PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

October 2004

FOREWORD

We are pleased to present the proceedings of the eleventh Sawtooth Software Conference, held in San Diego, CA, October 6-8, 2004. Our hotel on Shelter Island provided excellent views of the water, a fishing pier, and the busy boat traffic.

The focus of the conference was quantitative methods in marketing research. We were also treated to a thought-provoking paper on ethics. The authors were charged to deliver presentations of value to both the most and least sophisticated members of the audience. We were exposed to a variety of topics, including discrete choice, conjoint analysis, MaxDiff scaling, latent class methods, hierarchical Bayes, and IRT (Item Response Theory). The authors represented a mix of leading academics and practitioners, with the practitioners more heavily represented, as is typical for our conference.

The authors (and a few invited speakers) played the role of discussant to another paper presented at the conference. Discussants spoke for up to five minutes to express contrasting or complementary views. Some discussants have prepared written versions of their comments for this volume.

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

Sawtooth Software

January, 2005

CONTENTS

DATA COLLECTION ISSUES

COMPONENTS-BASED PRICING RESEARCH

CONJOINT ANALYSIS

DERIVED MEASURES OF PREFERENCE/IMPORTANCE

CONJOINT ANALYSIS CASE STUDIES

Applied Choice Models

Discrete Choice Design

Quantitative Approaches with Soft Attributes

MCMC Methods

Advanced Topics in Choice Modeling

SUMMARY OF FINDINGS

Nineteen presentations were delivered at the eleventh Sawtooth Software Conference, held in San Diego, CA. We've summarized some of the high points below. Since we cannot possibly convey the full worth of the papers in a few paragraphs, the authors have submitted complete written papers within this 2004 Sawtooth Software Conference Proceedings.

It's Ethical Jim, But Not in the Way We Used to Know It! (Ray Poynter, UK): Ray asserted that there is a near-global shift in attitudes regarding people in positions of authority, expert opinion, and research (whether dealing with "hard" or social sciences). People are much more skeptical today. At a broader level, these changes are due to very visible cases of personal/corporate corruption, citizen cynicism, lack of respect for public leaders, and the ubiquitous spin that is self-serving and often counterfactual. Within our own industry, there is an increase in direct marketing, falling response rates, an overall increase in polling, and a perception of convergence of direct marketing and market research.

Ray questioned how good the data are when we pursue some respondents relentlessly, after they have already refused multiple times to participate. He condemned improper uses of marketing research, such as SUGGING (selling under the guise of research) and push-polling. Legislation is increasing which may impact our ability to conduct research among the general population. As a result, the use of permission-based panels will likely increase. However, Ray questioned how ethical it is to report a 60% response rate for a panel survey, when the panel includes fewer than 1% of the population. Ray suggested we spend more resources training our staff in professional standards and to act ethically. He stressed that we consider the consequences of our research, and whether each project we undertake truly benefits respondents/consumers. And, we shouldn't necessarily "do unto others as you would have them do unto you" because their preferences may not be the same!

A Structured Approach to Choosing and Using Web Samples (Theo Downes-LeGuin, Doxus): Theo pointed out that the use of online panels will continue to grow as representativeness of phone-based sampling deteriorates. Web-based research is increasing, and timelines for projects continue to shorten. He suggested that many researchers no longer distinguish sampling and data collection mode decisions—a web panel is a simultaneous choice of sample source and mode. As a result, the traditional process of balancing survey costs and errors as a basis for sampling decisions is compressed or eliminated. Theo reviewed the pros and cons of probability-based sampling vs. non-probability based sampling. He also spoke of snowball (referral sampling) to define a frame for hard-to-reach populations.

Theo suggested that we estimate the likely sources of error, and use this information to decide, separately, regarding sample source and data collection mode. We should find the appropriate balance among manageability, cost efficiency, and reduction of error. Practical questions to ask include: "If a frame-based approach is available, does it offer a compelling advantage?" "If an intercept-based approach is the only alternative, have we made the risks/pitfalls known to our clients?" "Can we combine more than one sampling approach, and then compare the results to guide future decisions?"

The Options Pricing Model: An Application of Best-Worst Measurement to Multi-Item Pricing* (Keith Chrzan, Maritz Research): Best/Worst (or MaxDiff Scaling) is a relatively new technique forwarded by Jordan Louviere and colleagues back in the early 1990s. Keith showed how this technique could be applied to estimate the relative demand and price sensitivity for automobile options (such as sunroof, anti-lock brakes, etc.). Best/Worst questionnaires generally show a subset of the possible options, and ask respondents to indicate which is the best within the set and which is the worst. Eleven automobile features were tested at four price points each, among 202 consumers.

Keith analyzed the data using HB analysis, which yielded individual-level utility parameters (on a common scale) for each of the options at each price point. This permits the researcher to forecast which options (and at which price levels) would be most accepted in the marketplace. Utilities were converted to relative probabilities of choice by taking the antilog (exponentiating). Correlation between model prediction and self-reported actual purchase of options on the most recently purchased vehicle was 0.92. This suggested that the client could use the best/worst data to reasonably project future sales for not-yet-offered options and determine appropriate prices for each.

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

Conjoint Analysis: How We Got Here and Where We Are: An Update (Joel Huber, Duke University): Joel traced the history of conjoint analysis, from its roots in rigorous psychometric theory, to its eventual practical application in market research. Initially, conjoint measurement was proposed as a method to develop interval-scaling of multiple items from ranked observations. Early researchers applied conjoint measurement with human respondents, but found that many of the axiomatic tests (such as independence) were consistently violated. Even so, practitioners saw that the new measurement technique offered significant value, especially due to market simulation capabilities.

Practitioners replaced the rankings (early card-sort conjoint) with ratings (and later, choices), and replaced the full factorial designs with highly fractionated ones. Joel suggests that conjoint analysis has worked well because the simplification that respondents do in the conjoint survey often mirrors choice in the marketplace. Conjoint reflects how individuals might choose, given full information and more experience in making choices.

Choice has dominated ratings-based conjoint lately due to a number of factors. It is argued to better relate to market behavior, it emphasizes the competitive context, it is better for dealing with price/cost, and people are willing/able to make choices about just about anything. Surprises over the years have been the power of market simulators to account for differential substitution, the success of HB in predicting individual choices, and the difficulties of finding adaptive designs that outperform orthogonal ones.

The "Importance" Question in ACA: Can It Be Omitted? (Chris King, Aaron Hill, and Bryan Orme, Sawtooth Software): For all the success of ACA (Adaptive Conjoint Analysis), a potential weakness has been the "self-explicated importance question" used in ACA's "priors." Sometimes, self-explicated importance questions are difficult for respondents to understand, and the responses typically show limited discrimination, with many ratings grouping at the high point of the scale. Historically, importance questions were needed to estimate final utility scores under

OLS, for potentially dozens of attributes. However, with the availability of HB estimation, the importance questions might not be needed to stabilize individual-level ACA parameters.

The authors conducted a test to see if the importance questions could be skipped. Respondents were randomly assigned to receive a version of ACA that either included or didn't include the Importance section. Respondents who didn't receive the Importance section completed an extra six Pairs questions, though the total interview time was over a minute shorter than the respondents who completed traditional ACA. The authors found that traditional ACA (with the Importance question) achieved slightly better hit rates (for holdout choice tasks), but the share predictions were better when importances were omitted (in favor of six more pairs). Also, the final utility results without the pairs showed more discrimination among attributes ("steeper" importances) and provided different information from the utilities using the self-explicated importance information. The authors concluded that the Importance question could, if using HB analysis, be omitted.

Scale Development with MaxDiffs: A Case Study (Luiz Sá Lucas, IDS-Interactive Data Systems) Luiz investigated how well MaxDiff scaling questionnaires could be used to quantify "seriousness of offenses" (crimes, such as stealing, committing violence, or even murder). A series of studies already available in the literature (from the 1960s and 1970s), used a questionnaire that asked respondents to answer using a ratio scale, where a "10" was assigned to the seriousness of bicycle theft. If an offense was perceived 10 times as bad as a bicycle theft, respondents were instructed to assign 100 points, etc. Using the data, researchers had developed a "power rule" that related the ratio of the seriousness of an offense to the scale of that offense.

Luiz found an excellent match between the scales he developed from MaxDiff, and the "power rule" relationships found in the previous studies. Luiz also investigated the use of latent class to develop segments of individuals, based on their MaxDiff judgments of the seriousness of offenses. He also collected attitudinal statements to classify people in different psychological segments, as suggested by previous research. The segmentation developed from MaxDiff seemed consistent with that in the literature. Luiz's case study lends additional credibility to the claim that MaxDiff results are both ratio scaled (after exponentiating the logit-scaled coefficients) and reliable. And, a MaxDiff questionnaire is probably easier for respondents to complete than the questionnaires employed by the previous authors.

Multicollinearity in CSAT Studies (Jane Tang & Jay Weiner, Ipsos-Insight): Jane and Jay began their presentation by demonstrating how multicollinearity fouls our ability to derive stable betas under multiple regression. With synthetic data, correlated independent variables lead to unstable estimates of, say, drivers of customer satisfaction. The authors turned their discussion to a family of analytical methods that do a better job dealing with multicollinearity in CSAT studies: Kruskal's Relative Importance, Shapley Value Regression, and Penalty & Reward Analysis. These methods investigate all possible combinations of independent variables, and derive the contribution to the model of each variable, measured by the difference in some measure of fit when the variable is included vs. not included.

For both simulated and real CSAT data cases, the authors analyzed bootstrap samples to demonstrate that these methods result in much more stable estimates of drivers of satisfaction than standard OLS or stepwise OLS. They concluded that when the objective is to establish relative importance, rather than forecasting the dependent variable, methods that take into consideration all possible combinations of the explanatory variables are much more robust. Moreover, the margin of victory for these techniques increased as sample size decreases.

Modeling Conceptually Complex Services: The Application of Discrete Choice Conjoint, Latent Class, and Hierarchical Bayes Analyses to Medical School Curriculum Redesign (Charles E. Cunningham, Ken Deal, Alan Neville, and Heather Miller, McMaster University): McMaster University recently conducted a study to determine how to improve the quality of its medical school program. Due to increased class sizes, continuing to offer small tutorial group sizes of 5 students was becoming more expensive. The authors designed a CBC study to investigate other enhancements to the curriculum that would result in greater student satisfaction overall, and compensate for planned increases in tutorial group size.

Using qualitative research, the authors developed a list of attributes important to the quality of the program. Fourteen attributes each on four levels were used in a web-based, partial-profile CBC interview. The authors employed Latent Class to investigate the preferences of segments of students. Two segments emerged: students whose preferences better aligned with McMaster's small group, problem-based tutorial curriculum, and a smaller group of students who seemed to favor a more traditional medical school program Based on market simulations, the authors made specific recommendations regarding lower-cost (but significantly preferred) options that could improve the existing curriculum, despite an increase in tutorial group size. The results also underscored the need to identify prospective students during the admission process that are a better fit with McMaster's curriculum.

Over-Estimation in Market Simulations: An Explanation and Solution to the Problem with Particular Reference to the Pharmaceutical Industry (Adrian Vickers, Phil Mellor, and Roger Brice, Adelphi International Research, UK): Conjoint analysis is often used in pharmaceutical research. Due to the nature of product development in pharma, there is a strong demand for conjoint methods to deliver share predictions, in addition to utility estimates, sometimes three to five years prior to launch. The authors indicated that they typically see overestimation of share for new product entries using conjoint analysis. The overstatement of shares can often be 2x to 3x higher than market knowledge would suggest.

The authors described their typical questionnaires, including how they strive to establish a realistic setting (asking the physician to consider a specific patient when completing choice tasks), the use of partial profiles if more than six attributes, and HB estimation to obtain individual-level estimates. In addition to those standard procedures, they establish a cut-off threshold, applied at the individual level, and external effects, also applied at the individual level, to reflect other market realities for releasing a new drug. The full model takes into account any reluctance the physician may have about prescribing a new drug and also the volume restriction that results from third party payers and/or a physician's own consideration of what is a "fair" allocation among available product options. The amount of reduction depends upon such things as how serious and common the condition is, and the competitiveness of the market. The model still assumes 100% awareness of the new product, and the authors believe it represents an important step closer to predicting a realistic market share.

Estimating Preferences for Product Bundles vs. *a la carte* Choices (David Bakken and Megan Kaiser Bond, Harris Interactive): Bundling is a common pricing strategy which under many conditions can yield greater overall revenues than by selling the components separately (*a la carte*). Common examples include "value meals" at fast-food restaurants. "Mixed bundling" involves offering buyers a choice between bundles versus buying the components *a la carte*. However, standard choice-based conjoint approaches do not capture the choice process for mixed bundling strategies.

Based on a real client need, David and Megan designed a more complex CBC-like choice task that incorporated the notion of mixed bundling. Respondents (web-based survey) were instructed that they could either purchase the components as a bundle from a single manufacturer, or they could purchase the components separately from different manufacturers. Such a complex task required careful questionnaire design, pre-test, and detailed explanations and examples. The key to their solution was to develop two types of models from the data: a set of part worths predicting purchases of the bundles (including a coefficient for non-purchase of bundle), and a set of part worths predicting selections of each *a la carte* item given rejection of the bundle. The authors built a spreadsheet simulator that included a simple "if" condition to determine whether the shares of preference for each individual would be captured by the bundled alternatives or the *a la carte* alternatives.

The Importance of Shelf Presentation in Choice-Based Conjoint Studies (Greg Rogers, Procter and Gamble, and Tim Renken, Coulter/Renken): The grid-based approach and the shelf display are two common layouts for studying packaged goods using CBC. In the grid-based approach, a subset (often 9 to 12) of SKUs (graphics or text-only) are displayed on the screen. With the shelf-based display, all SKUs (often 20 to 30) under study are shown in each choice task, represented graphically on rows that look very much like store shelves. Shelf display seems to more accurately reflect the purchase experience, both in terms of available products and the general look of the environment. Greg and Tim hypothesized that if the display of products in the choice task were more like a store shelf layout, respondents would behave more like they really do when they shop.

The authors fielded a CBC study using the different layout approaches in CBC, covering multiple product categories. The criteria for success were: 1) ability to capture the same price elasticity estimates as are obtained using an IRI Marketing Mix Model (regression-based model using actual scanner sales data), and 2) ability to predict actual market shares. They found that the shelf layout provided slightly better fit to econometric models of price sensitivity, but that the grid layout provided slightly better fit to share. Importantly, the authors found that the estimates of price sensitivity from CBC were on average unbiased with respect to the scanner data models, though they often missed by a significant margin when any one product or brand was considered. They concluded that both exercises yield similar results, though the shelf display seems to have greater face validity, and is therefore easier to sell to clients.

The Effect of Design Decisions on Business Decision-Making (Curtis Frazier and Urszula Jones, Millward Brown-IntelliQuest): There has been a significant amount of research done probing the strengths of different flavors of discrete choice and methods for estimating part worths. Most studies have focused on hit rates and prediction accuracy for holdout shares. Curtis and Urszula focused their research on how different design decisions for discrete choice studies might affect the outcome in terms of business decision-making. The authors fielded a discrete choice study among 2400 respondents to a web-based survey. Experimental treatments were: full profile vs. partial profile; and no "None" alternative vs. including a "None" alternative vs. using a follow-up "None" question to each task. The research spanned three different product categories: Digital Cameras, HDTV and MP3 Players.

The authors found that partial profile tended to dampen the relative importance of price and increase the importance of brand, relative to full-profile. Including a "None" concept (either within the task, or as a second-stage question following the choice task) tended to increase the relative importance of price. To test how these differences might affect business decisions, the authors created hundreds of potential simulation scenarios, and used the part worths from the various treatments to determine "optimal price" points to maximize revenue for a client's hypothetical offering. As suggested by the findings with regard to "importances," partial profile designs lead to a significantly higher optimal price point. Including a "None" option in the questionnaire yielded the lowest derived optimal price points. Also, asking the "None" as a separate follow-up question produced much higher overall "None" usage, relative to when "None" was included in the choice task. Curtis and Urszula hypothesized that when "None" is included in the choice task, respondents may wish to appear cooperative by avoiding use of the "None." The authors concluded that different design decisions often have modest effect on holdout hit rates and share prediction accuracy, but can have a much bigger impact on business decisions, such as finding the right price points and projecting overall demand by relying on the scaling of the "None."

Application of Latent Class Models to Food Product Development: A Case Study (Richard Popper, Jeff Kroll, Peryam & Kroll Research Corporation, and Jay Magidson, Statistical Innovations): Food manufacturers need to understand the taste preferences of their target consumers, but taste preferences are often not homogeneous. Preference segments exist, and recognizing these differences may lead to better products that appeal to different segments and increase overall sales. The authors studied crackers using 18 flavor attributes, 20 texture attributes, and 14 appearance attributes. A trained sensory panel of 8 individuals rated the fifteen test crackers on the various attributes using 15-point intensity scales. The average ratings from the sensory panel were used as independent variables. 157 respondents rated all 15 crackers (dependent variable) over a period of three days. A completely randomized block design balanced for the effects of day, serving position, and carry-over.

Different models were used to detect consumer segments according to their liking ratings for the crackers. Four main models were tested: Latent Class (LC) Cluster model (nominal factor), LC Factor model (discrete factors), LC Regression model with a random intercept (nominal factor + one continuous factor) and a parsimonious non-LC regression model (two continuous factors). Latent GOLD software was used. The authors concluded that there was clear evidence of segment differences in consumers' liking ratings. Respondents reacted similarly to the variations in flavor and texture, but differed with regard to how they reacted to the products' appearance. Many other details regarding the relative strengths of the different models are covered in the full written paper.

Assessing the Impact of Emotional Connections (Paul Curran, Greg Heist, Wai-Kwan Li, Camille Nicita, Bill Thomas, Gongos and Associates, Inc.) The authors suggested that more advertisers these days are relying on forging emotional connections with their audience rather than relying principally on value propositions. The feeling is that in some markets, emotional connections are greater drivers of purchase and loyalty than other influences (such as utilitarian). The challenge faced by the researchers was how to quantify consumers' emotional connections with specific brands of automobiles. The authors conducted a great deal of qualitative research prior to designing the final quantitative instrument. The qualitative stage involved asking respondents to bring a collage of images and/or words to the interview that illustrated their feelings toward an ideal brand/product. Based on the preliminary qualitative effort, the researchers assembled characteristic images and words into vignettes that could capture specific emotional values. These stimuli were then used in a discrete choice exercise.

The authors felt that a key aspect to the research was a "negative priming" exercise. Based on previous research in psychology, the idea was to mentally overload and distract respondents in order to *hinder* overly-conscious thinking. After the negative priming exercise, respondents completed a discrete choice task involving eight emotional drivers on two dimensions for each vehicle: "How do you want to feel about a vehicle?" (Importance) and "Which do you associate with the <brand>?" (Brand Association). The authors examined the data by respondent segments based on automobile ownership. The most important emotional drivers were "peace of mind" and "smart and practical." A perceptual map (using correspondence analysis) displayed the results of the brand associations. VW was strongest on "fun to own" and "happy and carefree," Honda owned "peace of mind," and Saturn was strongly associated with "care for others." The authors concluded that "Independent or self-reliant" represented a positioning opportunity that no one vehicle measured currently fills.

Item Response Theory (IRT) Models: Basics, and Marketing Applications (Lynd Bacon & Jean Durall, LBA Ltd., and Peter Lenk, University of Michigan): Item Response Theory (IRT), also called *latent trait models*, originated in the educational testing literature to measure jointly subjects' latent traits or abilities and test item difficulty. Lynd, Jean and Peter described the simplest IRT model, the Rasch model, which assumes that a subject's performance on a test is determined by his or her latent trait and the difficulty of the test items. The latent trait is a random effect that varies across subjects, and item difficulty is a fixed effect for each item. The authors point out that the Rasch model is strongly related to CBC hierarchical Bayes models that are derived from random utility theory (RUM).

The authors suggested that IRT provides a rich framework for test item construction, which may have potential in marketing research. Test items can be characterized by their two parameters: discrimination and difficulty. Both of these parameters are combined in the item information function (IIF), which summarizes how much information an item has in estimating the latent trait for different values of the latent trait. One can easily imagine developing a large bank of marketing research items for different concepts, such as loyalty and satisfaction, where the items are indexed by their IIF. A marketing researcher could then select items from these banks to construct survey instruments for various purposes, such as studying highly loyal customers or dissatisfied customers. IRT enables the design of adaptive, online surveys. After obtaining an initial estimate of a subject's latent trait, an adaptive survey might select items to better estimate the trait with fewer responses. Instead of using a "shot-gun" approach to survey design, marketing researchers could be more strategic and systematic by employing the IRT framework.

Avoiding IIA Meltdown: Choice Modeling with Many Alternatives (Greg Allenby, Ohio State University, Jeff Brazell, The Modellers, Tim Gilbride, University of Notre Dame, and Thomas Otter, Ohio State University): IIA (Independence from Irrelevant Alternatives) is a property of logit models that can sometimes pose problems for market simulations-especially where there are many choice alternatives being studied. Greg and his co-authors investigated a new model using data from a discrete choice project involving over 1000 different automobile concepts. In typical choice modeling, a unique error term is associated with each alternative, and when the choice set is large, adding a near-duplicate offering results in nearly a doubling of predicted share. The authors' solution involved restricting the error space (the number of unique error terms) by assigning the same error realization to choice alternatives that share common important attributes (e.g., brand name). This avoids the "proportional draw" property of the logit model for these alternatives, resulting in a model with more reasonable predictive properties.

Predictive accuracy of actual market shares for automobiles showed a small improvement for the error-restricted models relative to traditional error specifications, despite the restrictive assumptions made about the error terms. The authors also illustrated their solution using a packaged-goods problem.

A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs) (Rich Johnson, Sawtooth Software, Joel Huber, Duke University, and Bryan Orme, Sawtooth Software): This presentation featured a second trial of a new adaptive design method for choice-based conjoint questionnaires. In a previous conference, Rich and his co-authors reported improved predictions of holdout choice shares (compared to standard CBC) when using customized (adaptive) experimental designs. The adaptive algorithm creates a unique design for each respondent, where the questions are chosen to maximize D-efficiency. Because part worths affect D-efficiency (utility balance yields greater statistical efficiency), preliminary estimates of part worths are needed for each respondent. In previous research, ACA-like self-explicated priors were used. In the current research, the authors dropped the self-explicated attribute importances and used only within-attribute level ratings. They also investigated both full-profile and partial-profile questionnaire formats.

1009 respondents completed a web-based study, randomly receiving one of four questionnaires (ACBC vs. CBC crossed by full profile vs. partial-profile). The hit rates and share predictions for the ACBC vs. CBC treatment were similar. However, holdout predictions were better for full profile than partial profile when predicting choices of full-profile holdouts, probably due in part to methods bias. The authors were puzzled why ACBC didn't perform better. Upon further investigation, they discovered that the adaptive designs were indeed about twice as efficient with respect to the *information they were given* (initial self-explicated part worths). However, the self-explicated information used to generate the designs was not accurate enough to produce efficient designs with respect to the final CBC-derived part worths. Omitting the "importance" question was apparently a mistake. The authors suggested that rather than depending on self-explicated information, a better method in future research might include using part worth information from previous respondents (through HB or Latent Class) as initial estimates, and updating those estimates after each respondent answer.

