Thomas A. and J. A. Fiedler (1997), “Current Practices in Perceptual Mapping,” Sawtooth Software Conference Proceedings, 259-270.