DATA COLLECTION ISSUES

IT'S ETHICAL JIM, BUT NOT IN THE WAY WE USED TO KNOW IT!

RAY POYNTER
UK

INTRODUCTION

Throughout the 70s, 80s, and early 90s, there was a growing, global homogenisation of what constituted ethical standards in marketing research. The key elements were: protection of anonymity, good procedures in terms of data files, membership of national and international bodies, the creation of professional standards and codes, and respect for statistical approaches that had been borrowed from established authorities and which had seemingly stood the test of time.

However, over the last 10 years this gentle and positive trend has been turned on its head, and replaced by a cacophony of change, fueling heated debates about what is, and what is not ethical. Across the globe we have seen Governments, media companies, advertising agencies, consumer groups, and populists attacking the methods, principles, objectivity, and ethics of marketing research. The two clearest consequences of this change in climate for marketing research are: the increased number of laws which regulate or restrict what researchers can do (for example Data Privacy, Do Not Call, and Anti-Spam laws), and the decline in co-operation rates.

This paper looks at this issue of ethics from a variety of angles, and recommends an adaptable and forward looking approach to deal with the changes that are already with us, and those that will arrive shortly. The paper is divided into 3 sections:

1. The drivers of change; an examination of the elements that are creating the change that is engulfing marketing research;
2. The respondents' perspective; a view of what respondents say they want from the research process, drawing on multi-country research conducted by the author;
3. Ethical strategies for the modern researcher.

DRIVERS OF CHANGE

There are many reasons why the context of ethics for marketing research has changed radically over the last few years, and why it will continue to change. Indeed, it is because the reasons for change as so numerous and diverse that we can be sure the process of change will continue—there is no likely, remedial action that will turn things back to the way they used to be.

The drivers of change can be broadly divided into exogenous (those from outside the realm of marketing research) and endogenous (those from inside the world of marketing research).

Exogenous Drivers

The exogenous drivers are largely beyond the control of marketing researchers, but their impact is immense.

Corporate Corruption

Since the dotcom collapse, and the spectacular corruption alleged at companies such as Enron, WorldCom, Arthur Anderson, and Parmalat there has been a growing feeling that the large corporations are not to be trusted. Much of the research conducted by the marketing research industry is for and on behalf of brands from the big corporations, something which does not improve our standing with potential respondents.

Consumer Awareness

Consumers are much more aware of how companies create and maintain brands. Just about every aspect of brand building and marketing has been reported in the news, debated on the radio, and done to death by television. As a consequence, potential respondents know more about what we are asking, why we are asking it, and what our clients are likely to do with their answers. Indeed, they often assume we will do more with their answers than we actually do (many assume their responses will enter an identified database, and that they will receive personalised sales propositions).

The traditional marketing research assumption that we are interviewing naïve respondents is not accurate; the modern respondent is complicit in the research process.

Anti-globalisation, Anti-GMOs, Anti Almost Anything!

If the zeitgeist of the twentieth century was ‘isms’ (Marxism, Fascism, Keynesism, Monetarism, etc), then the growing spirit of the times is ‘anti’. Groups campaign against Free Trade, against abortion, against the Euro and increased European integration. The impact on marketing research is that it makes more people suspicious of what we are and what we do.

Death of Respect

Thirty years ago bankers, doctors, lawyers, scientists, and judges were all held in high respect, their expertise was recognised and deferred to. However, this respect has been eroded by the exposure of mistakes and scams. From the blasé claims of the British Government about mad cow disease, through to the shuttle mistakes of NASA, to the collapse of fund managers run by Nobel prize laureates, the public’s confidence in, and hence their respect for, experts has been firmly dented. The impact on marketing research is that our claims about our ‘scientific methods’ and our independence are given less and less credibility.

Spin, Spin, and More Spin

As politicians, sports team managers, media stars, and even captains of industry increasingly replace facts and news with spin, the public become less and less inclined to believe the things they are told. Again, the impact on marketing research is that our claims of scientific method, independence, and trustworthiness are all tarred with the same brush.

Endogenous Change

As well as the changes which affect marketing research from outside, there are a wide number of changes which originate from within the world of marketing research.

Increased Volume of Marketing Research

The volume of marketing research has grown and grown, which means more researchers are chasing an ever reducing number of respondents (ie those citizens still prepared to be interviewed). Indeed, there is a view that in the US, 50% of all responses in marketing research come from just 4% of the population.

The main consequence of this change is that respondents are much more aware of the research process than we often allow for. People on access panels are often replying to multiple studies per month (partly facilitated by the fact that they are frequently members of more than one access panel).

This phenomenon raises two issues:

a) we should not assume that a respondent is not aware of the technique we are attempting to use with him or her;

b) if we annoy a respondent (for example by using poor survey design or by using an interview that is too long) we effectively reduce the pool of people available for future research.

Increase in Direct Marketing

The increase in Direct Marketing (using mail, telephone, email, and even text messaging) has mirrored and exceeded the growth in marketing research. This has resulted in more phone calls, more junk mail, more online contact, exacerbating the problems created by an increase in marketing research—more and more people are simply fed up with being interrupted. As a consequence we see a continued decline in cooperation rates, and increased calls for legislation to protect the peace and quiet of the citizen.

Convergence of Direct Marketing and Marketing Research

Many clients have, for a long time, wanted to combine their marketing research data with their marketing data in a way that would improve one-to-one marketing. However, the growth of the Internet as a research medium has acted as a catalyst to push this process along.

There are two major forces at play in the convergence of marketing research and Direct Marketing:

a) the people, client-side, who ‘own’ the databases are not the warm and friendly brand managers, they tend to be hard-headed e-commerce, CRM, or database managers—and often consider marketing research that does not enrich the data to be a wasted opportunity;

b) the Internet has reduced the barriers to entry of the ‘survey business’, and we are seeing CRM companies, website companies, e-commerce companies, and database companies pitching to our clients, offering low prices for inferior work, but work which can be integrated with the client’s databases.

In the UK, the Market Research Society has responded by creating a new category of research (called Category 6) which effectively permits the use of marketing research to build marketing databases. The question for marketing research professionals is how far can we, and should we, amend out procedures to allow us to move into this database building space.

Falling Response Rates

Response rates have been falling for years, but have now reached startling levels, according to the US organisation CMOR, response rates for RDD telephone calls in 2004 is about 11%. When we are using our conventional sampling statistics, we are using techniques that need response rates to be in excess of 70%—but do we make this clear to our clients? Indeed, are we clear about why we are using sampling theory based on high response rates with our low response reality.

Sensitive Uses of Marketing Research

Marketing research is used for a variety of sensitive purposes, for example: mystery shopping can result in significant consequence for the staff of branches visited, customer satisfaction surveys are frequently used as part of the bonus or compensation calculations of staff, TV ratings research is used to assess the viewing figures of different TV stations—which in turn is used to set advertising rates. These uses all raise ethical problems, as has been recently highlighted by the furore over the viewing figures for minority programmes in the US.

SPECIFIC ISSUES

In order to highlight the changes and challenges that the issue of ethics raises, this section will highlight a range of specific issues. It should be noted that this section is not intended to be an exhaustive list, but simply examples and indicators.

No Call, No Spam

Across the Western world there has been a plethora of new legislation introducing and extending restrictions on the use of the telephone and the use of email (a process sometimes referred to as legislative creep). These new laws tend to take one of two forms, the opt-out form (which tends to be preferred by the US), and opt-in (which tends to be the preferred form of the European Union).

Opt-out legislation is typified by the creation of lists, for example a Do Not Call List. The idea behind opt-out is that if somebody does not want to be contacted they can tell the caller, or some central list, that they do not wish to be called. Most current telephone legislation is based on opt-out.

Opt-in legislation is typically an attempt to limit contact to those who have already indicated they wish to be contacted. Much email related legislation is based on the presumption that emails should only be sent to people who have indicated they wish to receive them.

The growth of this sort of legislation (opt-in and opt-out) has tended to increase the cost of research and is another driver forcing research toward the use of access panels.

Marketing research organisations, such as ESOMAR and CMOR, have lobbied heavily to try to exclude marketing research from the reach of this legislation. However, this raises the question of whether we should be seeking to contact people who have said they do not wish to be contacted. And it raises the practical question about how reliable is the data elicited from people who did not wish to do the survey who have either been bribed (incentivised) or persuaded (rung/emailed repeatedly) to do the survey.

Invisible Processing

In traditional research, most of the data that were collected resulted from responses given willingly by respondents. Exceptions might be things like sex and ethnicity, which may have been entered directly by the interviewer.

However, when the Internet is used to collect data there are a wide range of data that can be collected without the respondent being aware of the process—these tend to be generically referred to as invisible processing. Examples of invisible processing include:

- Cookies, to track repeat visits;
- Web bugs, to track whether certain actions have been completed;
- Querying the browser to find out browser, screen height, operating system etcetera;
- Querying the link to establish IP address, ISP, referring URL etcetera.

To what extent should the respondent be alerted to the collection of this invisible data? Failure to indicate that invisible data are being collected is contrary to international guidelines (such as the ESOMAR guidelines) and is not consonant with the concept of respondent permission. However, attempting to list and explain all the invisible processing that might be going on may confuse and potentially alarm respondents, raising ethical and practical issues.

Improper Use of Marketing Research

Alongside the general growth of marketing research, there has been a growth in the 'improper' use of marketing research. Within the marketing research profession the chief area of concern is SUGGING, or selling under the guise of research. In SUGGING the citizen is exposed to something which starts as a questionnaire, but whose intention is actually to sell some service or product. This process alienates respondents and means they are less likely to take part in real research.

Another use, generally considered improper, is push polling. In a push poll the citizen is contacted, often by telephone, and again the process starts as if it were a conventional piece of opinion polling. However, contentious and potentially inaccurate prompts are then used to try and change the mind of the respondent. For example, the interviewer might ask "Does the news that some people believe X is a bigamist make you more or less likely to vote for him?".

Clearly the growth of these improper uses will reduce respondent co-operation rates, and increase the call for additional, restrictive legislation.

Disability Access

One specific trend in legislation is worthy of being separately highlighted, namely that of disability access. Over the last decade there has been a growth in legislation enshrining the rights of those who need special help to ensure access to services, for example people who are blind or who use wheelchairs. Some of this legislation will have a profound impact on marketing research, particularly research conducted via the Internet.

Many countries have passed legislation stating that the Internet needs to be accessible to people with specific needs, for example those who are colour blind, those who are partially sighted or blind, those with learning or cognitive problems, and those with problems in accessing the keyboard. One example of this sort of legislation is the US Americans with Disabilities Act, usually referred to as the 508 regulations. The 508 regulations insist that Government websites (and surveys on Government websites) are ‘accessible’.

One particularly demanding piece of legislation is the UK Disability Discrimination Act (1995), which came fully into force in October 2004. This law defines all websites as public places, and requires these websites to at least meet the WC3’s A standard for accessibility (which can be accessed via http://www.w3.org/WAI/). Surveys conducted on websites must also meet these standards.

The Respondent’s Perspective

Historically, many of the rules and shibboleths guiding the custom and practice of marketing research have been created through a debate conducted entirely within the marketing research community. This is why many of the rules are based, in a loose way, on the conventions that govern areas such as anthropology and psychology. More recently, there has been a realisation that the perspectives and wishes of the respondents need to be considered.

As part of this process the author, in conjunction with UK agency Virtual Surveys, has conducted research into what the respondents believe, and what the respondents want.

How Best to Contact Respondents

Most of the laws, and much of the debate, about how to contact respondents has concentrated on telephone and email. In Europe, for example, the assumption has often been that the gold standard for marketing research is still door-to-door interviewing, conducted with a randomised sampling design. However, when a sample of online respondents was asked how they wanted to be contacted, this view was sharply contradicted.

	UK	Germany	USA
Base	*258*	*257*	*768*
In the street	29%	28%	15%
At your door	8%	6%	7%
Phone you at home	8%	18%	12%
On your mobile phone	5%	4%	3%
By post	57%	45%	54%
By email	43%	53%	45%

Source: Virtual Surveys, October 2002, collected via RAWI.

As the table above shows, the public (or rather those of them who respond to an online survey) do not wish to be contacted at the door, nor in the US do they wish to be contacted in the street. The most popular methods of being contacted are post and email—perhaps because they are the easiest to ignore.

Who do Respondents Trust with Their Data?

The table below asked “Which of the following organizations do you trust to look after your personal information?”

	UK	**Germany**	**USA**
Base	*304*	*353*	*978*
Official Market Research Societies	7%	19%	2%
The Government	13%	10%	16%
Microsoft	16%	12%	11%

Source: Virtual Surveys, RAWI, October 2002

The survey included a longer range of organisation, but the extract above uses the Government and Microsoft as reference points. One point of interest is the very different figures for Marketing Research Societies in Germany and the USA. This may reflect the greater adherence to rules and procedures in Germany.

Protecting Anonymity

For many years one of the bedrocks of marketing research as been that the respondent's anonymity should be protected. The two tables below asked what people thought happened and what they wanted to happen.

	UK	**Germany**	**USA**
Base	*270*	*260*	*811*
Your name remains confidential & only your views passed to the company who asked for the research.	32%	50%	20%
Your secrecy is compromised (Net)	54%	39%	61%
Don't know	14%	11%	19%

Source: Virtual Surveys, RAWI, October 2002

The question, asked of those who had agreed to complete the survey, was "When you are interviewed by a proper marketing research company, operating under official rules, what do you think happens to your personal details?".

The table suggests that even amongst those who are happy to take part in the survey, a majority believe their anonymity is not being protected. Presumably, those who declined to take part in the survey had views that were at least as negative.

The table below is based on the UK only, and asked what people wanted to happen to their data.

Base, UK only	***1177***
Your data & your name to be given to the company paying for the research so they can try & contact you.	3%
Your name and details to be sold to people creating marketing lists.	3%
Your name to remain confidential & only your views passed to the company who asked for the research.	57%
Asked each time if you want personal details & comments passed to company paying for the research.	34%

Source: Virtual Surveys, RAWI, April 2003, UK only

When asked what the respondents wanted to happen to their data, a majority wanted their name to remain confidential, and their views to be processed in a way that protected their anonymity. However, a third of respondents were happy for their personal details to be used, provided they were asked for permission.

This piece of research suggests that the growing trend of permission-based research is in accord with what some respondents' wishes.

Ethical Strategies and Questions

This paper has highlighted a changing world, and one where ethics is of increasing importance. This section sets out some strategies, and then applies them to a few typical situations.

Permission-Based Research

Seth Godin has pioneered and popularised the concept of permission-based marketing. Permission-based research is based on similar concepts. Rather than being based on a provider-focused vision of what we can get away with, permission-based research is based on informed consent.

One very clear example of permission-based research is the Internet access panel. People sign-up to a panel knowingly, they can unsubscribe whenever they want, and they can decline individual surveys. Over a period of time the 'informed' part of their agreement becomes increasingly developed—which has implications for the research we are conducting with them.

There are several issues about adopting permission-based research, in particular:

- Mystery shopping;
- Invisible processing;
- Studies where letting the respondent know the details would change the results.

Professional Standards

Various professional organisations, for example ESOMAR and CASRO, have their own professional standards and guidelines. Researchers need to be aware of the relevant rules and engage in the updating of these rules in the context of new media and the changing ethical context.

One problem that the researcher faces when conducting international research is that the rules and guidelines tend to differ from one country to another. In a perfect world these rules would be harmonised, but in the real world it is necessary to be aware of these differences.

Staff Development

Most senior staff have acquired a knowledge of research ethics over the years of their employment, and by having to make the difficult decisions associated with their senior level. However, there can be a misguided assumption that this awareness and these views are shared throughout the organisation.

If a company wishes to ensure that it behaves in an ethical way it needs to ensure that the views and values are communicated throughout the organisation. This means training, documenting, and ideally engaging junior staff in discussion about the ethics of research.

The Consequences Test

One method of assisting researchers to survive in the ethical morass that is research in the current times is by using a technique that the author has termed the Consequences Test.

The first element of the Consequences Test is to assume that the key details of a project will at some stage enter the public domain. In the age of the Internet most information leaks out somewhere, sometime. Companies such as Microsoft, Enron, and Arthur Anderson have all found out that details and comments in emails from one employee to another can end up as incriminating evidence, and plastered across the media.

The second element is to consider the consequences of the proposed research on the relevant stakeholders. Indeed, the true second stage is the identification of the relevant stakeholders. Typical stakeholders include: the client, respondents, yourselves, consumers, employees of the client company, and shareholders of the client company. But the list of stakeholders can be wider, for example in the furore over television ratings in the US in 2004, viewers of television stations catering to minority groups were a relevant stakeholder group.

Having determined the likely and possible consequence of the research for the key stakeholder groups, the key ethical issues can be addressed.

Applying the Ethical Strategies

This section looks at applying these strategies to a cross-section of issues.

NPD Research

In most NPD research the focus of the research is to find out what consumers want and to see if new product ideas meet these wants. The interests of the respondent and the client company are often well aligned when conducting NPD research. This alignment of interest means that the ethical issues are few and are the ones that have been established for some time, including:

- Keeping the client's ideas confidential;
- Using techniques that are reliable, and ensuring that the client is aware of the probability that the market results will be different from the test results;
- Ensuring that any material used by the respondents and the interviewers is safe.

Pricing Research

In most pricing research the focus of the study is to find out how much the client can charge before demand is choked off. The interests of the client (to charge as much as the market will bear) and the consumer (to pay as little as is viable) are often not well aligned.

If one were to take the principle of informed consent at face value, we would need to explain to respondents that answering honestly could result in the prices of the products they want to buy being sold at a higher price. This would clearly run the risk of changing the respondent's responses.

The practical solution is to rely on Access Panels, whose members have entered into an agreement to respond to surveys in exchange for rewards and the chance to influence outcomes.

Their general permission could reasonably be assumed to include the specific for a piece of pricing research.

Customer Satisfaction Research

At its most straightforward level, customer satisfaction research can be used to enhance the service to clients. But there can be a number of concerns:

- Is there really a causative link between the variables being measured and customer satisfaction? If we doubt this, have we alerted clients to our concerns?
- What impact might the research have on the staff of a branch or franchise which is being evaluated, and does the client understand the reliability of the research at the branch level?
- If customer satisfaction scores are used to assess annual bonuses or compensation levels, are we sure that the statistics have sufficient reliability?

CONCLUSION

The context for ethics is changing faster and in more diverse ways than was the case for the past forty years. Increasingly, marketing researchers cannot assume that they are passive observers. Increasingly they must accept that they are part of the process.

By being aware of the ethical issues, by keeping up to date, and by applying the Consequences Test, researchers can ensure that they will stay positively engaged with the realities of marketing research and its connections with society and in particular its relationship with other parallel operations such as direct marketing.

Postscript

Do not do unto others as you would they should do unto you; their tastes may not be the same as yours. —George Bernard Shaw

A STRUCTURED APPROACH TO CHOOSING AND USING WEB SAMPLES

THEO DOWNES-LEGUIN
DOXUS

INTRODUCTION

Web surveys offer immediate gratification for researchers in the form of cost savings, ease of recruitment and, in most cases, immediacy of response. For this reason, we believe that use of the web as a data collection method will only increase with time, as will the attendant web sampling approaches—currently dominated by panels and non-probability list sources.

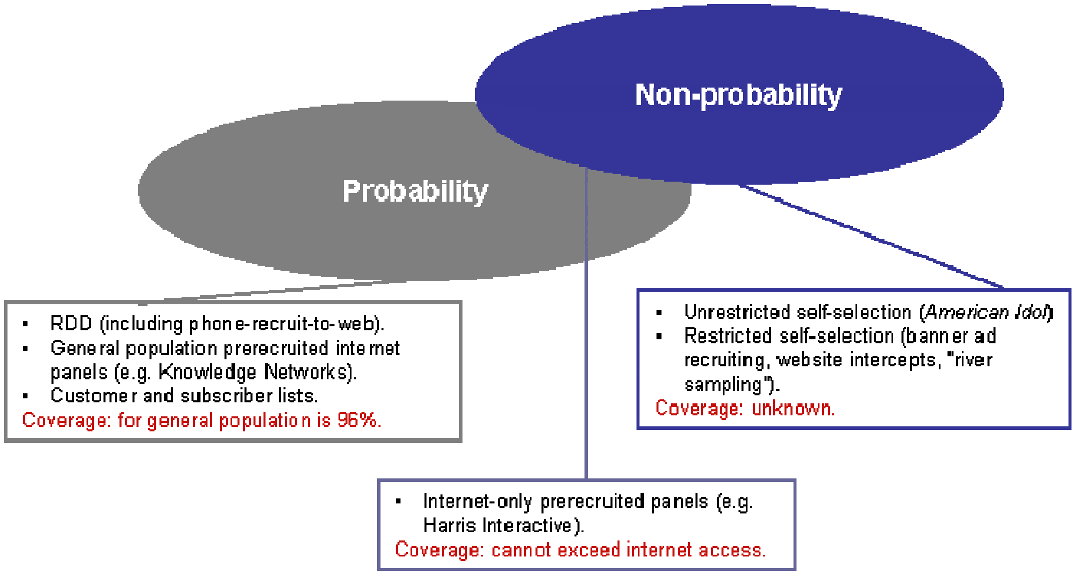
One of the most profound effects of the surge in web-based data collection is that many researchers and their clients no longer draw a crisp distinction between sampling and data collection mode decisions. For example, the decision to use a web panel is both a sampling and a mode choice. In almost all cases a web panel is not used without choosing web survey as the mode (though phone surveys from web panels are possible). And in many cases, because of the lack of traditional sampling frames and sources for the web, you also can't choose web as the data collection mode without also choosing a panel as a sample source.

Sampling source and data collection mode have, of course, always been inextricably linked—and advances in data collection technology have traditionally forced practical advances in sampling. The development of the frames from which random-digit dial (RDD) samples are drawn was borne of the obvious appeal of the telephone as a replacement for face-to-face interviewing and the obvious shortcomings of the common commercial sampling practices of the day.

We can look forward to the day when internet access is near-universal and a relatively stable sampling unit similar to phone numbers is available for the web. But for the foreseeable future, email and IP addresses are not viable means for traditional sampling approaches. Both are inherently flexible, non-unique means of identification that can identify people, places or devices that change constantly. They lack any universal authority or organizing body and population database aggregators (the role that the Bureau of the Census has played for geographic/in-person sampling frames and that the Bell system played in the early days of telephone frame development). Given the inherently decentralized quality of the internet, we may never see anything approach an internet sampling frame in the traditional sense that researchers use it for general population surveys.

TOWARDS A TAXONOMY FOR WEB SAMPLE SOURCES

Even though web research offers an even mixture of opportunity and anarchy, however, we can still make intelligent choices about how to sample. Clients and researchers need to know which kinds of samples are available on the web in order to make an active, rather than passive choice. In a useful paper, Couper (2000) created a taxonomy of web samples which we have adapted below. Web samples fall into two broad approaches, probability-based and non-probability based.



- RDD (including phone-recruit-to-web).
- General population prerecruited internet panels (e.g. Knowledge Networks).
- Customer and subscriber lists.

Coverage: for general population is 96%.

- Unrestricted self-selection (*American Idol*)
- Restricted self-selection (banner ad recruiting, website intercepts, "river sampling").

Coverage: unknown.

- Internet-only prerecruited panels (e.g. Harris Interactive).

Coverage: cannot exceed internet access.

Probability or frame-based web sample sources are those for which a frame exists before the sample is drawn, and from which the researcher *can* draw a probability sample in advance. Common examples include any internet panels based on probability sampling techniques and telephone recruits, such as Knowledge Networks, or self-contained databases specific to a certain population such as CRM (customer relationship management) databases or membership and subscriber lists. Although the population coverage of these sources is often unknown, they nevertheless represent a frame from which a probability sample can be drawn.

Non-probability sources are those for which no defined frame exists before the research is conducted. No probability sample is possible. This includes web site intercepts, email or web ads placed with the goal of research recruitment, and “river sampling” conducted for the purpose of single survey vs. panel participation. All of these approaches have in common the fact that we are going to places that we expect qualified respondents to hang out, intercepting them in the act of doing something else, and recruiting them to research. In other words, these approaches represent traditional convenience sampling similar to the approaches used for mall intercepts and focus group recruiting. For intercept- or advertising-based sources, the process of “drawing a sample” doesn’t apply because there is no defined frame, or even population, from which a sample can be drawn—instead, the researcher works with a profile of the target respondent and seeks to find that respondent in the most cost-efficient way possible. The trade-off for the cost efficiency of a non-probability, intercept-based web sample is that we are unable to define or measure sampling error or address the broad issue of whether the resulting sample is, broadly-speaking, truly representative.

All large, commercial internet-only pre-recruited web panels in the United States today are, strictly speaking, non-probability samples and in many regards are no different from a convenience sample. In our adaptation of Couper’s taxonomy, however, we position larger commercial web panels such as Harris Interactive’s Harris Poll Online (HPOL) in a “grey area” between probability and non-probability web sample sources. These panels seek to build a frame (using a mixture of probability and non-probability samples) from which a probability sample can be drawn. The basis for that frame may have unknown coverage of the population, but the panels are stable relative to intercept-based methods and one could in theory draw a probability sample from any such panel.

This fact notwithstanding, the fact that a frame exists doesn't mean that the sample drawn is "representative" or offers any advantage over a convenience sample. Most frame-based sampling sources actually rest all or in part on intercept-based methods. Harris, Greenfield and others typically rely on a combination of recruiting from RDD telephone surveys, list purchasing and website intercepts to build their panels. Since the source of the sample is never identified to research buyers, and customers cannot partition their sample requests by recruiting source (i.e., ask for a sample based only on RDD recruits), our ability to assess the sampling error is fundamentally no different from that of a website intercept. Despite sophisticated and very useful weighting processes that panel providers use, the basic selling point of most panels is not "we're a high quality sample by traditional definitions," but rather "the vast size of our panel means it must be representative."

Nevertheless, internet-only pre-recruited panel sources offer several attractive qualities for the researcher that lead Doxus to favor this approach in most of our work. One increasingly important benefit is the fact that privacy protection for respondents can be built into the frame construction. Such panels, for example, are typically multiple-opt-in sources, while website samples are essentially intrusive and usually only semi-opt-in. But the most common advantage of frame-based approaches is that, once we're willing to trust that a big panel must be a "good" panel, web panels allow us to more efficiently manage data collection and convert ad hoc contacts into research relationships.

An excellent example of this from business-to-business research is probability-proportional-to-company-size sampling. In this approach, we give a higher probability of selection to larger companies as a proxy for spending power. This addresses the concern that a "random" or interval sample of businesses from any high-coverage business source will produce a huge preponderance of small businesses. Of course, we could draw an interval sample, set quotas to over-represent larger businesses, and simply screen out all small businesses after a certain point—which is precisely the approach used in the approach most surveys web use. But from any angle, this is a very inelegant and inefficient approach. Quotas are statistically unattractive, expensive to manage, and place burden on the respondent to exclude himself based on responses rather than on the researcher to exclude segments of the population based on known characteristics. In a panel that has company size or revenue as a measure, we find it significantly smarter to take account of the appropriate measure of size in how we draw the sample. This allows us to talk to more people we care about and fewer that we don't.

AN APPLICATION OF SNOWBALL SAMPLING TO THE WEB

An interesting, and for internet research potentially relevant example of frame-based samples is snowball sampling in order to define a frame for a small or rare population. Like web panels, this method exists at the intersection of the two approaches we've defined. The term "snowball sampling" is typically used as a synonym for convenience sampling, and in particular for the form of convenience sampling (mostly discouraged, even in qualitative recruiting) of using one qualified respondent to locate another in order to fill out quotas for survey completion. In the survey research literature, however, snowball sampling or network sampling has been offered up as a potentially defensible approach to building a frame from which a sample can then be drawn.

The approach has been used (albeit infrequently) in studies of small, rare and mobile populations such as the homeless.

In this approach, the researcher accumulates qualified individuals who nominate others, snowballing from one referral to another in a manner similar to a multi-level marketing scheme. Once referrals become self-referential enough—that is, the same people get mentioned repeatedly as referrals—the researcher assumes that the population has been adequately defined and a frame is built (the literature offers some practical criteria for when to stop the referral or networking process and assume that the frame has been built). The frame presumably has less than 100% coverage of the population, but we can estimate the non-coverage and draw repeated samples from the frame with far more knowledge of the sources of survey error than a simple intercept sample would yield.

Within certain bounds, this approach is not only extensible to the web but in some ways perfectly suited. The internet is, by definition, a great networking tool: paired with email and messaging, it is the first communications tool that makes both private and broadcast communication available to anyone, and as the technology literature has amply demonstrated, it is subject to network effects (put simplistically, once it has a critical mass of users, growth of new users is exponential). The method could be used to build a high-quality sampling frame of a narrow population that might otherwise be hard to find without a privately-owned list (e.g., public school teachers in Des Moines).

Summary

Following are the practical questions we ask ourselves, and discuss with our clients, before choosing a web sampling approach:

1. Is a probability-based approach available? If so, does it offer a compelling advantage in terms of how we can draw the sample or manage the study? Frame-based alone is no guarantee of sample quality.
2. If an intercept-based approach is the only one available, have we made the potential pitfalls of the approach (without coming across as slope-shouldered) clear to our customers?
3. Can we combine more than one sampling approach (for example, two cells, one recruited and interviewed via an RDD telephone survey and one from an internet-only, pre-recruited web panel) and benefit from a comparison of the results to understand better—both for this project and for the future—which approach offers better manageability, cost efficiency and reduction of sampling error?

References

Biernacki, P., & Waldorf, D. 1981. “Snowball Sampling: Sampling and Techniques of Chain Referral Sampling.” Sociological Methods and Research. 10:141-1963.

Couper, M. P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*. 64(4), 464-494.

COMPONENTS-BASED PRICING RESEARCH

THE OPTIONS PRICING MODEL: AN APPLICATION OF BEST-WORST MEASUREMENT TO MULTI-ITEM PRICING

KEITH CHRZAN
MARITZ RESEARCH

INTRODUCTION

Marketing researchers often employ one or another type of conjoint analysis to model price sensitivity: typically price is one attribute among several and calculated utilities for its levels allow simulation and price sensitivity analyses. The model described below has this special property: it can be used to estimate separate price sensitivity curves for individual attributes.

BACKGROUND

Some history on maximum difference scaling and best-worst "conjoint" analysis will set the stage for the research problem and the proposed solution described below.

Maximum Difference Scaling

Developed by Finn and Louviere (1992), maximum difference scaling (maxdiff) has several benefits over traditional perceptual scaling:

- It is a general attitude scaling method – Sá Lucas, in this volume, illustrates its use in measuring the perceived severity of criminal offenses and as an alternative to a Likert agreement scale (Sá Lucas 2004)
- It is a more discriminating way to measure attribute importance than either rating scales or the method of paired comparisons (Cohen 2003)
- It has greater predictive validity as an importance measurement than either ratings scales or the method of paired comparisons (Cohen 2003)
- Like other choice-based methods, it prevents scale-use heterogeneity, making it ideal for cross-cultural studies, or as the basis for needs-based segmentation (Cohen 2004).

Maxdiff is the multinomial extension of the traditional method of paired comparisons (Thurstone 1927, David 1988). Whereas a paired comparison question asks a respondent to make a binary choice ("Would you rather have a TV with a flat screen or with a built-in DVD player?") maxdiff has the respondent specify "best" and "worst" choices from sets of three or more objects:

“Which of the following features would MOST make you want to buy the TV described below and which would LEAST make you want to buy it?

	Most	Least
Built-in DVD player	[ ]	[ ]
Flat screen	[ ]	[ ]
Cable-ready	[ ]	[ ]
5 year warranty	[ ]	[ ]
Made by Sony	[ ]	[ ]

Design strategies for maxdiff question sets range from using traditional orthogonal fractions of 2^n factorial designs to the Sawtooth Software B/W Designer (Sawtooth Software 2003) that creates balanced designs with fixed numbers of alternatives per question.

Analysis options include:

- Simple counting and arithmetic transforms when using perfectly balanced designs (Finn and Louviere 1992; Swait, Louviere and Anderson 1995)
- Aggregate MNL, including SAS and SPSS programs
- Latent class MNL with Latent Gold or Sawtooth Software’s CBC Latent Class Module
- HB MNL using Sawtooth Software’s CBC/HB product

An important twist in coding maxdiff questions is that one reverse-codes the design matrix for the “worst” choices (i.e. multiplies all design codes by -1) and then concatenates the best and worst data matrices into a single data set. Though this method produces good results in practice, it is theoretically unappealing, because it frequently happens that the utilities would be different if you ran separate “best” and “worst” models. Most practitioners just ignore this slight contradiction and combine the “best” and “worst” choices anyway.

Optionally, one could further simplify the design coding task by modeling the “best” choice as a multinomial choice among the several alternatives, and dividing the “worst” choice into a series of pairwise choices wherein each non-best and non-worst alternative is, in turn, shown to be preferred over the “worst” alternative. For CBC users, this means you can use standard .cho file coding rather than creating a user specified design coding matrix. A more complete description of coding options in CBC appears in the Appendix.

Best-Worst “Conjoint” Analysis

Swait, Louviere and Anderson (1995) describe how to extend maxdiff scaling to a conjoint-like application they call Best-Worst Conjoint Analysis or B-W. One could argue, since the technique produces no decomposition of the utility of multiattribute stimuli, that it is not a type of conjoint analysis at all.

Taxonomy aside, the respondent’s task in B-W is the same as in maxdiff, but the experimental design differs. Using an orthogonal main-effects design or an efficient design generated by a SAS design macro, one creates a set of profiles. Each profile becomes a separate choice question, where respondents choose the “best” and “worst” attribute/level combination in each question:

"Which of the following features would MOST make you want to buy the TV described below and which would LEAST make you want to buy it?

	Most	Least
Built-in DVD player: No	[ ]	[ ]
Flat screen: No	[ ]	[ ]
Cable-ready: Yes	[ ]	[ ]
Warranty: 5 year	[ ]	[ ]
Manufacturer: Sony	[ ]	[ ]
Price: $299	[ ]	[ ]

In direct comparisons with choice-based conjoint models containing the same attributes and levels,

- B-W experiments have been found to contain less respondent error (Swait, Louviere and Anderson 1995, Chrzan and Skrapits 1996)
- Mixed results occur as to whether or not B-W and choice-based conjoint produce equivalent part-worth utilities, after adjusting for the difference in respondent error, with one finding that they do (Swait, Louviere and Anderson 1995) and one that they do not (Chrzan and Skrapits 1996)
- B-W may have a slight edge in predictive validity, though the test that showed this was a weak one (Chrzan and Skrapits 1996)

A further advantage of B-W is that it puts all attribute levels on the *same* interval scale; this is different than choice-based conjoint, which puts all attributes' levels on different scales with different arbitrary origins. This property uniquely allows cross-attribute level comparisons, an advantage illustrated below.

The Options Pricing Model

Some products are sold with optional features that are available at additional cost. For example, many personal computer manufacturers sell their products online, with base models that can be customized with optional features and upgrades for incremental cost. Likewise, automobile manufacturers offer optional features, at separate additional costs.

In these cases, each optional feature has a separate price and, conceivably, a separate sensitivity to differences in price. A client wanting to understand the price sensitivity attributable to each feature may be ill-served by using a typical choice-based conjoint model:

Which of these cars would you rather buy?

Car A	Car B
Honda	Ford
Minivan	Sedan
$26,000	$34,000
Sunroof @ $400	Sunroof @ $350

In this case, the cost difference in the base prices swamps that attributable to the sunroof, making it difficult to measure the sunroof price adequately. Even leaving the base price and the other attributes out of the experiment, however, price sensitivity of individual options may be poorly measured:

Which of these cars would you rather buy?

Car A	Car B
Antilock brakes @ $200	Antilock brakes @ $400
Sunroof @ $600	Sunroof @ $300
CD Player @ $200	CD Player @ $400
Heated Seats @ $300	Heated Seats @ $200
Total: $1,300	Total: $1,300

Respondents would understandably be indifferent between the above packages. Moreover, the reality of the choice process to be measured is that respondents can pick and choose individual options, not packages as in this question. Perhaps better would be to do paired comparisons, like

Would you rather buy

[ ] A sunroof for $400, or
[ ] A CD changer for $350

The multiple choice extension of this is, of course, a type of B-W model we can call the Options Pricing Model (OPM).

In OPM, attributes' levels are their price points. Using an efficient experimental design, one can create a set of profiles to be evaluated with maxdiff scaling. If the number of attributes is less than about eight or ten, these may be full profile questions. Partial profile questions may be used when the number of attributes is larger. Analysis can be done to produce aggregate (MNL), segment (latent class MNL) or respondent level models (hierarchical Bayesian MNL).

CASE STUDY

Research Design and Analysis

The client needed to test price sensitivity for 11 options available on automobiles, each at 4 price points. Attributes and price points (partially disguised) were:

- **6 Disk In-Dash CD Changer** at $(600, 675, 750, 900)
- **Sunroof** at $(800, 900, 1,000, 1,200)
- **Anti-Lock Brakes** at $(400, 500, 600, 800)
- **MP3 Player** at $(410, 470, 530, 650)
- **Rear Side Airbags** at $(300, 350, 400, 500)
- **Heated Front Seats** at $(300, 350, 400, 500)
- **Fog Lights** at $(250, 300, 350, 450)
- **Cassette Player** at $(150, 200, 250, 350)
- **Keyless Entry** at $(210, 260, 310, 410)

- **Cruise Control** at $(135, 175, 215, 295)
- **100,000 Mile Powertrain Warranty** at $(900, 1,200, 1,500, 2,100)

202 household automobile purchase decision makers were qualified and surveyed using an internet panel sample source and a web-based interview.

Given the large number of attributes, we employed a partial profile design. Using an orthogonal fraction of a 2^{11} design, we settled on a 15 run design to determine which attributes would be present in each of 15 choice questions (the 16th was the null set with no attributes present). Order of questions was randomized, and levels were randomly assigned to each attribute, in each question, for each respondent. Across respondents this creates a near-orthogonal design. This would have been a poor design had the client wanted respondent-level utilities, because the design would be unbalanced and inefficient for individual respondents. An aggregate model fit the client's need (and budget) so this potential limitation was not injurious to the study's objectives.

Results

Utility Model

Two possibilities are that price could be generic (have the same utility for single dollar differences across all attributes) or attribute-specific. The utility of a single dollar differed across attributes by as much as a factor of 10, however, so the simple generic model described the data poorly and was abandoned. The resulting utility model appears in Table 1.

Table 1

	Utilities			
Attribute	Lowest Price	Mid-Low	Mid-High	High Price
CD(600, 675, 750, 900)	1.289	1.197	0.973	0.837
Sunroof(800, 900, 1,000, 1,200)	1.493	1.066	0.830	0.907
Brakes(400, 500, 600, 800)	1.351	1.242	1.055	0.648
MP3(410, 470, 530, 650)	1.266	0.994	1.002	1.034
Airbags(300, 350, 400, 500)	1.214	1.197	1.030	0.855
Seats(300, 350, 400, 500)	1.184	1.149	1.152	0.811
Lights(250, 300, 350, 450)	1.265	1.106	1.114	0.811
Cassette(150, 200, 250, 350)	1.173	1.173	1.107	0.843
Keyless (210, 260, 310, 410)	1.218	1.218	1.013	1.203
Cruise (135, 175, 215, 295)	1.203	1.297	1.039	0.757
Warranty (900, 1,200, 1,500, 2,100)	2.418	1.134	0.744	0.000

Note that the last level of the last attribute is the reference level—set to the arbitrary zero of the interval scale. All other levels of all attributes are measured on this same scale, with this same origin. This is different from all other forms of conjoint analysis, in which each attribute's levels are interval scales with separate origins. McFadden's ρ^2 was .217, suggesting a reasonably good model fit; put another way, the root likelihood of the aggregate model was .253, versus an expected likelihood of .175 from respondents choosing randomly.

Simulation

Because all attributes' levels are on the same interval scale, it makes sense to measure the relative appeal of two different options at two different price levels. For example, take the Anti-Lock Brakes at $400 (utility 1.35) and the MP3 Player at $650 (utility 1.034). From these we can use the standard MNL choice rule to calculate that demand for the MP3 Player at this price will be 42% that of the Anti-Lock Brakes at $400:

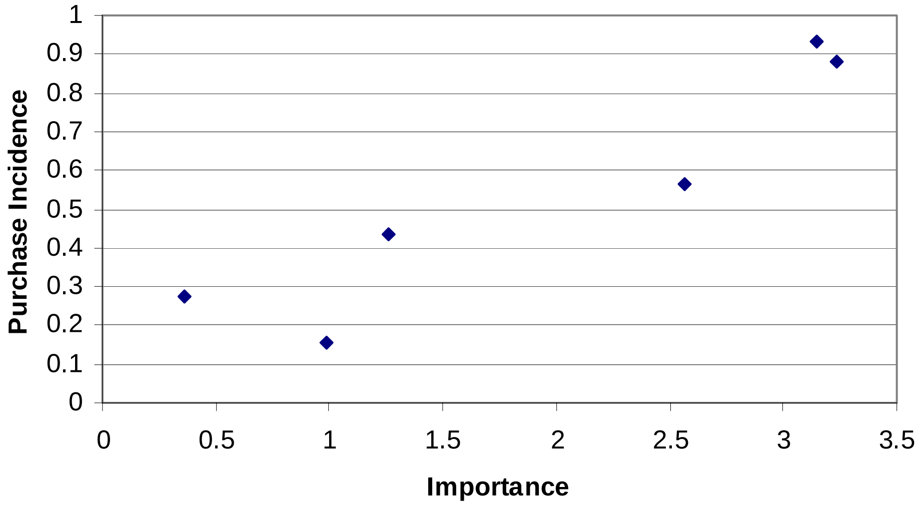
- Exponentiating the utility of the Anti-Lock Brakes yields $e^{1.35} = 3.857$
- Exponentiating the utility of the MP3 Player yields $e^{1.034} = 2.812$
- The logit choice rule would put the relative share for the MP3 player at 2.812/(2.812+3.857) = 42%

This allows the client to forecast demand for new features.

Validation

The utilities of existing options at existing price point should be related to the actual demand for those options. Prior to the B-W exercise, respondents reported which options they had added to their most recent vehicle purchases. The correlation between self-reported purchases and OPM utilities was 0.92. See the scatterplot in Figure 1

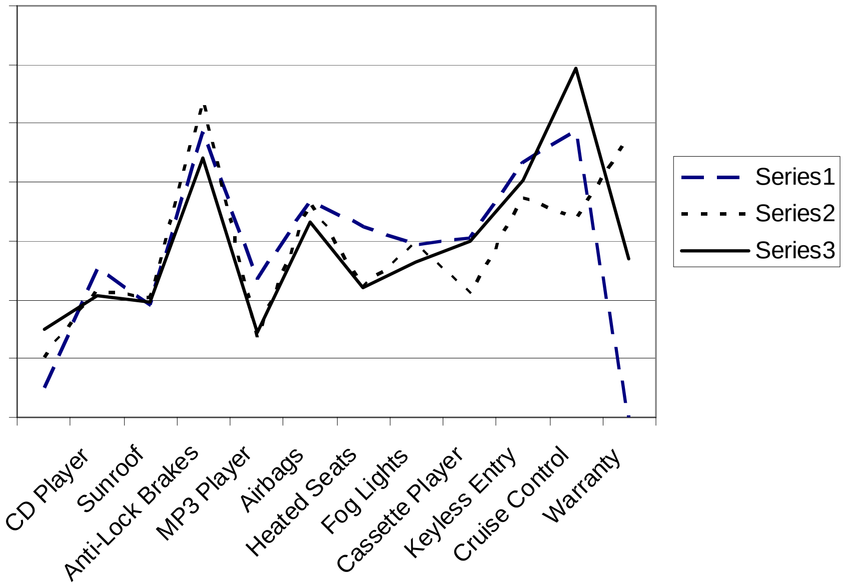
Figure 1



Segmentation

Using latent class MNL, or HB with cluster analysis, respondents could be grouped into segments. Though this study was not designed for respondent-level estimation, the client requested an exploratory segmentation, using cluster analysis of attribute utilities (rather than part-worth utilities). Since utilities were estimated using the second method described in the Appendix, attribute utilities were computed as the mean of the levels' utilities for each attribute. One solution that resulted appears in Figure 2, though it is not the solution the client actually used.

Figure 2



Segment 1 is most interested in heated seats, cruise control and airbags, and least interested in the warranty. Segment 2 seems more safety conscious: they are the least interested in entertainment attributes and most interested in anti-lock brakes and the warranty. Segment 3 showed the least interest in anti-lock brakes and the highest interest in the CD player and in cruise control.

Conclusion

Based on maxdiff scaling and more specifically on B-W conjoint analysis, the Options Pricing Model is a designed choice experiment for a specific pricing situation. OPM looks like a reasonable solution to this very specific price sensitivity modeling need.

Likely OPM could be extended to handle more general menu price problems like those that restaurants face, perhaps with separate OPM exercises for appetizers, soup and salad, entrees and desserts.

APPENDIX: MAKING A .CHO FILE

Some ways of organizing OPM data for analysis via Sawtooth Software's CBC utility estimation tools appear below.

One could use the Swait, Louviere and Anderson (1995) recommendation for coding to produce attribute utilities in addition to part worth utilities. For the case of the 4^{11} experiment, this would produce 10 = (11-1) attribute utilities and 33 = (11[4-1]) part worth utilities, for a total of 43 parameters. The ability of both CBC/HB and Sawtooth Software's Latent Class Module to handle user-specified design coding is vital for the effects coding and for the reverse coding of the "worst" levels used in this strategy.

An easier way is to consider each of the 44 levels to be separate objects and to code them as you would the objects in a standard maxdiff experiment: 43 dummy codes for 44 objects. Each attribute utility is just the mean of its levels' part worth utilities. This still requires user-specified design coding to reverse code the design matrix for the "worst" choices.

As the CBC/HB software is currently programmed, neither of these methods allows you to impose monotonicity constraints, a handy feature indeed for a pricing study. To do this you have to model all the instances of attribute/level combinations being better than others, and this involves breaking down each B-W question into several inequalities. For example, in a set of objects A – E, say a respondent chooses A to be the best and D to be the worst. In this case we recast these best and worst choices into four choice observations:

- Choose A from A – E
- Choose B from B & D
- Choose C from C & D
- Choose E from D & E

One can now use standard CBC .cho file attribute coding and a standard constraint (.con) file.

Another way to impose constraints is to save the HB draws and use the manual method of tying the draws and collapsing into a point estimate per individual, as described by Rich Johnson in his "Monotonicity Constraints" paper for HB (Johnson 2000).

References

Chrzan, Keith and Mike Skrapits (1996) "Best-Worst Conjoint Analysis: An Empirical Comparison with a Full Profile Choice-Based Conjoint Experiment," paper presented at the INFORMS Marketing Science Conference, Berkeley.

Cohen, Steve (2003) "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," 2003 Sawtooth Software Conference Proceedings, 61-74.

David, H. A. (1988) The Method of Paired Comparisons. New York: Oxford University Press.

Finn, Adam and Jordan J. Louviere (1992) "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Public Policy and Marketing, 11: 12-25.

Johnson, Richard M. (2000) "Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes," Technical Paper available at www.sawtoothsoftware.com.

Sá Lucas, Luis (2004) "Scale Development with MaxDiffs : A Case Study," 2004 Sawtooth Software Conference Proceedings, in press.

Sawtooth Software (2003) "Best/Worst Designer," white paper available at www.sawtoothsoftware.com.

Swait, Joffre, Jordan J. Louviere and Don Anderson (1995) "Best Worst Conjoint: A New Preference Elicitation Method to Simultaneously Identify Overall Attribute Importance and Attribute Level Partworths," working paper.

Thurstone, L. L. (1927) "A Law of Comparative Judgment," Psychological Review, 4: 273-286.

CONJOINT ANALYSIS

CONJOINT ANALYSIS: HOW WE GOT HERE AND WHERE WE ARE — AN UPDATE

JOEL HUBER
DUKE UNIVERSITY

This paper was originally published in the 1987 Sawtooth Software Proceedings. Because of the excellent quality of this work, we asked Joel to present this paper again at the 2004 Sawtooth Software Conference. We have published the original paper below, but with added footnotes by the author and myself, given a 2004 perspective. –Bryan Orme, Editor.

Conjoint analysis has had a profound effect on the conduct of research in many facets of business, particularly in the areas of product positioning and new product development. It is a field approaching the maturity stage of its life cycle. However, with the coming of inexpensive, user-friendly programs for conjoint, we can expect its use to increase substantially. Indeed, we will soon see the day when virtually all market research firms will offer conjoint studies as part of their standard repertoire. Managers will use conjoint not just for special projects, but as an indispensable tool enabling them to test the impact of proposed actions on the market. Conjoint is becoming less elite, its secrets no longer the property of a few, but available in its simpler versions to all.[1]

Today I would like to present my personal perspective on the history of conjoint analysis. The field is shaped by two fundamentally conflicting forces. First, there are the idealistic psychometric forces that started the field. Opposing these, while at the same time arising from them, are the pragmatic forces, practitioners who have determined the way conjoint is used. The tension between these forces has shaped the growth of the field and will continue to guide its future development.

THE PSYCHOMETRIC TRADITION

The term "conjoint[2]" has itself contributed to the mystery of the field. The term arose out of an attempt to apply extensive measurement to preference judgments. Extensive measurement refers to a method to build a scale by comparing relative lengths (extensions) of objects. For example, by comparing the lengths of different rods put end to end, one can form a scale on which it is appropriate to perform such operations as addition and subtraction. While such interval-level scales are relatively easy to generate from physical quanta such as weight, size and time, they have been notoriously difficult in the case of human preferences.

The difficulty arises because we know what it means to say that we like potatoes better than rutabagas, but generally not what it means to say that our liking for potatoes over rutabagas is greater than our liking for artichokes over eggplant. This indeterminacy poses a problem to

[1] Indeed, the use of conjoint analysis has dramatically increased. Based on a 2004 Sawtooth Software customer survey, we project that between 5,000 to 8,000 conjoint analysis projects were conducted by Sawtooth Software users during the previous 12-month period. The relative proportion of projects by conjoint flavor among our users was CBC (61%), ACA (27%) and CVA (12%).

[2] Many researchers today believe, incorrectly, that the term conjoint comes from the idea that respondents are asked to "CONsider features JOINTly." Conjoint means "to join or become joined together," as in "conjoined twins."

psychometricians who want the same solid base for measuring the psyche as physicists had for measuring weight. Without interval scales of preferences, it is difficult to specify what it means to have an additive model of preference.

The psychometricians reasoned that while our ordinary language pronouncements of preferences do not directly produce interval scales, certain kinds of preference judgments had to be based on utility values that do. One set of preference judgments that requires metric underpinnings refers to compound or conjoint objects.

Consider the statement that one prefers a $10,000 convertible to an $8,000 sedan. This statement implies that the benefit of a convertible over a sedan is greater than $2,000. Psychometricians were able to show that by putting together a number of such preference statements, it is possible to derive intervally scaled additive partworth utilities that could underlie these preferences. Further, they specified a number of tests to determine if such an interval scale is justified, given the preference orderings.

Conjoint measurement provided a theory for creating a measurement scale from judgments on compound or conjoint objects. It generated a great deal of excitement when first proposed. Conjoint was a "psychometric conjurer's stone"—a way to transform the dross of ordinal preferences into the gold of interval scales. At last the measurement of preferences might be put on par with measurement in the exact sciences.

The early contributions focused on finding sets of elegant axioms and/or conditions required to uncover the latent interval partworths. Some of these conditions, such as independence, are well known, while others, such as double cancellation, are less well known. The axiomatizations are best summarized in the classic Foundations of Measurement, Volume 1 (1971) by Krantz, Luce, Suppes, and Tversky. In the preface to that volume, reference is made to Volume 2 on applications. It is ironic and significant that that volume has not yet been published[3].

What happened? As soon as the psychometricians applied their models to human behavior, they found that the axioms were consistently violated. It was very similar to what is now occurring with respect to the Von Neumann-Morgenstern axiomatization of choice under uncertainty (Thaler, 1985). Virtually all the axioms were violated in relatively minor but systematic ways. Initially, it appeared that random error could account for the intransitivities and lack of additivity found. However, as more elegant and precise tests were devised, this escape was also blocked (e.g., see Falmange, 1976).

In hindsight, it is not surprising that if people cannot give consistent interval partworth values directly, then such metric rigor is unlikely to be hidden beneath more complex judgments on conjunctive stimuli. There is no intervally scaled ruler hidden in the brain that can account for complex preference judgments.

Still, the psychometricians provided a clear and coherent tradition, aspects of which are still important today. That tradition includes the following components:

First, the belief that individual preferences can be expressed in numerical terms that lead to behavior.

[3] Since then, a follow-up volume did appear: Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations (Foundations of Measurement) by Patrick Suppes, David M. Krantz, R. Duncan Luce, Amos Tversky, Academic Press 1989.

Second, the focus on comparisons among conjunctive stimuli, defined on multiple attributes, so that the response requires trading off high levels on one attribute with low levels of others.

Third, the tradition of using factorial designs in which the attributes to be tested are statistically independent of one another.

Fourth, the emphasis on testing the assumptions, such as additivity, as a prior condition to estimating the partworth utilities.

Finally, the orientation to ordinal responses from subjects as the primitive behavior being modeled, rather than direct magnitude or interval scales[4].

From Psychometric Swords to Market Researcher's Plowshares

The psychometric tradition is rigorous and idealistic, whereas its adoption by the market research community has been approximate and pragmatic. The market research community began with the same rigorous models, but soon found that the partworth utilities were managerially very useful despite the fact that the tests did not work. In effect, the operation failed, but the patient thrived. Useful aspects of the original conjoint measurement framework were adapted and less useful ones were dropped.

Rich Johnson's succession of conjoint models is perhaps most illustrative of the changes that occurred. Rich was trained as a psychometrician, and his original trade-off analysis used 3-by-3 trade-off matrices, in which respondents were to rank order alternatives defined on various levels of the two attributes (Johnson, 1974). Then, by computerizing the approach, he was able to avoid certain redundant questions and speed up the task. However, the price of this additional speed was a lessened ability to make consistency tests at the individual level. His next step expanded the task from one of categorical preferences to graded-pair comparisons. This permitted more information to be collected from respondents with very little additional cost in time or effort. Finally, he used the personal computer to merge direct attribute judgments with paired comparisons and guide the selection of "optimal" pairs during the conjoint task[5]. All these changes helped to obtain information from respondents more efficiently and to formulate a better predictive model of their preferences. Still, these changes represent a substantial departure from the psychometric tradition.

Much of the ambivalence between idealism and practice in the marketing research community is found in Green and Srinivasan's (1978) classic review article on conjoint. In that article they differentiate conjoint analysis from the older conjoint measurement in order to make appropriate separation between the two fields. A dualism is evident in their discussion of the various ways to perform conjoint analysis, sometimes focusing on what is theoretically justified, while at other times succumbing to practical reality.

[4] It is interesting to note that we have come full circle. The early researchers in conjoint measurement favored ordinal (ranked) data, rather than ratings judgments. In the late 70s and throughout the 1980s, researchers tended to favor the increased information (and ease of data collection and analysis) provided by ratings scales. Now, in 2004, interest has shifted back to non-metric scaling, this time embodied in choice (as in CBC).

[5] Joel is referring to Rich Johnson's Adaptive Conjoint Analysis (ACA) product, which went on to become the most widely-used conjoint method and software system for conjoint analysis is the late 1980s and throughout the 1990s.

What did the marketing research community take from the psychometricians and what did they change? Generally, the trends are evident in Cattin and Wittink's (1987) review of practices in conjoint.

First, the field continues to consider behavior as captured by partworth utilities and simple additive models[6].

Second, in keeping with the psychometric tradition, they use compound stimuli which force individuals to trade off conflicting attribute levels.

Third, they still rely on orthogonal arrays[7], although highly fractionated designs have replaced the original full factorials.

The first three components of the psychometric tradition have been passed down relatively unchanged. The last two, the structural tests and the nonmetric orientation, have rapidly eroded[8].

Consider, first, the tests of the structural composition rules (such as additivity) that were the major focus of the axiomatic systems. These tests are now virtually ignored, or worse, assumed away. Consider, for example, the common use of fractional main-effects design. While these offer far greater efficiency and permit main-effects estimates of many more attributes, they assume that interactions are zero. If there are interactions, the preference function will be biased or wrong. Further, because no test is possible, the analyst will never know that the results are biased[9].

The other major deviation from the psychometric tradition has been a move from a nonmetric to a metric orientation. This has occurred both in the kinds of data collected from respondents and routines used to analyze it. The original reason to use rank-order inputs over quasi-metric ones stemmed from a legitimate uncertainty about what respondents meant in responding to, for instance, a ten-point strength-of-preference scale. A number of nonmetric analysis packages, such as Kruskal's (1965) MONANOVA and Srinivasan's and Shocker's (1973) LINMAP permitted relatively easy analysis of input data about which only ordinal properties could be assumed. The shift from ordinal to metric inputs has been largely pragmatic. For example, putting 25 profiles onto a ten-category sort board is both easier for subjects and provides more reliable inputs than an exhaustive rank-order task. Using a rating scale permits one to generate predicted choices with equivalent reliability but fewer judgments.

Metric methods have also become more popular as methods of analysis. This shift is in part due to the ease of use of ordinary least squares over nonmetric routines. But it also stems from a realization of the value of the weak but errorful metric information in a rating or a sort-board task. Of course, such data can be analyzed by nonmetric procedures, such as monotone

[6] Interestingly enough, with respect to model specification, we have again come full circle. Early conjoint researchers favored parsimonious models that featured part-worth estimation at the individual level. Through the 1980s and 1990s, there was increased interest in linear estimation of terms, higher-order effects, and (largely due to necessity) pooled estimation from choice models. With the advent of HB analysis, evidence is in favor of simple main-effects models, with part-worth representations of attributes' utilities.

[7] Largely due to efforts by Warren Kuhfeld (and co-authors), researchers today have recognized that strict orthogonal arrays are often not optimal (especially in cases involving asymmetric designs). Computerized searches that sacrifice some orthogonality in favor of increased level balance are able to achieve overall greater precision of estimates.

[8] Again, see the discussion in footnote 3 regarding the recent strong renewal of interest in non-metric data (choice).

[9] Lately, conjoint analysts have come to appreciate that many interaction effects are actually due to heterogeneity, and individual-level part-worths used in choice simulators can reflect a variety of interaction effects (through sensitivity analysis), even though the effects were not directly specified in the model.

regression or LINMAP. The nonmetric procedures find a monotone transformation of the dependent variable that best fits the model. With such routines this transformation is very linear, indicating that the transformation provides little additional information. More significantly, the monotone transformation generally degrades the predictive ability of the model (e.g., see Huber, 1975). Quasi-metric data has some interval properties that nonmetric routines treat as noise, but metric routines are able to use. Nonmetric routines may be losing popularity simply because they do not appear to help predictions. However, this is simply a pragmatic criterion; we lack a good theoretical reason why metric routines should work better. Indeed, most theoretical considerations lead to the championing of a nonmetric orientation[10].

To summarize, the marketing research industry has taken some of the external trappings of conjoint measurement, but has generally deleted or modified its elegant inner workings—those dealing with hypothesis testing on strictly ordinal inputs. While these modifications may not, in a strictly predictive sense, matter, it is appropriate to inquire as to the kind of offspring that has emerged from this union of pragmatic and idealistic parents.

WHAT REALLY GOES ON DURING A CONJOINT EXERCISE?

If we are not tapping a latent interval utility scale during conjoint, what are we doing? The answer depends on whether the stimuli are unitary or decomposable. Unitary stimuli are those which respondents cannot easily break into component parts, such as foods, scents, or esthetic objects. With these kinds of stimuli, response to factorially designed stimuli is holistic and generally frustrates attempts to build simple models of that response. Main-effects designs are at a particular disadvantage in that the assumption of additivity is usually violated. Specialized designs are appropriate if the source of the interactions can be localized within a few variables, and if there is not too much heterogeneity across subjects, conditions that are sadly not often satisfied.

Because of the design problems with unitary stimuli, most conjoint analysis uses readily decomposable attributes and displays attributes in ways that make it easy for respondents to separate them. Respondents are given repeated questions with predictably arrayed attributes, and anyone who has watched a conjoint exercise knows the result. Respondents simplify the task by focusing on a few attributes. Each profile is evaluated by scanning these attributes and adjusting the valuation of the alternative accordingly. This process results in a very good fit of the additive model at the individual level. Typically, a small number of attributes are strongly significant and the rest are nonsignificant. Interactions are very rare—they require extra processing. Different tasks produce slightly different patterns of responses. For example, paired comparisons may produce more significant attributes. However, the general pattern of a strongly simplified evaluation strategy that has accurately captured an additive model emerges regardless of the particular task.

Evidence for such a simplification process appears in an anomalous result which I have found in my work, and I suspect many of you have as well. Following a conjoint exercise we often include a choice task, either using actual brands or profiles where the attributes have been scrambled to reduce the likelihood that the conjoint choice process will be trivially repeated. In these studies it is possible to note the correspondence between the internal fit of the conjoint and

[10] Again, the orientation toward metric conjoint has waned, and researchers are using choice-based conjoint flavors more often now than metric responses. Even if using metric responses, attention has turned from OLS to HB analysis.

its accuracy in predicting the holdout choices. If the conjoint task is a correct representation of the holdout choices, then the better the fit of the conjoint and the more it satisfies the axioms, the better it should predict the holdouts. In other words, respondents who are well modeled by the conjoint should be ones we can best predict. However, my experience has not been consistent with this expectation. Correlations between conjoint fit and predictive accuracy are very low and often negative, particularly if one screens out the totally random subjects. How could this be?

The simplest account focuses on differences in the conscientiousness of respondents. The conscientious ones try to consider as many attributes as possible in their conjoint task. Being mortal, they make mistakes, resulting in preference reversals and greater error levels. When they make the holdout choices, they conscientiously try to be consistent with their earlier judgments, resulting in greater predictive accuracy. For the less conscientious respondents the reverse holds. They simplify the conjoint task greatly in order to get through it, focusing on one or two attributes. This simplification permits remarkable fits to be the additive conjoint model. In the holdout choice task with distorted or scrambled attributes, using the same simplified decision rule is not easy. These less conscientious respondents shift strategies, basing their choices on different attributes than in the conjoint, resulting in a poor correspondence between the two.

The point is not to resolve this anomaly, but to raise an issue about conjoint. The professional success of conjoint practitioners attests to how well it works. It produces intuitively pleasing results that managers find very useful, although we do not have a clear account of why it works. Indeed, its problems could lead many to discard it. After all, it has shed its theoretical roots and appears only to capture a simplified and truncated version of choice behavior[11]. The next section considers why conjoint works, and this leads naturally into ways in which it might be improved.

WHY CONJOINT WORKS

1) Conjoint requires tradeoffs that are similar to those in the market.

A conjoint task is valuable because it forces the respondent to evaluate conflicting attributes, as between the type of car and its price. People typically try to avoid making such judgments by searching for unambiguous solutions involving dominance or following a relatively clear heuristic. However, the marketplace also requires such judgments and people make them when they must in the marketplace or the conjoint task.

The conjoint task, in which alternatives are compared on a number of dimensions, can be usefully contrasted with a direct elicitation approach, in which attribute utilities are directly assessed. There are two problems with direct elicitation that are ameliorated with conjoint. First, it is difficult with the direct elicitation approach to keep a respondent from seeing everything as important. Certainly, $2,000 is very important in selecting a car, but it may not be more important than the difference between a convertible and a sedan. Second, direct elicitation does not directly relate to a choice in the marketplace[12], but is a summary measure of those behavioral decisions. In contrast, the conjoint task is more directly analogous to market choice.

[11] With the resurgence of interest and application of the non-metric tradition through choice, the conjoint analysis applied in 2004 is truer to its theoretical roots. Rather than use ratings to predict choices, we employ choices to predict choices. Still, true to Joel's assertions, our modern CBC questionnaires still reflect a simplified (truncated) representation of real-world choice behavior.

[12] Joel discounts the predictive validity of direct elicitation (self-explicated) methods. The debate continues nearly 20 years later, though the vast majority of the industry still favors tradeoff methods. Some well-known researchers have published papers

2) The simplification in conjoint may mirror that in the market.

The simplification found in conjoint to a small number of attributes is only misleading if there is a very different kind of simplification in the marketplace. There is evidence that the decisions in the market are based on remarkably few dimensions (Olshavsky and Granbois, 1979). If so, then conjoint may indicate those few attributes on which the consumer bases his or her decisions.

Further, to the extent that conjoint is used to predict aggregate shares, it does not matter that an individual's selection of attributes is unstable over time. As long as conjoint captures an unbiased selection of attributes at the time, the aggregate market shares will also be unbiased. The criteria for conjoint to work at the aggregate level are considerably less stringent than for individuals.

To summarize, conjoint works by forcing respondents to make trade-offs among attributes. They then simplify the task by selecting a small number of attributes on which to base their judgments. To the extent that this pattern of simplification is mirrored in the marketplace, then conjoint market shares will predict quite well.

3) Conjoint profiles are orthogonal.

The use of orthogonal arrays is an aspect of the original psychometric formulation that has resisted modification by the marketing research community[13]. In particular, main-effects fractional factorials have been heavily used because they permit more attributes and levels. In the case of decomposable stimuli, the simplification by respondents typically assures that interactions will not be present.

The orthogonal nature of these designs is important in a way not generally appreciated. An orthogonal design is simply one in which the levels of different attributes across profiles are uncorrelated. Such designs assure that an estimate of one attribute is unaffected by the estimate of other attributes. It might appear that we could suffer moderate levels of multicollinearity without much harm. Most econometric models seem to thrive with much higher levels of multicollinearity. However, the fact that respondents regularly simplify the conjoint task leads to substantial difficulties if any attributes in the design are correlated. Let me illustrate with an example.

I was involved in a conjoint study dealing with snowmobiles. We were concerned with the impact of engine size on the acceptability of the snowmobiles. So that the profiles would be realistic, we increased the price by an appropriate amount for each of the engine sizes. Price was positively correlated with engine size in the design. This was not expected to be a problem, except that multicollinearity might render the estimates marginally less efficient. However, we found that for many respondents the coefficient of price had the wrong sign (high price is preferred) and for others the coefficient of engine size had the wrong sign (small size is preferred). We believe that respondents, in simplifying, had tended to focus on one of these two variables. For example, those who focused on engine size gave higher evaluations of profiles

defending the predictive validity of self-explicated methods, placing it on par or in some cases above that of conjoint methods. Improvements in the manner of collecting self-explicated judgments have led to its increased performance. For example, self-explicated judgments now require respondents to trade-off one attribute for another by techniques such as a constant-sum scale.

[13] See footnote 7 regarding the enlightened frame of reference with regard to orthogonal arrays.

with larger sizes, but these, by our design, had higher prices. Thus, high price appeared to be desired by these subjects.

This account points to an advantage inherent in orthogonal arrays. For orthogonal arrays, the main-effect estimate for each attribute is independent of the others, whereas in the correlated case, this independence does not hold. If attributes are correlated, misspecification results in biased estimates. The particular misspecification that so often occurs in conjoint is simplification, where a number of attributes are effectively ignored. With orthogonal arrays, the estimated coefficients for the attributes remain unbiased. In the correlated case, misspecification results in distortions in the coefficients for both the attributes focused upon and those ignored. Thus, orthogonal arrays play an important role of increasing the robustness of conjoint by making it less likely that coefficients have counter-intuitive signs. This robustness contributes to much of the managerial satisfaction with conjoint.

4) Conjoint Simulators Account for Heterogeneous Tastes in a Market.

A final reason conjoint works relates to the way it is used. Typically, partworth functions are estimated at an individual level, then these are aggregated to produce estimates of market share under various conditions or scenarios. These simulations implicitly reject the notion that one homogeneous customer can account for choices in the marketplace. Instead, one is forced to deal with each customer having an idiosyncratic preference function, or at least with an explicit clustering in which strongly differing tastes are represented in different clusters. This practice of preserving the heterogeneity of individuals in simulators facilitates the representation of two important properties of markets that are difficult to achieve with other market research techniques. These are the properties of differential substitution and dominance.

Differential substitution refers to the notion that a new competitor in a market tends to take share differentially from those brands with whom it is most similar. For example, New Coke took most share from Classic Coke and Pepsi, and had relatively little impact on the lemon-lime soft drink category. Dominance refers to the idea that a brand that is equal on most attributes but slightly worse than its competitor on others gets very low share.

Managers understand these two phenomena and expect simulations to reflect them in positioning and new-product studies. Unfortunately, most models of market structure can account for differential substitution and dominance only in a very awkward fashion. By contrast, both phenomena arise naturally out of a conjoint simulator. For example, differential substitution occurs because individuals who like Pepsi tend to like Coke. Generally, changes in any brand will have a greater impact on similar brands than dissimilar ones. Dominance is represented in a conjoint simulator since a brand that is dominated consistently loses out to that competitor and achieves almost no share.[14]

To summarize, conjoint works because it is derived from a task that forces respondents to trade off attributes in ways that may parallel actual buying behavior. The orthogonal design provides not only efficiency, but a strong degree of robustness against misspecification. Finally,

[14] Joel’s discussion of the success of simulators based on individual-level data well captures why HB analysis (which yields individual-level estimates) has been such a boon for choice researchers. Later developments such as Randomized First Choice have leveraged these principles of differential substitution and dominance, but achieving individual-level part-worth models has provided probably the most incremental benefit.

the preservation of the utility function at the level of the individual or segment permits us to simulate a market that behaves in ways we expect.

The Future of Conjoint

There are three areas in which substantial changes are anticipated. The first area of change involves conjoint theory, the way we think about and organize the field. The second involves the task. The final area involves the ways conjoint is used.

1. Conjoint Theory: From Estimation of Utilities to Emulation of Behavior

In terms of theory, the psychometric framework has one fatal flaw—it assumes that utilities exist that account for preferences. Reality, unfortunately, is far more complex. Preferences between profiles are better described as being constructed, using various heuristics from the information at hand. The additive models in conjoint may reflect this process and at times correspond to it quite well, but conjoint certainly cannot reveal an interval scale in the brain. Instead of thinking that conjoint estimates latent utilities, it is more appropriate to consider that it emulates choice behavior.

There is a significant loss associated with giving up the idea that conjoint reveals a latent preference scale. The existence of a scale means that it is possible to formulate optimal experimental designs that result in the most efficient estimates of that preference scale. If, instead, conjoint is viewed as a paramorphic emulation of behavior, then it is no longer clear what makes a good design.[15]

There are some advantages with viewing conjoint as an emulation of choice behavior. First, we are no longer permitted to beg the question about the applicability of conjoint to the marketplace—something we can do if the same utility scale is presumed to underlie both choice and conjoint. Instead, we are forced to ask whether the conjoint task corresponds to the choice in the marketplace. For example, it is relevant to assess the number of dimensions that are actually used in the market, then choose a conjoint task that results in similar depth of processing. Second, in the behavioral perspective, one is freed from a rigid adherence to a particular question form to capture appropriately the choice process.[16]

In contrast to utility, behavior can be captured in a number of paramorphic ways. Indeed, to get at some behavior it may be better to use different kinds of questions, such as combining direct elicitation and paired comparisons, rather than focusing on a particular question type. Switching to a different kind of question may discourage respondents from getting into a response pattern. Further, the differences in responses across question types will reveal the stability of the choice behavior.

2. Conjoint Task: From Monolithic and Rigid to Multifaceted and Adaptive

As our way of thinking about conjoint changes, so will the task we ask of subjects. Two important changes will occur, both adding new kinds of questions to the traditional conjoint task.

[15] The search for ways to incorporate prior beliefs about latent utilities into the design of choice experiments has proven elusive over the years (See Johnson's paper on ACBC within this same volume). Greater utility balance usually leads to greater fatigue and respondent error. And, if prior beliefs aren't re-introduced when estimating final part-worth utilities, utility-balanced designs may also decrease precision.

[16] These points clearly argue in favor of choice-based experiments, with multiple product concepts available per choice task. Had there been a method for successfully estimating idiosyncratic models of preference from choice data in the late 1980s, Joel no doubt would have embraced choice-based conjoint methods at that time.

The first involves the ability of routines to adapt to the idiosyncratic behavior of respondents, while the second adds a relatively realistic choice task at the end of the conjoint task to better assess the correspondence between conjoint and market choice.

We have already begun to see ways in which conjoint can profitably adapt to the needs of individual respondents. The Sawtooth Software ACA System uses "priors" to construct pairs so that paired comparisons are as closely balanced as possible. While this reduces the strict statistical efficiency of the design, it makes the questions more challenging and increases the correspondence between conjoint and subsequent choice (Huber and Hansen, 1986). The flexibility of the personal computer in administering conjoint will certainly lead to other adaptive mechanisms. Two such applications are particularly exciting—hierarchical conjoint and interaction testing.

Hierarchical conjoint permits the modeling of rich decision making despite simplification of the conjoint task by respondents. If respondents can only cope with two or three attributes at a time, the routine determines the partworth functions for these most important attributes, then fixes their levels. Subsequent test profiles only differ on the remaining attributes. For example, the values of location and price might be estimated first in an apartment study, followed by furniture style and room layout. Under standard conjoint, these less important attributes might not be revealed, whereas in hierarchical conjoint both their position in the hierarchy and relative importance could be assessed.

A second adaptive mechanism concerns the search for interactions. A promising technique might work as follows: First, a main-effects design would make rough estimates assuming no interactions. The residuals from these initial judgments would be tested for weak (and confounded) evidence of interactions. Then these potential interactions would be tested through specially designed questions. Such a method would avoid the current problem of having to assume that interactions are zero, and could be very helpful in studies where different interaction patterns are expected across respondents.

Perhaps the greatest prospect for improving conjoint involves the inclusion of a relatively realistic choice task at the end of the exercise. These choice tasks sometimes take the form of asking respondents to choose brands from a simulated store or having them evaluate actual products. A much less costly, if somewhat less realistic option, is to add choice questions at the end of the conjoint exercise. These might take the form of choices among alternatives defined on different attributes than those displayed in the original conjoint. This is easy to do within the Sawtooth Software ACA System by adding Ci2 System questions after the ACA System section. Further, new developments in personal computers will permit potential choice objects to be displayed in color video, thereby increasing the realism of the holdout choices even more.

The value of having a holdout choice task is twofold. First, in the field this holdout task is useful in identifying respondents whose conjoint responses are unlikely to correspond to their behavior. These respondents can then be given less weight in the simulation. Second, it permits an immediate assessment of the relevance of conjoint to choice. Where a conjoint model appears not to correspond to choice, it can be changed or improved[17]. This permits the testing of different forms of conjoint and leads to continuous improvement in its predictive validity.

[17] Here Joel argues in favor of adjusting metric-based conjoint methods to better predict limited choice holdouts. Again, lacking a robust methodology to estimate individual-level models from experimentally-designed choice tasks alone, it naturally wasn't

3. Conjoint Simulators: From Complex and Opaque to User Friendly and Understandable

The third area in which we can expect conjoint to progress involves the way data are used in choice simulators. Elaborate choice simulators currently exist, permitting the analyst to ask virtually any question of the data. The positioning of a product can be optimized with respect to sales or profitability. Alternatively, one can assess the impact of changes on positioning or competition on the behavior of various segments. Unfortunately, these simulators are not particularly user-friendly. With time they will become easier to use and their use by managers should increase[18].

Even if made more user-friendly, there is still a problem. While simulators permit the manager to cope with heterogeneous tastes in the market, they remain a black box. The only way for a manager to understand a simulated market is by experimenting with a large number of runs. These simulation runs give managers a feel for the behavior of the market in the face of different positionings or competitive offerings. Developing this understanding is hard, relatively unstructured work. We need to develop ways to permit managers to understand more directly the behavior of the market being simulated[19]. Preference spaces may provide part of the answer, but it is very difficult to represent both respondent heterogeneity and central tendency on one map. Working with a small number of segments also helps. We would like to know how to specify a small number of segments, such that their aggregate behavior closely approximates the entire market. Defining such segments remains an unsolved question that evades simple solutions.

Just as the computer continues to make the conjoint task itself more appealing to respondents, it will also increase the ease by which the output of conjoint can be applied by managers. Once again, we may have been saved by the computer. We have come to realize that choice behavior cannot be captured by a simple scale of utilities. As we come to accept a behavioral base to conjoint, we can never return to the elegance and simple unity that characterizes the psychometric framework. However, with the computer we have a tool that may be powerful enough to mirror the complexity of behavior and display it in a manageable and understandable way.

apparent at the time that we might forego the ratings-based exercise altogether and just focus on the more realistic choice scenarios as calibration tasks!

[18] Certainly, simulators have become easier to use and ubiquitous. Spreadsheets were available (Lotus 123) in the mid-1980s for developing conjoint simulators, and Microsoft's Excel was being released in the year Joel presented this paper at the Sawtooth Software Conference. Still, in 1987, it was hardly widespread practice to build conjoint simulators within spreadsheets. Today, researchers commonly build attractive spreadsheet applications for market simulations in Excel. Sawtooth Software's ASM (Advanced Simulation Module) has made product optimization within its commercial market simulator (SMRT platform) very straightforward.

[19] These problems by in large have not been solved. However, the application of cluster analysis or particularly Latent Class to the development of "needs-based" segments and the use of segment membership as banner variables for "cross-tab" style display in simulation output may increase the level of understanding of market structure and preference. Still, it takes a talented manager to absorb so much detail and come to sound conclusions.

REFERENCES

Cattin, Philippe and Dick R. Wittink (1987) "Commercial Uses of Conjoint Analysis: An Update" Working Paper, Graduate School of Business, Cornell University, Ithaca NY.

Falmange, Jean-Claude (1976) "Random Conjoint Measurement and Loudness Summation" Psychological Review, 83, 65-97.

Green, Paul E. and V. Srinivasan (1978), "Conjoint Analysis in Consumer Behavior: Issues and Outlook" Journal of Consumer Research, 5, 103-23.

Huber, Joel (1975) "Predicting Preferences on Experimental Bundles of Attributes: A Comparison of Models" Journal of Marketing Research, 12, 290-7.

Huber, Joel and David Hansen (1987) "Testing the Impact of Dimensional Complexity and Affective Differences in Adaptive Conjoint Analysis" Advances in Consumer Research, 13, in press.

Johnson, Richard M. (1974) "Trade-Off Analysis of Consumer Values" Journal of Marketing Research, 11, 121-217.

Krantz, D.H., R. D. Luce, P. Suppes and A. Tversky (1971) Foundations of Measurements, Vol 1, New York: Academic Press.

Krusksal, Joseph B. (1965), "Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data" Journal of the Royal Statistical Society, Series B, 27, 251-63.

Olshavsky, Richard W. and Donald H. Granbois (1979), "Consumer Decision Making—Fact or Fiction" Journal of Consumer Research, 6, 93-100.

Srinivasan, V. and Allan D. Shocker (1973) "Linear Programming Techniques for Multi-dimensional Analysis of Preferences" Psychometrica, 38, 337-69.

Thaler, Richard (1985), "Mental Accounting and Human Choice" Marketing Science, 4, 199-214.

.

COMMENT ON HUBER: PRACTICAL SUGGESTIONS FOR CBC STUDIES

JON PINNELL
MARKETVISION RESEARCH

Joel Huber has provided a unique and very insightful recounting of the evolution of conjoint and discrete choice. I'd like to think that Joel provided the answers to the "what," "where," "why," and "when" questions of conjoint and discrete choice.

What I would like to offer is the "how" of conjoint or discrete choice, or a guide to best practices in commercial research. Inevitably, I am not able to do the "how" without including the "who." So, for the "who" — these are my perspectives based on the experimentation that we have conducted at MarketVision; and the perspectives that others have shared with me or have shared at conferences, especially at the conferences that Sawtooth Software has sponsored over the years. Our approaches may not be the most commonly used, and may be debated heartily, but we have adopted these as well founded as well as field tested and proved.

Most of the suggestions that follow are based on the guiding principle of increasing efficiency. The first two suggestions appear to run counter to that objective. First, let me provide some context on how we think about efficiency. There is clearly a statistical efficiency, often evaluated via D-efficiency, which expresses the statistical "goodness" of the design. However, there are other efficiencies that should be considered as well, such as the time it takes the respondent to provide a response, the amount of information in that response, and the design and analysis time required by the researcher.

Below, I offer ten practical suggestions based on our beliefs about best practices in conjoint research. They are…

PRACTICAL SUGGESTION #1: USE CHOICE-BASED TASKS.

The most recent *Sawtooth Solutions* (Summer 2004) published results showing what percent of studies were conducted using CBC (61%), ACA (27%), and CVA (12%). This shows that among this audience choice-based designs have become the most prominent approach to conjoint analysis.

A decade ago there was a great deal of concern about the limited amount of information contained in a single choice—we knew which alternative a buyer preferred—but not by how much. We also knew that various methods (ratings versus choice) provided different answers (Huber, et al, 1992; Pinnell, 1994; Huber and Pinnell, 1995; Huber, 1997).

At the time, the tools commonly available to researchers prohibited individual-level utility estimation with choices. Instead, multinomial logit models were used to estimate utilities in aggregate, or maybe at the subgroup level. Several researchers who wanted individual level utilities suggested dual conjoint (Huisman, 1992) —with some converting ACA's individual level utilities to match those consistent with those from choice exercises (Huber, et al, 1992; Pinnell, 1994; Williams and Kilroy, 2000).

Also troubling was the difference between choices and ratings in the respondent task. Choices were a common customer and respondent task, while ratings were less intuitive and less natural.

Over the past decade, we have seen choice match in-market behavior more realistically than ratings. We have also seen the introduction of methods to allow estimation of individual level utilities from choice studies.

Choices are still less efficient than ratings, but are processed easily and quickly by respondents and appear to provide the best predictor of in-market behavior—most likely because it is the same task consumers must complete in the purchase process.

PRACTICAL SUGGESTION #2: USE RANDOMIZED DESIGNS.

Many early applications of discrete choice modeling relied on a fixed experimental design. However, with the growing popularity of computer aided interviewing over the past decade, randomized designs became feasible for a larger number of practitioners.

Mulhern (1999) found that randomized designs are nearly as efficient as fixed designs for symmetric choice experiments; for asymmetric choice experiments, randomized designs appear to be more efficient than fixed designs. In the specific example cited by Mulhern, the randomized design is 95% as efficient as the optimal fixed design for a symmetric choice experiment including ten attributes with four levels each. For an asymmetric choice experiment with eight attributes, one with seven levels, four with six levels, and three with five levels, Mulhern found the randomized design to be approximately 14% more efficient than the fixed design.

Chrzan and Orme (2000) compared the statistical efficiency of various fixed and random design strategies under different conditions. They found that one or more of the random design strategies was optimal or nearly optimal for all conditions except for a situation with "many" interactions.

The Sawtooth Software CBC Users Manual reports relative efficiencies greater than 90% for the randomized design. The median efficiency for the randomized designs reported is about 97% relative to a hypothetical orthogonal design.

Randomized designs can also reduce order and context effects relative to fixed designs.

Randomized designs are less efficient than fixed orthogonal designs, but they give the researcher more flexibility and are easier to implement than fixed orthogonal designs. The loss in efficiency of randomized designs relative to fixed orthogonal designs is minimal in most choice experiments, which leads us to recommend randomized designs over fixed designs due to the gain in flexibility and ease of implementation.

There is one caveat to the recommendation of using randomized designs. It is often beneficial to include at least one fixed task as a holdout task to allow an assessment of respondent reliability.

Practical Suggestion #3: Use a large(r) number of alternatives per task.

Designing choice tasks also leads to the debate over respondent fatigue and burden versus the researcher's desire to gather as much information as possible from the respondents. This leads to the questions of how many choice tasks each respondent can reliably evaluate and how many alternatives can be included in each task without overburdening the respondent.

The number of choice tasks to ask was addressed by Johnson and Orme (1996). They conclude "you can usually ask at least 20 choice tasks without degradation in data quality."

We think the number of alternatives per task is a more interesting discussion. Increasing the number of alternatives per task provides incremental information, though it may not be immediately clear how. While choices are relatively inefficient, multinomial logit is quite powerful, and it derives that power based on the number of pairwise inequalities it evaluates in a choice task. In the case of a choice task with six alternatives, five pairwise inequalities are created compared to just one pairwise inequality from the two-alternative task, or pairs. This implies statistical efficiency is greater for choice studies that include a greater number of alternatives per task. Louviere and Woodworth (1983) and Bunch, Louviere, and Anderson (1994) each conclude that paired comparison choice tasks are less efficient from a design perspective than designs that include more alternatives per task.

We evaluated (Pinnell and Englert, 1997) increasing the number of alternatives in choice tasks on cost, congruence, and efficiency criteria to determine if the added complexity of more alternatives is outweighed by higher statistical efficiency. The tasks with four alternatives took about 33% more time and those including seven alternatives took about 60% more time for respondents to evaluate than pairs. The cost in time of evaluating choice tasks with more alternatives is relatively small compared to the gain in information collected, as summarized in the following table (based on time equalized empirical findings):

Efficiency:	Sevens nearly 6X D-efficiency of pairs
Parameters:	Sevens 59% larger than pairs
Std. Error of Parameters:	Sevens 25% smaller than pairs
Validity:	Sevens higher hit rates and lower MAE
Cost (Time):	12 sevens = 20 pairs

A secondary benefit of using a larger number of alternatives per task with randomly generated designs is that the occurrence of dominating concepts will be greatly reduced. Eliminating dominated concepts, thereby increasing the utility balance of the alternatives in the task, will increase the information contained in each respondent choice. Utility balance has been shown to increase the effectiveness of each respondent choice (Huber, Zwerina, and Pinnell, 1995; Huber and Zwerina, 1996; Johnson, Huber, and Bacon, 2003; and Johnson, Huber, and Orme, 2004).

PRACTICAL SUGGESTION #4: ONLY USE FIRST CHOICES TO ESTIMATE UTILITIES.

We have seen in previous work (Pinnell 1999a, 1999b) that respondents take longer to provide a first choice when they know that they are being asked to provide a full rank order of the alternatives compared to when they are providing only first choices. Subsequent choices, such as ranked second or ranked third did not appear useful, as those utilities were smaller (lower scale and larger error) and, more importantly, were different. There appeared to be a processing difference between first choices and later choices, with the utilities estimated from later choices showing a kink when plotted against the utilities estimated from first choice. We attributed the difference to loss aversion on the part of the respondents. The first choices from a full ranking, however, provided greater predictive validity than when respondents were only asked to provide a first choice.

For a while, our recommendation was to capture at least a second choice for each task, but to disregard all choices except for the first choice for each task. There is a cost to such questioning in terms of the additional interview length for respondents. Subsequent investigation, now using HB, has shown the superiority of first choice from a ranking question over simple first choice to be greatly diminished than with an aggregate logit.

Rank methods or allocation are still seen by some as appropriate to account for choices that are context dependent. For example, consider three possible scenarios:

- A corporate IT department might have different PC requirements for engineers than for administrative staff.
- An individual might have different preferences for beer when drinking alone versus dining out.
- A physician might have different prescribing preferences based on the patient's tolerance or susceptibility to side effects.

If there are differences in preferences, we advocate asking situation specific choice tasks. For example, ask "when purchasing computers for engineers, which would you choose?" and in a separate series of choice tasks, ask "when purchasing computers for administrative staff, which would you choose?" Estimate separate utilities for each occasion. Then, by determining the relative frequency of each occasion, the unit of analysis can be changed from a person to an occasion. We have had good success with this approach.

PRACTICAL SUGGESTION #5: ESTIMATE UTILITIES WITH HB.

As discussed above, relying on respondent's choices rather than ratings had involved giving up individual level data. Sawtooth Software provided several methods to develop disaggregate solutions from choice data: k-logit, latent class, and ICE, but their introduction of readily available software to estimate Bayesian models truly revolutionized choice modeling.

Early applications of Bayesian methods to conjoint include Lenk, DeSarbo, Green and Young (1996), and Allenby, Arora, and Ginter (1995). The recent *Sawtooth Solutions* (Summer 2004) reports that 62% of CBC users are using HB to estimate their final model. Their adoption of HB is likely attributable to multiple success stories reported at this conference in the past.

At this conference in 2000, I reported the results (Pinnell, 2000) from six existing choice studies (summarized below) comparing hit rates using utilities developed with HB to hit rates using utilities developed with aggregate logit. In five of the six studies the results were positive, and strongly so. The anomalous sixth study was unique in that it was the only one of the six that was not a full profile study. The sixth study relied on choices from partial profile alternatives.

Comparison of Hit Rates with Disaggregation
Summary of Six Commercial Studies

	Aggregate Logit	**Hierarchical Bayes**	**Improvement**
Study One	75.8	99.5	23.8
Study Two	24.8	79.5	54.7
Study Three	60.5	62.6	2.1
Study Four	61.2	79.3	18.1
Study Five	59.2	78.8	19.6
Study Six	71.9	68.1	-3.8

While the hit rates shown above measure reliability, error in share predictions are a better measure of validity. The studies included in this analysis did not consistently include holdout tasks to allow the estimation of errors in prediction. However, others have shown improvements to share predictions from using HB.

Follow-up research (Pinnell and Fridley, 2001), which focused exclusively on partial profile designs, showed mixed results. Of the nine studies, HB showed a significant improvement in hit rate in three, no effect in two, and a statistically significant degradation in four. Our conclusion at the time was that HB was overfitting with the relatively sparse data that partial profile tasks can produce.

The data sets used in this meta-analysis were not designed specifically for this particular evaluation either and did not include hold-out tasks by which we could evaluate reductions in errors of share predictions.

However, we did find it interesting that the choice tasks with more alternatives per task did better relative to those that had fewer alternatives per task, reinforcing suggestion #3 (above).

Comparison of Hit Rates with Disaggregation
Summary of Nine Partial Profile Studies

Agg. Logit	**Hier. Bayes**	**DIFF**	**Std. Err**	**t-ratio**	**# of Alt./Task**
73.4%	66.5%	-6.9%	0.009	-7.24	3
68.0%	65.7%	-2.3%	0.004	-5.67	3
59.2%	57.0%	-2.1%	0.006	-3.58	3
64.1%	59.6%	-4.5%	0.015	-3.05	3
56.0%	56.5%	0.6%	0.011	0.54	4
52.2%	53.6%	1.4%	0.017	0.82	4
47.6%	51.7%	4.0%	0.017	2.38	4
46.6%	57.6%	11.0%	0.015	7.41	4
32.8%	48.6%	15.8%	0.011	14.84	5

Sawtooth Software, partly in reaction to this finding and working with Peter Lenk, modified their HB software to allow the influence of the prior covariance matrix to be tuned by the researcher. After such tuning, Orme showed that the apparent detrimental effects of HB could be eliminated. However, it was not the case that each partial profile study could be improved with HB but at least parity results were achieved in each case. Which leads to the next suggestion…

PRACTICAL SUGGESTION #6: DON'T USE PARTIAL PROFILE DESIGNS BLINDLY.

The text supporting the previous suggestion provides a cautionary tale regarding partial profile tasks, especially with disaggregate (HB) analysis.

Partial profile tasks provide a mechanism to execute choice tasks with large numbers of attributes. However, they have been shown to produce mixed results. Two new works reported in this volume show partial profile tasks to under-perform relative to full profile choice tasks.

Specifically, Frazier and Jones show that partial profile tasks produced MAE in share predictions nearly 50% larger than the standard full profile with a none option (when averaged across the three reported studies).

Separately, Johnson, Huber and Orme show hit rates for partial profile tasks to be 8 points lower than full profile tasks (71% vs. 63%), and MAEs in share predictions were more than 50% larger for partial profile than for full profile (7.1 vs. 4.6). The authors also showed that derived attribute importances from partial profile are flatter (more nearly equal) than those derived from full profile. This mirrors the result others have seen comparing utilities from ACA (which is also partial profile) to other full profile methods.

One other point on Johnson, Huber and Orme, the authors (almost apologetically) mentioned that the results predicting holdout tasks using utilities derived from partial profile tasks might have been hampered by a methodological bias as the holdouts were full profile. I personally don't believe this represents a methodological bias until we can buy partial profile products in the marketplace. One could argue that full profile stimuli produce a simplification heuristic in respondents and that might be detrimental to our ability to predict in-market behavior. However, it is not clear that the same simplification is not taking place in market. This is a topic where more research would be beneficial.

PRACTICAL SUGGESTION #7: ASSUME PRICE DEPENDENCE, BUT NOT LINEARITY.

Depending upon the category being studied, the base price of a product can vary markedly by market and channel. Moreover, within a category very similar products can have very different prices. These variations should be included in the choice design, with alternative specific pricing for each product, as well as by market and channel, as appropriate for the product category.

We are often asked by clients to include a large number of levels for the price attribute. Our inclination was often to code price as a linear variable (or to transform price such as taking the log of price). Our rationale was that by increasing the number of levels of price, ceteris paribus, we were decreasing the certainty in the parameter estimate of any one level. By solving for a single price attribute, rather than multiple parameters—one for each level—we believed we were smoothing out the error associated with any one level's estimated utility.

However, our practice has largely changed to include price as a part worth function rather than a linearly coded attribute. We have found this to replicate well and better match in-market pricing changes as well as support the notion of psychological pricing. It reinforces the finding Marder (1997) reported. In addition, we have often found it beneficial to include exogenous variables into the price utility—specifically, price cut-off constraints. We, and others (for example, Swait, 1998; Chrzan and Boeger, 1999; Frazier and Patterson, 2000), have incorporated a cut-off penalty into the price variable when estimating utilities. As Johnson and Orme showed, respondents can change their price orientation during a choice exercise. And as Huber et al (1992) suggest, the orthogonal representation of price and brand in conjoint exercises diminishes each one's usefulness to respondents about their respective cues. Including a soft penalty creates a kink in the demand curve, which we believe better mirrors consumer behavior in reality (though not necessarily in hold-out tasks).

PRACTICAL SUGGESTION #8: TEST EACH RESPONDENT'S RELIABILITY.

An earlier suggestion mentioned including one fixed task to test each respondent's reliability. In practice, we commonly hardcode the first task in a choice exercise and exclude it from utility estimation. This task can provide a check on the scaling of the utilities in a probabilistic choice model. This task also provides us one mechanism to test respondents' reliability. To effectively test respondent reliability, it is generally necessary to include more than one holdout task.

There is often a question about developing holdout tasks—should one make holdout tasks easy or hard for respondents? We think the answer is both. Making informed assumptions about the respondents' preferences, but without perfect information, we aim to have the most preferred alternative have about 50% greater probability of choice than the next most preferred. So for pairs, for example, we would target a 60:40 preference ratio between two alternatives.

One can gain a better understanding of respondents' reliability by repeating the same holdout and gauging respondents' reliability on the holdouts themselves. The interested reader is referred to Wittink and Johnson (1991).

To truly test respondents' reliability in practice we use re-sampling techniques. For example, we will estimate utilities using all choice tasks but the first, and use those (generally individual level) utilities to predict each respondent's choice to the first task. We will repeat this using several different tasks as the holdout. We will typically exclude 3 to 5 percent of our respondents due to poor reliability. After excluding these, we will re-run HB, though the results rarely change much.

PRACTICAL SUGGESTION #9: BEWARE (MAYBE JUST BE AWARE) OF SEQUENCE AND ORDER EFFECTS.

In this volume, Rogers and Renken share their results showing the importance of a more realistic shelf-like presentation. We often use a similar store shelf representation of products. We have found (not surprisingly) that the presentation of those products on the screen can influence their relative appeal. As a general rule, if we are doing a store shelf presentation of products, we will employ at least two and generally four rotations, varying what is in the upper left portion of the screen.

Also, when we are testing a wide range of prices, we will typically restrict the range of prices a respondent can see at the beginning of the choice tasks, and then broaden the range of available prices as the tasks continue. We think this is appropriate to gauge "near in" pricing and then test larger, but less likely, pricing changes.

It has been shown (Johnson and Orme 1996; Huber et al 1992) that price can change in importance during the course of a choice exercise. Therefore, building a simulator from such a sequential design involves an additional calibration step. Similarly, comparing the strength of brand preference from early tasks to late tasks will require a similar calibration.

PRACTICAL SUGGESTION #10: FURTHER OUR SCIENCE, AND THINK!

For all that Sawtooth Software has done to help the practicing researcher, and they surely have done much, they have yet to develop software that will think for us. Researchers, by our very nature, have a need for information and a healthy (albeit sometimes overbearing) skepticism. However, when we miss-take our purpose with the rote process, we can forget our true research objective. The fanciest of designs or most elaborate of models will not compensate for other errors or shortcomings in the design or execution of research.

These conferences that Sawtooth Software has produced have done much to develop our collective science. It continues to be up to each of us, though, to continue to test, experiment, and share findings regarding best practices with an open mind and an eager willingness to learn.

Please continue to question conventional wisdom.

REFERENCES:

Allenby, Greg, Neeraj Arora, and James Ginter (1995), “Incorporating Prior Knowledge into the Analysis of Conjoint Studies,” *Journal of Marketing Research*.

Bunch, David, Jordan Louviere, and Don Anderson (1994), “A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes,” Working Paper, Graduate School of Business, University of California, Davis.

Chrzan, Keith, and Leesa Boeger (1999), “Improving Choice Predictions Using a Cutoff-Constrained Aggregate Choice Model,” Presented at INFORMS Marketing Science Conference; Syracuse, NY.

Chrzan, Keith, and Bryan Orme (2000), “An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis,” Sawtooth Software Conference Proceedings.

Frazier, Curtis and Michael Patterson (2000), “Cutoff-Constrained Discrete Choice Models,” Sawtooth Software Conference Proceedings.

Frazier, Curtis and Urszula Jones (2004), “The Effect of Design Decisions on Business Decision Making,” This volume.

Huber, Joel, Dick Wittink, Richard Johnson and Richard Miller (1992), “Learning Effects in Preference Tasks: Choice-based versus Standard Conjoint,” Sawtooth Software Conference Proceedings.

Huber, Joel and Jon Pinnell (1995), “Consistent Differences between Experimental Choice and Ratings-Based Tradeoffs,” Presented at INFORMS Marketing Science; Sydney, NSW, Australia.

Huber, Joel, Klaus Zwerina and Jon Pinnell (1996), “Are Utility Balanced Choice Designs Really More Efficient?” Presented at INFORMS Marketing Science Conference; Gainesville, FL.

Huber, Joel and Klaus Zwerina (1996), “The Importance of Utility Balance in Efficient Choice Designs,” *Journal of Marketing Research*.

Huber, Joel (1997), “What We Have Learned from 20 Years of Conjoint Research: When to Use Self-Explicated, Graded Pairs, Full Profile or Choice Experiments,” Sawtooth Software Conference Proceedings.

Huisman, Dirk (1992), “Price-Sensitivity Measurement of Multi-Attribute Products,” Sawtooth Software Conference Proceedings.

Johnson, Richard and Bryan Orme (1996), “How Many Questions Should You Ask in Choice-Based Conjoint Studies?” Presented at AMA’s ART Forum, Beaver Creek, CO.

Johnson, Richard, Joel Huber, and Lynd Bacon (2003), Adaptive Choice-Based Conjoint, Sawtooth Software Conference Proceedings.

Johnson, Richard, Joel Huber, and Bryan Orme (2004), “A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs,” This volume.

Lenk, Peter, Wayne DeSarbo, Paul Green, and Martin Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Part-worth Heterogeneity from Reduced Experimental Designs," *Marketing Science*.

Louviere, Jordon, and George Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data." *Journal of Marketing Research*.

Marder, Eric (1997), The Laws of Choice: Predicting Customer Behavior. Free Press.

Mulhern, Michael (1999), "Assessing the Relative Efficiency of Fixed and Randomized Experimental Designs," Sawtooth Software Conference Proceedings.

Pinnell, Jon (1994), "Multistage Conjoint Methods to Measure Price Sensitivity." Presented at AMA's ART Forum, Beaver Creek, CO.

Pinnell, Jon and Sherry Englert (1997), "Number of Choice Alternatives in Discrete Choice Modeling," Sawtooth Software Conference Proceedings.

Pinnell, Jon (1999a), "Depth of Probing, Allocation and Rank Order Methods in Consumer Choice Analysis," Presented at INFORMS Marketing Science Conference; Syracuse, NY.

Pinnell, Jon (1999b), "Should Choice Researchers Always Use 'Pick One' Respondent Tasks?" Sawtooth Software Conference Proceedings.

Pinnell, Jon (2000), "Customized Choice Designs: Incorporating Prior Knowledge and Utility Balance in Choice Experiments," Sawtooth Software Conference Proceedings.

Pinnell, Jon and Lisa Fridley (2001), "The Effects of Disaggregation with Partial Profile Choice Experiments," Sawtooth Software Conference Proceedings.

Rogers, Greg and Tim Renken (2004), "The Importance of Shelf Presentation in Choice-Based Conjoint Studies," This volume.

Swait, Joffre (1998), "A Model of Heuristic Decision Making: The Role of Cutoffs in Choice Processes," Working Paper, University of Florida.

Willaims, Peter and Denis Kilroy (2000), "Calibrating Price in ACA: The ACA Price Effect and How to Manage It," Sawtooth Software Conference Proceedings.

Wittink, Dick and Richard Johnson (1991), "Estimating the Agreement Between Choices Among Discrete Objects and Conjoint-Ratings-Based Predictions After Correcting for Attenuation," Working Paper, Cornell University.

.

THE "IMPORTANCE" QUESTION IN ACA: CAN IT BE OMITTED?

W. CHRISTOPHER KING,
AARON HILL, BRYAN ORME
SAWOOTH SOFTWARE, INC.

BACKGROUND

Adaptive Conjoint Analysis (ACA) was developed by Sawtooth Software's founder, Rich Johnson, based on his early work in tradeoff matrices dating back to the late 1960s. During the 1970s, Rich and his colleagues refined the technique by applying early versions of ACA to various real-world business problems. In the late 1970s, the first versions of ACA were proprietary and developed on the Apple II platform. In 1983, Rich founded Sawtooth Software and released ACA as a commercial software product in 1985. ACA soon became the most widely-used conjoint analysis program and technique in the world (Wittink and Cattin, 1989). It held that position until about the year 2000, when choice-based conjoint (CBC) techniques became more widely used (Sawtooth Software, 2003a). Even still, ACA is a widely used method for studying more attributes than is prudent for full-profile approaches. It is especially valuable for product design research and segmentation studies. In this paper, we assume the reader is already familiar with the basic mechanics of ACA. For more information, please see the "ACA Technical Paper" available for download at www.sawtoothsoftware.com in the Technical Papers Library.

ACA has benefited from scrutiny over the years, particularly from leading academics. In a paper published in 1991 (Green, Krieger, and Agarwal, 1991), the authors criticized ACA's self-explicated "importance" question. They argued that the attribute importance ratings employed in ACA's "priors" section were "…too coarse; only four response values [were] permitted." Later that year, William McLaughlan conducted a split-sample research study that compared a 9-pt and a 4-pt importance question within ACA, finding virtually no difference in the quality of the final part worth utilities (McLaughlan, 1991). Green, Krieger, and Agarwal also questioned the scale compatibility of self-explicated importance ratings and the subsequent conjoint (pairwise) judgments. They argued that it was not proper to combine both types of responses (as dependent variables) within the same regression tableau. Based on that suggestion, Sawtooth Software modified the way ACA utilities were estimated under OLS to address these potential scale incompatibilities. The results were quite similar to the earlier estimates, though time has shown that neither utility estimation procedure clearly dominates. As an unexpected benefit, it was later discovered that the newer OLS method seemed to slightly reduce the number of level effects problem (Orme, 1998). Since the late 1990s, we at Sawtooth Software have recommended hierarchical Bayes (HB) estimation as the gold standard for ACA, though the modified OLS procedure, nearly identical to that developed by Johnson in the early 1990s, remains the default option in the base system.

One prevalent criticism of ACA has been that the self-explicated importance judgments probably reduce the discrimination among attributes, leading to "flatter" importances in the final utility estimates. Furthermore, many researchers have reported that respondents can have difficulty understanding the importance questions, and have a tendency to over-use the upper portion of the rating scale. Given the availability of HB and its proven superiority for estimating quality part worth utilities using less information from each individual (due to its powerful information-borrowing mechanism), we have wondered whether we might achieve equally good (or even better) results by skipping the importance questions altogether. This research seeks to answer this question. As will be shown, it appears for most research applications that the importance question *can* be skipped (if applying HB estimation), resulting in shorter questionnaires and less confusion and burden for respondents.

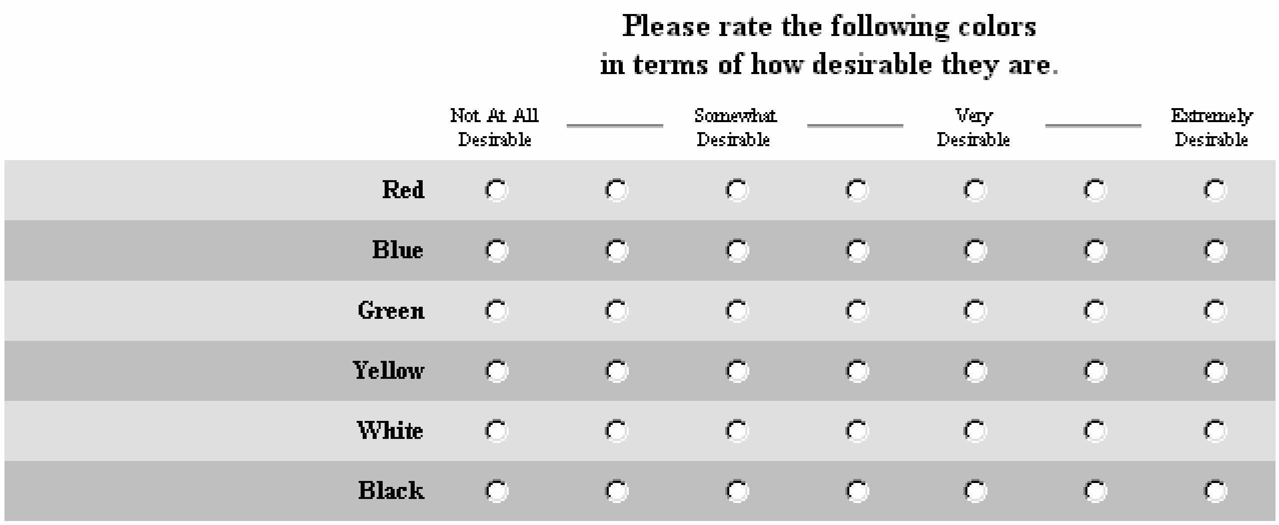
PARTS OF THE ACA INTERVIEW

ACA is a multi-stage hybrid conjoint instrument that features the following sections:

1. Unacceptables (optional, and rarely used)
2. Rating/Ranking of levels within attributes
3. Importance ratings of attributes
4. Pairs (the "conjoint" section)
5. Calibration Concepts (optional section)

The first three sections represent a simple self-explicated scaling exercise and are often called the "Priors." The first section allows respondents to indicate which levels of an attribute are "unacceptable," even if these levels were included in a product that was excellent in all other ways. This section is not available in ACA/Web and is rarely used, so we skipped it for the purposes of this study. In the second section, the respondent is asked to rate/rank the levels of any attributes for which we cannot know ahead of time the rational order of preference (such as brand, colors, etc.). Figure 1 is an example of a ratings style question.

Figure 1

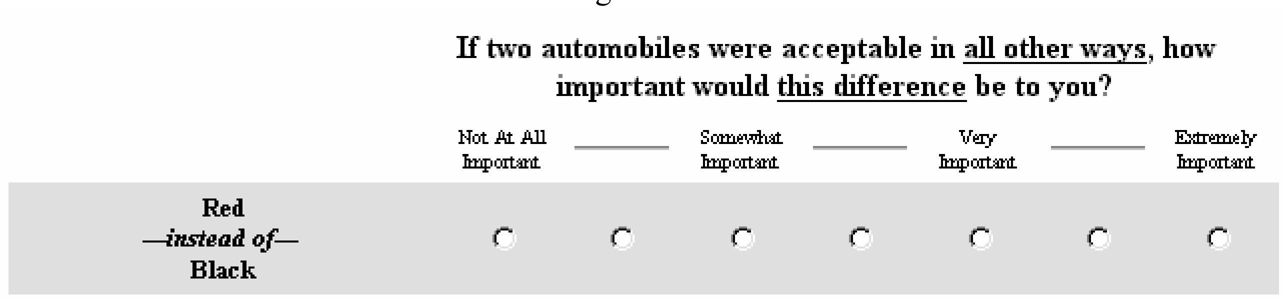


Please rate the following colors in terms of how desirable they are.

	Not At All Desirable		Somewhat Desirable		Very Desirable		Extremely Desirable
Red	○	○	○	○	○	○	○
Blue	○	○	○	○	○	○	○
Green	○	○	○	○	○	○	○
Yellow	○	○	○	○	○	○	○
White	○	○	○	○	○	○	○
Black	○	○	○	○	○	○	○

Next follow the Importance questions, which are the focus of this research. After the first ranking/ratings section (or based on other pre-specified attributes where we know ahead of time the rational order of preference), we display the best and worst levels for each attribute. We ask respondents to indicate how important it is for them to receive the best level instead of the worst level (there are other ways to phrase this question). Figure 2 shows an example, assuming the respondent has rated Red as the most preferred level and Black as the least preferred:

Figure 2



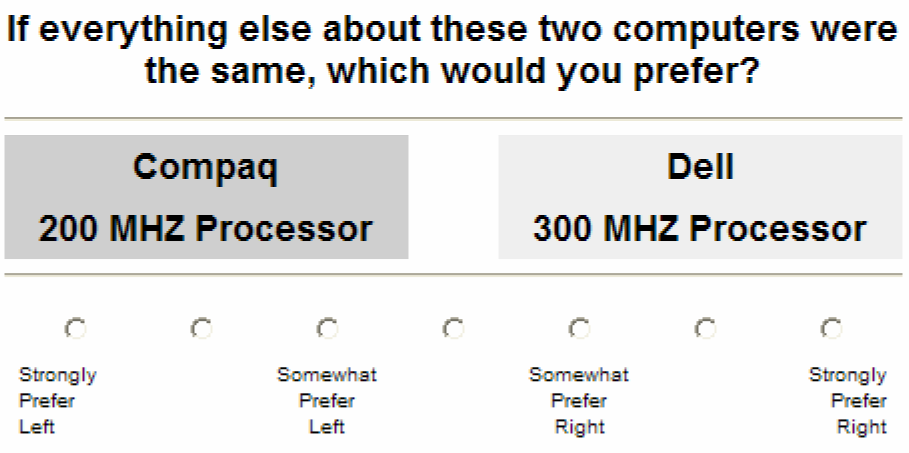
If two automobiles were acceptable in all other ways, how important would this difference be to you?

	Not At All Important		Somewhat Important		Very Important		Extremely Important
Red ***—instead of—*** **Black**	○	○	○	○	○	○	○

The importance rating is used to prioritize the order that attributes are investigated in the conjoint "Pairs" section. The metric information from this rating scale is used to develop the rough "prior" utilities used for the utility balancing in ACA's Pairs design algorithm, and also in the final utility estimates. The importance question is currently a *required element* in the ACA survey, though we investigate an experimental version of ACA that drops this section.

To this point in the ACA interview, no "conjoint" questions have been asked. Figure 3 shows an example conjoint pair, which is the workhorse of the ACA interview and that provides the critical tradeoff information needed for refining the part worth utilities. After each pair is asked, the information augments that previously collected (updates the prior utilities), and is used to select the next question asked of the respondent (respondents typically answer 12 to 30 Pairs).

Figure 3



If everything else about these two computers were the same, which would you prefer?

Compaq **200 MHZ Processor**	**Dell** **300 MHZ Processor**

○	○	○	○	○	○	○
Strongly Prefer Left		Somewhat Prefer Left		Somewhat Prefer Right		Strongly Prefer Right

After the Pairs section, an optional Calibration Concepts section may be included to gauge each respondent's purchase likelihood for different product concepts. Purchase Likelihood judgments were not important to this research, so we skipped this final section.

CONCERNS REGARDING THE IMPORTANCE QUESTION

Again, researchers have expressed a number of valid concerns regarding the Importance question, including:

- For some attributes, it is difficult to understand what is being asked—especially a binary attribute which either has or doesn't have a feature.
- Researchers sometimes have difficulty with formatting and wording.
- Lack of context can become an issue, particularly with attributes exploring sensitive topics (i.e. salary).
- Respondents do not answer in a truly ratio-scaled manner (a rating of "4" does not mean that the attribute is twice as important as one rated a "2"), yet the standard OLS utility estimation routine in ACA assumes we have captured ratio data.
- Respondents don't discriminate enough among the attributes using the provided rating scale. The responses typically load on the upper part of the scale.
- In studies with a large number of attributes, the importance questions can become tedious. (Perhaps a respondent's time would be better spent answering more Pairs questions. The advent of HB potentially removes the need for individual self-explicated importance information.)
- Importance scores may artificially constrain and usually do flatten the differences among attributes.
- It is difficult for respondents to establish the proper scaling (framing) for the first few attributes (though this problem might be reduced by first reviewing the list of attributes prior to starting the ACA process).

In their as-of-yet unpublished work "Adventures in Conjoint Analysis" (Krieger, Green, Wind, 2004), the authors state:

> "Considerable research, both conceptual and empirical, suggests that a respondent's self explicated attribute level desirability ratings are much more reliable than self explicated attribute importances.
>
> "A recent examination of reliability in a conjoint setting indicated that the median test/retest correlation for attribute-level desirabilities was 0.925, while that for importance ratings was only 0.574."

ALTERNATIVES TO THE IMPORTANCE QUESTION

- Initially, set all importance scores equal (for design purposes), but let the final part worths be estimated using level ratings and conjoint Pairs data.
- Use starting importance scores (for design purposes) based on the average of previous respondents' final derived importance values.
- Bayesian updating routine.

- The researcher might establish importance scores (for design purposes) based on previous research, focus groups, sales data, or client whim.
- Derive importances using the observed spread among self explicated ratings within each attribute.
- Use a different conjoint method (partial-profile choice?)

Research Study

Because of the confidentiality of the information, we cannot disclose the sponsor of this research or details regarding the product or features. We can say that the purpose of the research was to compare accessibility and content issues related to an existing electronic information service over the internet. We studied 20 attributes, where roughly half were binary (had two levels), and the others ranged in complexity to a maximum of five levels per attribute. None of the attributes had *a priori* level order, so respondents had to provide level ratings for all 20 attributes in the study. Graphics were used to reflect the attribute levels to make the content (which at times was quite complex) more accessible to the respondents.

We randomly split respondents among three design cells. The study was conducted over the internet, and respondents were recruited in a variety of ways, including an opt-in list and banner ads, according to the client's specifications. A total of 1,419 respondents were used in the final analysis, after data cleaning. Respondents were discarded if they completed the survey too quickly (fastest 5%), if their Pairs ratings showed little discrimination (>=85% identical), or if they did not complete the holdout tasks. Less than 15% of the data was discarded.

Respondents in each cell received different ACA treatments. The three design cells were as follows:

> **Cell 1 (n=463):** Standard ACA procedure, as currently available in ACA/Web v5. The questionnaire included 20 level ratings grid questions (one for each attribute), 20 Importance questions, and 14 Pairs questions (6 at 2 attributes and 8 at 3 attributes).
>
> **Cell 2 (n=508):** Modified ACA where all importance ratings were set to be equal (for Pairs design purposes). **The Importance section was skipped**, and final part worths were developed using HB estimation (fitting the Pairs, and constrained by within-attribute level ratings). Updating of part worths during the Pairs section used the standard ACA procedure. Attribute importance was therefore recalculated after each Pairs question, so the initial importances had less influence as the survey progressed. The questionnaire included 20 level ratings grid questions (one for each attribute), no Importance questions, and 20 Pairs questions (8 at 2 attributes and 12 at 3 attributes).
>
> **Cell 3 (n=448):** Modified ACA where the importance ratings were projected for each respondent based on the derived importances from respondents that had already completed the survey. All other aspects were identical to Cell 2.

In addition, we included partial-profile holdout choice questions in the survey. After completing the ACA sections, each respondent answered six of a total of eighteen possible holdout tasks (selected randomly). Each task included 4 of the 20 attributes, with 3 alternatives per task. Across the tasks, all 20 attributes and all levels were represented. To conserve time, no

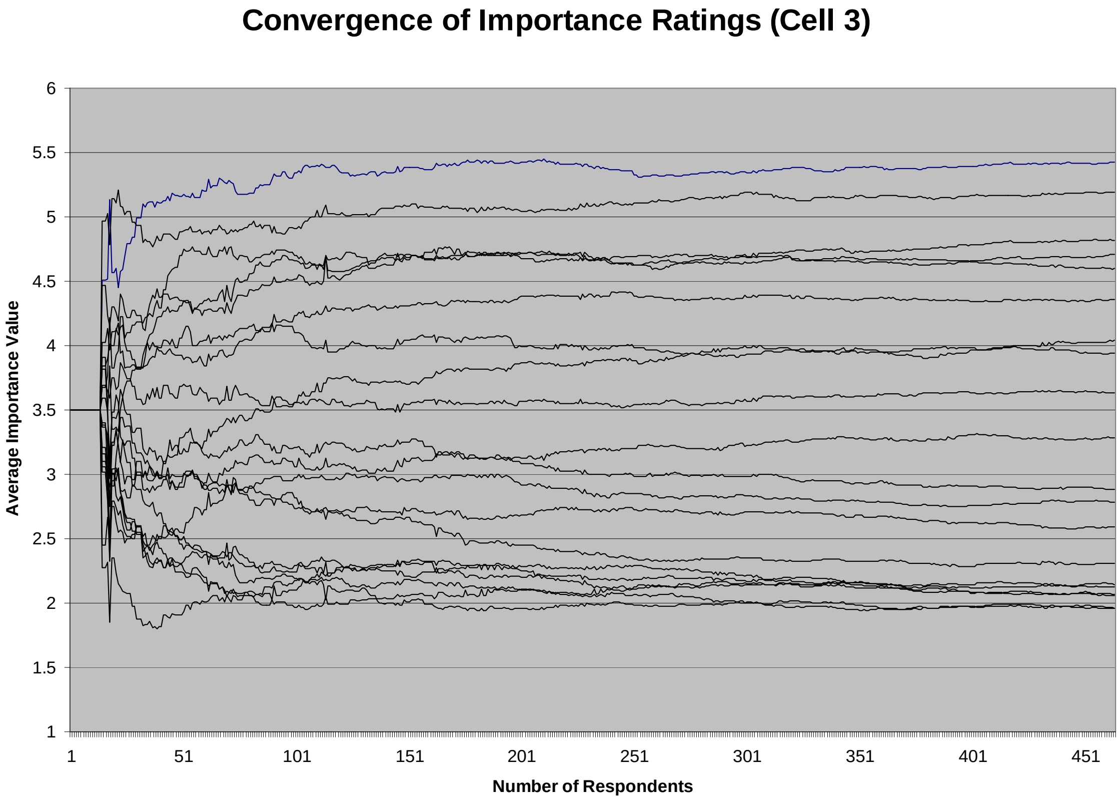
Calibration Concepts were asked (Hill, Baker & Pilon, 2001, found that calibrating utilities did not help share predictions).

Prior to fielding the study, we hypothesized that the Importance questions might not be very useful, and might actually misinform utility estimation. We hypothesized that the time savings from skipping the Importances could be used more profitably in asking additional conjoint Pairs questions. We expected that the importances based on prior respondent averages would provide reasonable starting points for design purposes in the Pairs. We also hypothesized that skipping the Importances would lead to higher completion rates.

Convergence of Attribute Importances for Cell 3

As previously described, Cell 3 of our design used the utility information from previous respondents as initial importance scores for design purposes for the conjoint Pairs section. Because there was no information available when the first respondent entered the survey, we initially set the importances equal. After 12 respondents, we began to use the simple average of all previous respondents to initialize the design procedure for the Pairs algorithm (for the next respondent). We were curious to know how quickly the averages would stabilize for the respondent population. Figure 4 displays the average importance ratings for all 20 attributes by number of respondents who had taken the survey.

Figure 4



As can be seen, the average importance scores stabilized quite rapidly. After 50 respondents had completed the survey, only minor switches in the rank order of attribute importances can be seen. We should note that we used the prior group importances in a deterministic manner for arranging the attributes in rank-order (for purposes of the round-robin sequence used in designing Pairs with two attributes at a time). In hindsight, we recognize that we may have done slightly better by using the prior importances in a probabilistic way for establishing the initial rank-order of attributes for these first Pairs questions. This would have lead to greater variation in the design for the first Pairs (shown at two attributes at a time) across respondents.

Time to Complete and Completion Rates

Because we were interviewing using computers, we were able to capture the time (in seconds) to complete each section of the survey. As shown in Figure 5, it took 2.6 minutes for respondents in Cell 1 to complete the 20 Importance questions. Cells 2 and 3 omit the Importance questions and substitute six additional pairs. We can further see in Figure 5 that the overall interview time for the ACA section was *lower* for Cells 2 and 3 relative to Cell 1 (though the Pairs section took slightly longer to complete). We could have substituted even more conjoint Pairs questions to match the interview time of Cell 1.

Figure 5

	Median Overall Survey Completion Time	**Median time per ACA section**			
(in Minutes)		Ratings	Importance	Pairs	Total ACA
Cell 1 (14 pairs)	23.0	6.5	2.6	2.9	11.9
Cell 2 (20 pairs)	21.4	6.5	0.0	4.0	10.5
Cell 3 (20 pairs)	20.9	6.6	0.0	4.2	10.8

More than half of the people that started the survey ended up quitting at some point (there was no monetary incentive to complete the survey). This figure is disappointing. The percent of respondents that completed the survey once they started were 39.6%, 42.9% and 42.6% for Cells 1, 2, & 3 respectively. The shorter interviews (Cells 2 and 3) indeed had directionally higher completion rates, but the margin of difference is relatively small.

Part Worth Utility Analysis

We used the following methods to estimate the part worth utilities.

ACA/OLS: Standard OLS estimation, as offered in ACA (only applicable for Cell 1 respondents).

ACA/HB with Importance Scores: ACA/HB procedure, fitting the metric information of the conjoint pairs, constrained by both the importance scores and the within-level ratings (only applicable for Cell 1 respondents).

ACA/HB without Importance Scores: ACA/HB procedure, fitting the metric information of the conjoint pairs, constrained only by the within-level ratings*.

(*Some ACA/HB users may wonder how we employed ordinal constraints using only the within-attribute ratings. To do this, we modified the .ACD file, and artificially set the importance scores to be equal for all 20 attributes. This "trick" has worked well for us in a few data sets, where the self-explicated importances seemed suspect.)

FINDINGS—HIT RATES

As previously explained, respondents completed six (of eighteen) choice tasks including three alternatives each (shown in partial profile). We asked respondents to indicate the favored and least favored alternative within each task (a best/worst approach).

Using the part worth utilities, we predicted which product alternatives we would have expected respondents to choose and reject within the holdout choice tasks. We scored each product alternative (by adding the part worth utilities corresponding to the levels included in that alternative) and projected that the respondent would choose the alternative with the highest utility and reject the alternative with the lowest utility. We compared predicted choices to actual choices, and scored the "hits" and "misses." The hit rates, by design cell and different utility estimation methods, are given in Figure 6.

Figure 6

Hit Rates	Cell 1	Cell 2	Cell 3
ACA/OLS	67.1 (0.63)	N/A	N/A
ACA/HB with Importance scores	67.9 (0.64)	N/A	N/A
ACA/HB without Importance scores	66.7 (0.61)	66.2 (0.61)	65.6 (0.66)

(Standard errors are shown in parentheses. A difference of 1.7% in hit rate is required between groups for significance at 95% confidence.)

The best hit rate occurs for traditional ACA (Cell 1) with ACA/HB estimation, using the importance scores. However, this is only marginally better than the design cells that omitted the Importance question (and substituted six more Pairs questions). Three key points occur to us: 1) Omitting the importances and substituting six more pairs (an overall time savings) results in nearly the same hit rates, 2) Previous research suggests that Importance ratings reduce measurement error within each individual at the expense of bias, which often benefits hit rates, 3) Hit rates are not as important to managers as accuracy of share predictions, which we'll report in the next section.

FINDINGS—SHARE PREDICTIONS

As mentioned previously, share prediction accuracy is of most importance to managers. To compute share prediction accuracy, we used the Randomized First Choice simulation model as offered in Sawtooth Software's market simulator (though the standard logit simulation approach would produce very similar findings). We used part worth utilities for the respondents in each cell to predict the overall shares of first choices for all respondents and all eighteen holdout choice tasks. This, importantly, is a more stringent test of internal predictive validity, as it holds out both choice tasks *and* respondents. We tuned the shares (using the Exponent, or scale factor) to achieve highest predictive accuracy within each cell. This ensures that observed differences in predictive accuracy are due to substantive differences in the utilities rather than arbitrary differences in scale.

Figure 7 reports the predictive accuracy in terms of Mean Absolute Error (MAE). MAE characterizes by how much (on average) the predicted shares differ from the actual choice shares. Lower MAE values indicate more accurate predictions.

Figure 7

Share Predictions (MAE)	**Cell 1**	**Cell 2**	**Cell 3**
ACA/OLS	7.1%	N/A	N/A
ACA/HB with Importance scores	6.8%	N/A	N/A
ACA/HB without Importance scores	6.7%	6.4%	5.9%

Cells 2 and 3 (which omitted the Importance questions and substituted six additional Pairs) show the best predictive accuracy. The additional Pairs questions seem to benefit this analysis, and the lack of individualized importance information for designing the Pairs hasn't seemed to hurt. (We cannot, unfortunately, test for statistically significant differences). We expect that the margin of victory for omitting the Pairs would have been even greater if additional pairs had been asked with the time savings from skipping the Importance questions.

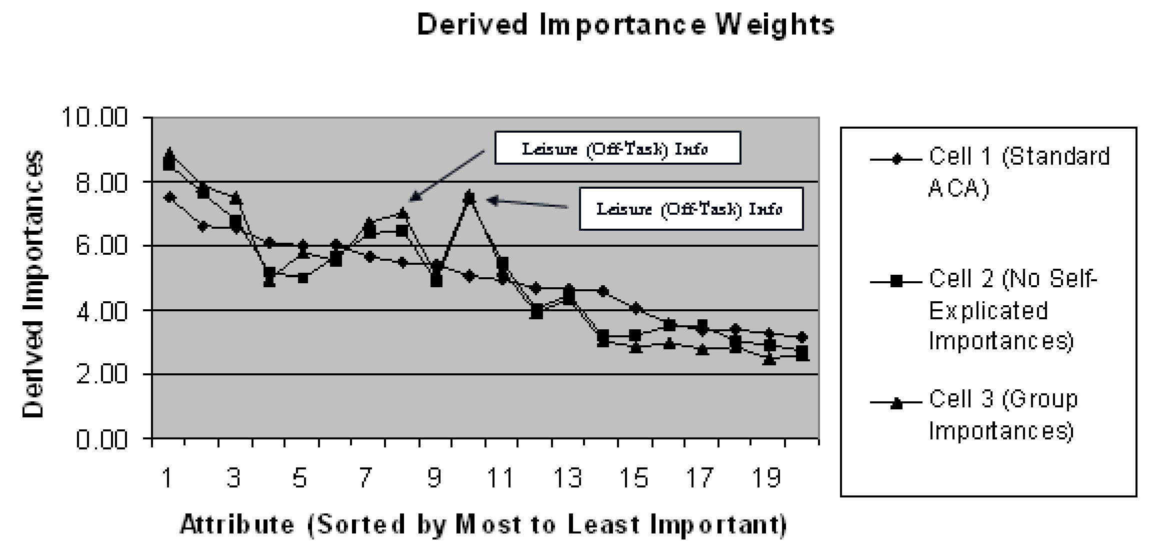
It is important to note that the use of the Importance ratings in Cell 1 seem to provide no benefit for share prediction accuracy. Throwing out the Importance ratings for Cell 1 results in directionally lower errors in share prediction (6.7%) than when using the Importance information during HB estimation (6.8%).

As has been seen in other studies (Sawtooth Software, 2003b), HB estimation provides more accurate predictions than the OLS procedure.

FINDINGS—IMPORTANCE SCORES

We may wonder why the modified versions of ACA that skipped the Importance questions resulted in more accurate share predictions. Were the part worth utilities more precise due to the additional pairs asked? Was the information provided by Importance ratings different from that implied by the tradeoffs in the conjoint pairs (and holdout choice tasks)? We compared the *derived* Importance scores (by comparing the best to worst utilities for each attribute, and percentaging) for the three design Cells in Figure 8. In each case, we used HB estimation. We sorted the attributes in order of derived Cell 1 importances.

Figure 8



Of note:

- The importances for Cell 1 (standard ACA) are “flatter” than for Cells 2 and 3.
- The derived importances are nearly identical for Cells 2 and 3 and are quite different in some instances from Cell 1.

This suggests that the information provided by the self-explicated importances not only flattens the differences in derived importance among the attributes (relative to the conjoint Pairs information), but also leads to different conclusions regarding the order of importance of attributes. It turns out that the two attributes that show the highest negative deviations for Cell 1 importances versus Cells 2 and 3 involved leisure (off-task) information. It is plausible that it was not socially desirable to state (in a self-explicated importance question) that the off-task leisure information was as important as on-task information, and that the truer importance was revealed in the tradeoff tasks. Anecdotal evidence supports the derived importances suggested by the modified ACA procedures. The client commented that previous research at the firm suggested that the off-task leisure information was quite important.

CONCLUSIONS

This research suggests that for most situations ACA researchers may be better off by skipping the Importance questions and adding more Pairs questions. For this study, using the Importance questions during estimation only provided modest benefit in terms of hit rates. And, the Importance questions may actually hurt share predictions.

In summary, skipping 20 importance questions, but adding six additional pairs:

- provides nearly the same overall hit rate accuracy,
- seems to improve the share prediction accuracy,
- yields “steeper” and <u>different</u> derived importances,
- reduces the average length of the survey by over one minute.

Using aggregate information as proxy for prior importances yields directionally lower hit rates but directionally better share predictions. This experiment has not conclusively shown that this more elaborate procedure adds any value, though it is theoretically pleasing. Perhaps applying the prior group information in a non-deterministic way in the first stage of the Pairs design might tip the scales in favor of this approach rather than naïve prior importances (Cell 2).

Using within-attribute ratings information seems enough to initially provide informative (non-dominated) tradeoffs in ACA.

FUTURE DIRECTIONS:

Based on this research, it would be valuable to:

- Study other methods of incorporating prior importance scores (particularly, using prior group importances in a non-deterministic manner),
- Validate the findings with additional studies,
- Integrate the ability to drop the Importance question within ACA software (it is currently impossible to do without customization).

REFERENCES

Hill, Aaron, Gary Baker and Tom Pilon (2001), “A Methodological Study to Compare ACA Web and ACA Windows Interviewing,” *Sawtooth Software Conference Proceedings*: Sequim, WA.

Krieger, Abba, Paul Green, and Yoram Wind (2004), “Adventures in Conjoint Analysis: A Practitioner’s Guide to Tradeoff Modeling and Applications,” Unpublished Manuscript.

Green, Paul, Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions", *Journal of Marketing Research*, (May), 215-22.

McLaughlan, William (1991), “Scaling Prior Utilities in Sawtooth Software’s Adaptive Conjoint Analysis,” *Sawtooth Software Conference Proceedings*: Ketchum, ID.

Orme, Bryan (1998), “Reducing the Number-of-Attribute-Levels Effect in ACA with Optimal Weighting,” Sawtooth Software Technical Paper, available at www.sawtoothsoftware.com.

Sawtooth Software (2003a), “More Evidence CBC Is Most Popular Conjoint Method,” *Sawtooth Solutions*, Summer 2003.

Sawtooth Software (2003b), “ACA/Hierarchical Bayes v2.0 Technical Paper,” Technical Paper available at www.sawtoothsoftware.com.

Wittink, D.R. and P. Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53 (July) 91-6.

COMMENT ON KING, HILL AND ORME

RICH JOHNSON
SAWTOOTH SOFTWARE, INC.

My thanks to Chris, Aaron and Bryan for solid work that points in a useful direction for future development of ACA.

I was not personally involved in this research, so I was pleased to see that their results verified those of other investigations of importance questions. They are also compatible with results from a study that I'll be reporting later in this conference.

ACA was invented to permit conjoint estimation at the individual level in the context of large numbers of attributes and levels. It seemed to make sense to ask respondents to provide desirability ratings for levels within each attribute, and some kind of importance ratings of attributes, so those could be put together to provide initial estimates of partworths.

At the time, researchers seemed fond of another kind of attribute importance question, such as:

> "When choosing an airline how important is safety?"
>
> "How important is on-time performance?" etc.

We knew that generic questions like this would not be useful for our purposes, because we couldn't know what respondents had in mind when answering them. I took particular pride in the sort of importance question used in ACA, which we called a "regrets" question, that asked a respondent to consider the difference between two specific attribute levels and register the amount of his or her regret in having to settle for one level instead of the other.

But it turned out that of all the kinds of questions asked in ACA, the importance questions were among the most problematic. Respondents who are not highly motivated or who are not very good at abstract thinking have a hard time producing good information with them. Such problems may occur especially with Web interviewing.

Before the availability of HB estimation, importance questions were an essential part of the ACA interview. But HB's ability to improve estimation for an individual by borrowing information from others appears to have overcome the need for importance questions. This is good news because, as the authors have shown, we can now get results that are the same or perhaps better with less interview time and with less difficulty for our respondents.

DERIVED MEASURES OF PERFERENCE/IMPORTANCE

SCALE DEVELOPMENT WITH MAXDIFFS: A CASE STUDY

LUIZ SÁ LUCAS
IDS-INTERACTIVE DATA SYSTEMS

ABSTRACT

This paper describes how MaxDiffs can be applied to evaluate the degree of seriousness of several offenses/crimes. This work also explores if the way people envisage reality impacts how they evaluate the degree of seriousness of offenses. The different visions of reality are obtained through the segmentation of the sample also using MaxDiffs.

INTRODUCTION

Scales are known to be a problem in Marketing Research. People tend to use our discrete scales in different ways: some tend to use the upper part of the scale (yea-sayers), the middle of the scale (middle-of-the-roaders) or even the lower part (nay-sayers). Some use all the scale, some use a smaller range. This can be particularly damaging, for example, in multi-country studies. Part of the MR folklore is a hypothetical segmentation study where two segments were found: Brazilians (yea-sayers) and Germans (middle-of-the-roaders).

Two approaches have been used more recently to overcome different scale usage by respondents:

- Rossi et al. (2001) developed a MCMC method that estimates respondent level position and range in the scale, making it possible to correct for these biases
- MaxDiffs, whose application is the object of this paper.

Our example, the degree of seriousness of offenses, is a difficult problem due to its abstract character. It is quoted, for example, in Marketing Research textbooks such as Aaker et al.(1988). Here we describe how a technique called Psychophysics could be applied to the problem, its shortcomings and how we think MaxDiffs can overcome the difficulties.

PSYCHOPHYSICS

The Degree of Seriousness of Offenses problem is described, for example, by Lodge (1981). In this work, that deals with the quantitative measurement of opinions, Lodge argues that:

- Ordinal discrete scales are widely applied and have a long history, dating back at least to 150 B.C. – Hipparchus used a six-point scale to judge the brightness of the stars
- However, when we use them, information is lost because of the limited resolution of categories:
    - suppose a five-point agreement scale (disagree strongly/disagree somewhat/.../ agree somewhat/agree strongly) is used, and the respondent agrees somewhat with

statements A and B, but a little more with A than with B, although he/she does not agree strongly with A. This respondent has no way to show it in this five-point scale

- Category scales represent only an ordinal level of measurement, denying researchers legitimate access to more powerful methods based on interval assumptions.

Lodge (1981) then describes how Psychophysics, a technique created in the middle of the XIXth century, could be applied to the problem of quantitatively measuring opinions.

Psychophysics is based on the way human observers are accurately able of using proportional judgments of stimulation levels for virtually all aspects of the five senses: loudness of sounds, brightness of light, sweet and sour tastes etc.

Typically results in Psychophysics obey the so-called ***Power Law***:

$$Response = Score = \alpha\, Stimulus^{\beta}$$

Following this law, the ratio between two stimuli should be related in a direct way to the ratio between the scales that measure them. If we double the stimulus, maybe we are not going to double the response measured in the scale (the score), since usually this beta coefficient is smaller than one.

Using this technique, data for the problem would be collected as described below (Lodge, 1981):

> *I would like to ask your opinion about how serious you think certain crimes are. The first situation is 'a person steals a bicycle parked on the street'. This has been given a score 10 to show its seriousness. Use this situation to judge all others.*
>
> *For example, if you think a situation is 20 TIMES MORE serious than the bicycle theft, the number you tell me should be around 200.*
>
> *If you think it is HALF AS SERIOUS the number you tell me should be around 5, and so on.*

SO COMPARED WITH THE BICYCLE THEFT, HOW SERIOUS IS:

- a parent beats his young child with his fists. The child requires hospitalization;
- a person plants a bomb in a public building. The bomb explodes and 20 people are killed

Several psychophysical studies like these were performed in the 60s and 70s:

- Sellin and Wolfgang – *The Measurement of Delinquency* (1964): juvenile court judges, parole officers, college students (representative of Everyman...) were instructed to assign numbers to sentences describing criminal offenses from 'stealing and abandoning the vehicle' to 'rape' and 'murder'
- The study was replicated several times in the United States, Canada and England (Figlio, 1978). We should mention here a study described by Figlio (1978), performed by the U.S. Bureau of the Census in 1977, with a sample of around 54,000 respondents and 200 descriptions.

Using Psychophysics, although we have a standard for position (10 for bicycle theft), we have other problems:

- We do not overcome the problem of conservative vs. exaggerated use of the scale/number estimation
- The use of the technique assumes respondents have the competence to make the proportional judgments. Can we be sure of that when we have illiterate respondents?

If we cannot be sure illiterate people are able to make proportional judgments, we can safely assume that they are able to say an offense is more serious than another, and this leads us naturally to MaxDiffs.

MAXDIFFS

Returning to the ordinal five-point agreement scale:

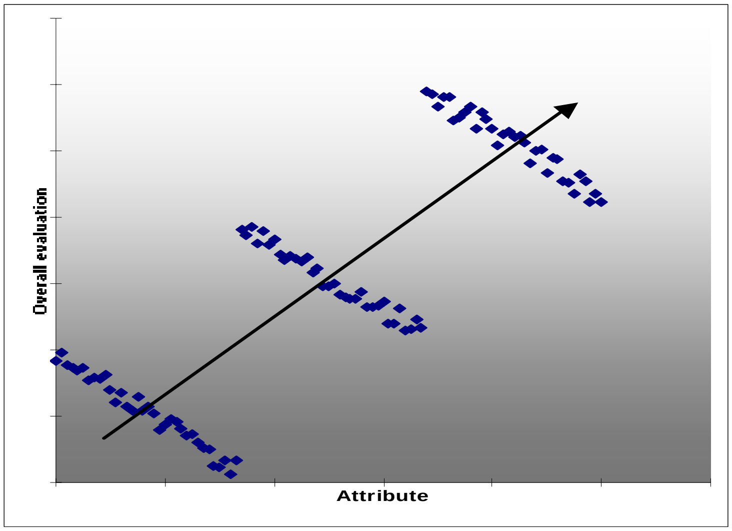
- Disagree strongly
- Disagree somewhat
- Neither agree nor disagree
- Agree somewhat
- Agree strongly

We can overcome the ordinal scale problem turning the scale into a semantic-differential one like this:

Disagree strongly □ □ □ □ □ Agree strongly

Here, although we have interval scales, neither the problem of limited resolution of categories is overcome nor the heterogeneity in the use mentioned before (use of extremes vs. middle of the scale) is prevented.

Another problem can be quoted. Let's take a look at Graph 1 below:



Graph 1: Scale Usage Heterogeneity – Yea/Nay/Middle-of-the-road saying

From left to right, we have first the nay-sayers, then the middle-of-the-roaders, and finally the yea-sayers on top right. In each group, the overall satisfaction decreases as the attribute score gets higher.

If we fit a single straight line to all data we obtain a positive coefficient for the regression line. This positive relationship is only due to scaling effects. The regression line above has an R^2 of 0.65.

Surely a careful analyst would graph data before performing a simple analysis like that, but these effects will not always be as evident as this extreme example, and can be hidden behind, for example, a 20-variable multiple regression application. So we really need a scaling technique that prevents this kind of problem. MaxDiffs is a natural candidate.

The technique is described by Cohen et al. (2002 and 2003). It is also presented by Chrzan (2004) in this Sawtooth Software Conference.

Briefly we can mention that:

- MaxDiffs originated from Best-Worst Conjoint Analysis (Louviere et al.(1992, 1995 and 2002), quoted by Cohen and Neira (2003))
- In each choice task one chooses not only the Best option but also the Worst one
- Following conventional Conjoint Analysis attributes are presented in choice tasks (groups of statements/attributes) where one chooses:
    - the sentence with which he/she agrees most/least;
    - the most/least important attribute;
    - the most/least serious offense;
    - etc.

- In the choice tasks, the respondent behaves as if in each choice task he/she is examining all the pairs of statements/attributes, choosing the pair with *Maximum Difference* from the best-worst point of view
- In each task, a trade-off is imposed on the respondent, so that it's not possible for the respondent to agree with/consider as most important etc. all the attributes: one should always be selected
- There is only one way to identify the best and the worst: there is no heterogeneity in the choice/usage of scale

SERIOUSNESS OF OFFENSES

For our exercise, also described in Sá Lucas (2004), we developed a set of 25 offenses, from "Disturbing the neighborhood with loud, noisy behavior" to "Planting a bomb in a public building. The bomb explodes and 20 people are killed". The full list can be found in the tables presented in the following pages. A particular problem concerns some of the offenses. Let's take the set:

- Stealing $10
- Stealing $100
- Stealing $1,000
- Stealing $10,000

Should we expect that 'stealing $10,000' is 1,000 times more offensive than 'stealing $10'? Surely not. But then, how much?

First let's describe the way we collected data:

- I will show you some crimes/offenses that can be committed by several people
- These crimes or offenses will be presented in groups of four sentences
- For each group of statements, please choose the crime/offense you think is the most serious and the one you think is the least
- Even if you think that two offenses/crimes are both very serious, you must choose just one as the most serious
- In the same way, you must choose just one as the least serious

A typical choice task would be:

- A person, using force, robs a victim of $10. The victim struggles and is shot to death
- A person steals, with no violence, property worth $100 from a household
- A father kills a drug addict son who had threatened to kill the entire family
- A person, using force, robs a victim of $10. No physical harm occurs

We should note that this choice task is really much more demanding than the usual "brand/pack/price" CBC choice task. Some respondents could think, for example, the third statement describes a very serious crime and some could think that it is not a crime at all. This is exactly what happened in our sample...

In our study, we had 200 cases in the sample. Fieldwork was conducted in the third quarter of 2003 in Rio de Janeiro, Brazil, and the sample was spread across all socio-economic classes: male/female, 18-60 years old. The scores were calculated using Hierarchical Bayes estimation, and are given in the following tables using Cohen's scale:

- Calculate the 'share of preference' for each sentence (similar to conventional conjoint)
- Assign an index of 100 for the average:
  - 100% of share of preference / 25 sentences = 4% associated with index 100
- So, if a sentence has an index of 586, it has a share of preference that is 5.86 times greater than the average

The average scores for the entire sample can be grouped into four categories:

- *Serious Crimes*: bombs, for example
- *Institutional Crimes*: stealing public money, for example
- *Quotidian Crimes*: stealing, without violence, $100 from a household, for example
- *Smaller Offenses*: stealing the best friend's husband, for example

The average scores are given in the tables below:

Table 1 – Serious Crimes

Average	*Serious Crimes*
586	A person plants a bomb in a public building. The bomb explodes and 20 people are killed
441	A father kills the wife and two children and commits suicide
312	A person, using force, robs a victim of $10. The victim struggles and is shot to death
285	A group of politicians and judges make a deal with drug dealers
246	A father kills a drug addict son who threatened to kill the entire family
161	A man forcibly rapes a woman. Her physical injuries require hospitalization
124	Street vendors sell drugs to teenagers in front of a school
107	A parent beats his young child with his fists. The child requires hospitalization

The bomb is 5.5 times more serious than spanking the son...

Table 2 – Institutional Crimes

Average	*Institutional Crimes*
71	A group of politicians steal $100 millions of public money and deposit it in fiscal paradise banks
61	A factory knowingly gets rid of its wastes in a way that pollutes the water supply of a city. As a result 20 people become ill but none requires medical treatment
42	A group steals money destined to medical assistance for a community that lacks medical care

Note the smaller range than in Serious Crimes.

Table 3 – Quotidian Crimes

Average	*Quotidian Crimes*
17	A person, using force, robs a victim of $10. The victim gets hurt and requires hospitalization
10	A person, without violence, steals property worth $10,000 from a household
7	A person, using force, robs a victim of $10. No physical harm occurs
7	A person, without violence, steals property worth $1,000 from a household
5	A person, without violence, steals property worth $100 from a household

Table 4 – Smaller Offenses

Average	*Smaller Offenses*
3	A person steals a car and later abandons it
3	A man steals a close friend's girlfriend
2	A street vendor sells stolen or false goods in the street
2	A man steals a close friend's wife
2	A woman steals a close friend's husband
2	A woman steals a close friend's boyfriend
2	A person, without violence, steals property worth $10 from a household
1	A person steals a bicycle parked on the street
1	A person disturbs the neighborhood with loud, noisy behavior

There are no significant differences among the average scores for economic class, gender, age group, or education level.

Power Law – Psychophysics

Remembering the problem:

- Stealing $10
- Stealing $100
- Stealing $1,000
- Stealing $10,000,

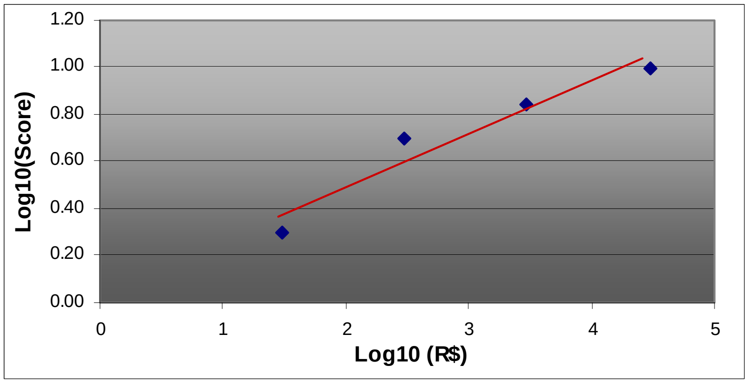
if the Power Law

$$Response = Score = \alpha \, Stimulus^{\beta}$$

holds, the Log relationship would be linear:

$$\text{Log (Score)} = \text{Log } (\alpha) + \beta \text{ Log (Stimulus).}$$

The graph below shows the relationships for the quantities ($10 / $100 / $1,000 / $10,000). Logs are taken for the quantities in reais (R$, the Brazilian currency):



Graph 2: Log-Log for Stimuli and Response

Sellin (1964) reported for a similar scale a beta coefficient of 0.17. Figlio (1978) reports a value of 0.24. Based on several studies, Lodge (1981) estimates this coefficient should be around 0.25. In our case, we obtained a coefficient of 0.22. That is, the scales are absolutely consistent: 25 years later, in a completely different socio-economical environment, the coefficients agree...

Visions of Reality

We saw that the scores did not vary significantly among the different economic classes, age group, gender etc. However it would be interesting to see if these scores would vary depending on the way people envisage reality.

Since we are dealing here with crimes, it would also be interesting to see if the degree of seriousness would be different depending on the way people view the possibility of reforming criminals, and on ethics in general.

To verify this, we segmented our sample, also using MaxDiffs, using twelve sentences describing attitudes towards:

- Possibility of reforming criminals
- Ethics/honesty
- Psychoanalytic Visions of Reality

The Visions of Reality were based on the work of Schafer (1976). This American psychoanalyst has derived, for therapeutic work, four visions representing different ways people could see reality:

- Comic
- Romantic
- Tragic
- Ironic

These visions will be detailed later. Let's begin with the reformation of criminals. We have selected two sentences describing opposite opinions about the matter:

Sentence	***Description***
1	*An outlaw can't be reformed: the only solution is death sentence*
2	*Penitentiaries should aim the reform of prisoners to reintegrate them into society*

The ethical point of view could be described as:

Sentence	***Description***
3	*Doing something slightly dishonest is not a problem: in one way or another everyone takes advantage of something*
4	*The most important thing for me is to be ethical and correct in everything I do*

We will describe briefly our understanding of Schafer's visions. The sentences we use to describe them were created by us, for segmentation purposes, and do not necessarily represent Schafer's point of view on the matter.

For each vision of reality, we created two sentences as descriptors. We are going to describe briefly how we see each one of the Shafer's visions, together with the corresponding sentences we created.

Comic Vision

- The comic vision seeks evidence to support unqualified hopefulness regarding personal situations in the world
- This serves to affirm that no dilemma is too great to be resolved
- This vision celebrates "the power of positive thinking"
- It is essentially pragmatic, aiming social cohesion (if not conformism)
- Their program can be summarized as "reform and progress"

To describe this vision, we have selected the sentences:

Sentence	Description
5	*No problem is too big that it cannot be solved*
6	*The world is in progress, something is always changing for the better*

Romantic Vision

- For this vision life is a quest or a series of quests: a perilous, heroic, individualistic journey that ends, after crucial struggles, with exaltation
- The standard cowboy would be, following Schafer, the commonplace American expression of this vision
- Implicitly, if not explicitly, it is regressive and childlike, particularly in its nostalgia for a golden age in time and space that is the final destination of the quest
- Here self-expression is equated with triumph

So we selected the sentences:

Sentence	Description
7	*I follow only my own values, this is the only way to succeed*
8	*There are no more those people capable of heroic acts that change the world*

Tragic Vision

- The person with this vision knows that renunciations are associated with gratification
- Recognizes the necessity to act in ignorance and bear the fear and guilt of action
- Is by far the most remorselessly searching vision, deeply involved, and along with the Ironic Vision, it has an impartial perspective of human affairs

We selected the sentences:

Sentence	Description
9	*A lot of times, in order to get what we want, we need to give up something we like*
10	*If needed I make a decision even if I am not absolutely sure of what should be done*

Ironic Vision

- This vision is ready to seek internal contradictions, ambiguities, and paradoxes (overlapping the Tragic Vision)
- Differently from the Tragic Vision, this vision aims at detachment
- A person with this vision takes nothing too seriously and defies not only well established traditions but also firmly held beliefs...

The sentences we selected were:

Sentence	Description
11	*If necessary I go against dogmas and well established traditions*
12	*For your own good, you can't take everything too seriously*

Schafer (1976) emphasizes the strong overlap between the Tragic and Ironic Visions, commenting that some authors often use the phrase “tragic irony” for the blending of these two visions.

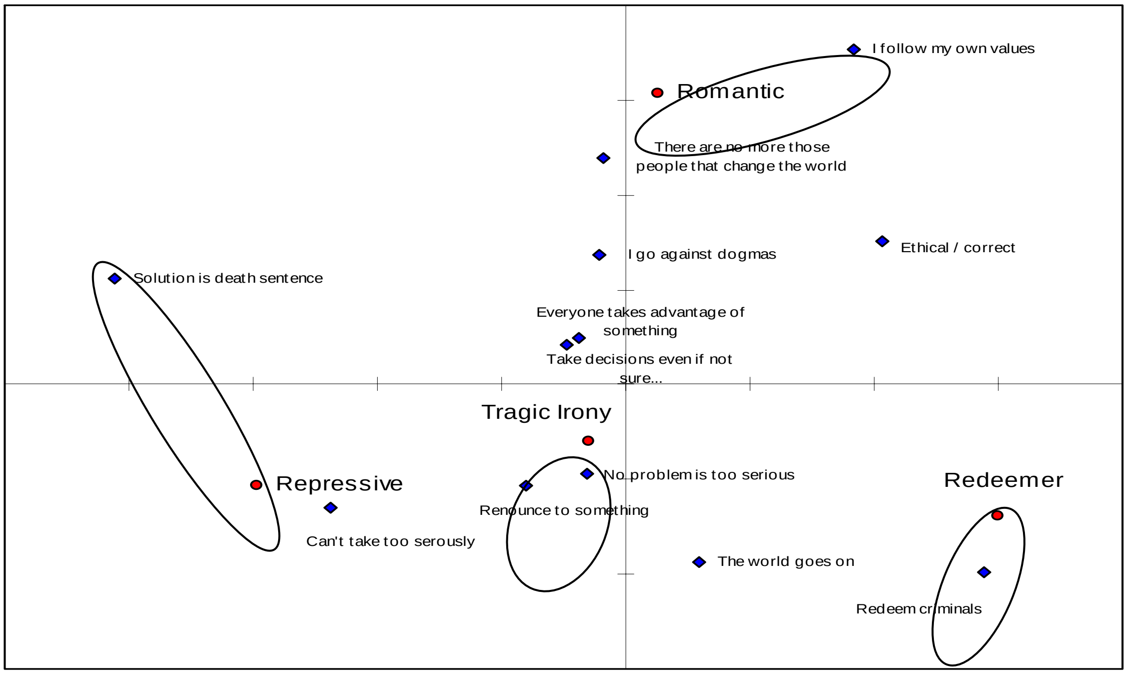
Those twelve sentences composed the set of sentences presented to respondents in choice tasks of four sentences each. In each choice task, the respondent was asked to choose the sentence with which he/she agreed most and the sentence with which he/she agreed least.

Using Latent Class analysis, we segmented the sample into four groups (percentages show the incidence of the segment in total sample):

Table 5 – Four-group Segmentation

Group	%	Typical sentence
Romantic	32	*I follow only my own values, this is the only way to succeed*
Repressive	22	An outlaw can’t be reformed: the only solution is death sentence
Tragic Irony	18	A lot of times, in order to get what we want, we need to give up something we like
Redeemer	28	Penitentiaries should aim the reform of prisoners to reintegrate them into society

The correspondence analysis mapping below shows the relationship among sentences and segments:



Graph 3: Correspondence Analysis Mapping – Attributes and Segments

For the four, five, and six group solution we analyzed, we were never able to split the Tragic Irony group, showing how overlapped these two visions really are.

If we were not able to find differences for age group, class etc. we can see differences in the table below among segments in the way they evaluate some offenses:

Table 6 – Differences among Segments

	Romantic	***Repressive***	***Tragic Irony***	***Redeemer***
Robbery of $10 with death	266	467	270	272
Selling drugs in front of school	98	233	102	82
Politicians stealing $100 M	95	39	60	76
Steal of $$ for medical care of unassisted population	54	18	39	49
Stealing $10,000 from household	10	6	14	9
Stealing $1,000 from household	6	4	10	6
Stealing and leaving a car	2	1	8	3
Stealing $10 from household	1	0	5	1
Stealing the boyfriend	1	0	5	1
Street vendor selling stolen / false goods	2	2	3	2

Respondents with a predominant Romantic Vision do care more about what we called Institutional Crimes, while those with a Repressive point of view care more about Serious Crimes. Tragic Irony respondents are more worried with less serious crimes/offenses.

HOW RELIABLE ARE THE RESULTS?

In our study, we had a very small number of choice tasks:

- 10 choice tasks for 25 attributes in the scaling step
- 8 choice tasks in the segmentation

We had 5 versions of the questionnaire and we calculated utilities for the degree of seriousness using Hierarchical Bayes estimation.

So the first question that arises is how reliable are the results. From the tables, we see that the results were consistent among themselves. Besides, the β coefficient in the Power Law is consistent with expected values in the literature. Finally, the segmentation was able to find the overlapped character of the Tragic and Ironic Visions. We had quite a few interesting external validations for our results.

CONCLUSION

Ideally, we should have connectivity of items within the design for each respondent, but this may lead to more choice tasks than we can afford in the questionnaire. In our case, we have intensively pre-tested the questionnaire, and arrived at the conclusion that our number of choice tasks was the maximum that could reasonably be allowed for the interview.

MaxDiffs has already been proved to be a very powerful scaling method, but the number of choice tasks may be an issue: we usually can't afford to have all the choice tasks we may need per respondent. Some possible ways to overcome this difficulty are to solve the problem with aggregate analysis (Latent Class instead of Hierarchical Bayes) or by specifyingusing HB a more appropriated user defined prior covariance matrix (Sawtooth Software, 2004).

Anyway, we need more work that could (or not...) validate results obtained with fewer choice tasks like the study presented here. The author firmly believes that if we want MaxDiffs to be a general purpose real world sized problems technique, we need to be able to have fewer choice tasks for a problem with, say, 30 to 50 attributes.

ACKNOWLEDGEMENTS

The author would like to thank the lawyer, Ana Lucia Lucas, for her help with the development of the offense statements and the psychoanalyst Renato Baraúna for help with Schafer's Visions of Reality. The author also thanks Bryan Orme, not only for the helpful comments on MaxDiffs, but also for help with the English text: English can really be a tricky language for a non-native speaker...

REFERENCES

Aaker, D., Kumar, V. and Day,G. (1988). *Marketing Research*. John Wiley & Sons.

Chrzan, C. (2004). *The Options Pricing Model.* Sawtooth Software Conference 2004 Proceedings

Cohen, S., and Markowitz, P. (2002). *Renewing Market Segmentation: Some New Tools to Correct Old Problems*. ESOMAR 2002 Congress Proceedings, 595 – 612

Cohen, S., and Neira, L., (2003). *Measuring Preference for Product Benefits Across Countries: Overcoming scale usage bias with Maximum Difference Scaling*. ESOMAR 2003 Latin American Congress Proceedings, 333 – 352

Cohen, S. (2003). *Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation. Sawtooth Software Conference 2003 Proceedings*, 61-74

Figlio, R. (1978). *The National Survey of Crime Severity: Result of Pretest*. Monograph, Department of Criminology, University of Pennsylvania

Lodge, M.(1981). *Magnitude Scaling-Quantitative Measurement of Opinions*. Sage Publications

Louviere, J. (1992). *Maximum Difference Conjoint: Theory, Methods and Cross-task Comparisons with Ratings-based and Yes/No Full Profile Conjoint*. Non-published paper. Department of Marketing, Eccles School of Business, University of Utah, Salt Lake City

Louviere, J., Swait, J., and Andreson, D. (1995). *Best-Worst Conjoint: a New Preference Elicitation Method to Simultaneously Identify Overall Importance and Attribute Level Partworth*. Working Paper. University of Florida, Gainsville, FL

McIntosh,E. and Louviere, J.(2002). *Separating Weight and Scale Value: an Exploration of Best-attribute Scaling in Health Economics*. Paper presented at the Health Economics Study Group, Odense, Denmark

Rossi, P., Gilula, Z. and Allenby, G. (2001). *Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach*. Journal of the American Statistical Association, 96, 20 –31

Sá Lucas, L. (2004). *Medindo com Eficiência a Opinião Social* , in Portuguese. ABEP 2004 Brazilian Congress of Marketing Research Proceedings , São Paulo, Brazil

Sawtooth Software (2004). *The CBC/HB System for Hierarchical Bayes Estimation – Version 3.2 Manual*. Sawtooth Software. http://www.sawtoothsoftware.com

Schafer, R. (1976). *A New Language for Psychoanalysis* . New Haven, Yale University Press

Sellin, J. and Wolfgang, M. (1964). *The Measurement of Delinquency*. New York: John Wiley

MULTICOLLINEARITY IN CSAT STUDIES

JANE TANG, JAY WEINER
IPSOS-INSIGHT

ABSTRACT

This paper examines the strengths and weaknesses of the commonly used tools for modeling customer satisfaction data. Most CSAT studies are plagued with multicollinearity. Compound that with the fact that most CSAT studies are tracking studies, you have a significant challenge on how to model the data and deliver stable, actionable results to your clients. We need be sure that differences in results from one wave to the next are true differences in the market and not just the result of a small number of respondents checking 8 instead of 7 on the next wave of the questionnaire.

This paper compares several traditional CSAT modeling techniques (Logistic regression, Partial Least Squares, Ordinary Least squares, Shapley value regression and Kruskal's Relative Importance). First, a simulated data set is used to show the relative impact of multicollinearity on the results. In addition, bootstrap samples are pulled from this data set and used to show the relative stability of the various techniques. An actual case study will be used to show how the various methods perform with a real data set.

INTRODUCTION

In Customer Satisfaction (CSAT) studies, we often conduct driver analysis to understand the impact of explanatory variables on the overall dependent variable. That is, we need to provide the client with a list of priority items that can be improved and have a positive impact on overall satisfaction or customer loyalty and retention. Typically, the goal is to establish the list of priorities and relative importance of the explanatory variables, rather than trying to predict the mean value of customer satisfaction if these improvements are implemented. Since most CSAT studies are tracking studies, the results can be monitored over time to determine if the desired changes are occurring.

We must be sure that changes in the results are in fact consumer response to the marketing efforts and not just phantoms of the analytic tool used to build the model. Multicollinearity is a serious problem in many CSAT studies and presents two challenges in modeling CSAT data. The first problem is how to accurately reflect the impact of several independent variables that are highly correlated. The second challenge is how to insure that the results are consistent wave to wave when you are trying to track a market over time. This paper will illustrate the problems that multicollinearity presents in modeling data and then compare various modeling tools.

THE ISSUE OF MULTICOLLINEARITY

Perhaps the best method for dealing with multicollinearity is to design questionnaires that are well designed and tested to avoid this issue. For most researchers, this is a common desire, but difficult to achieve. In most CSAT studies, we measure a variety of attributes that are often highly correlated with each other. For example, in evaluating the service in a customer call center, we frequently ask respondents to rate satisfaction with the friendliness of the operator, and also of the operator's ability to handle the problem the first time. We often see that these two attributes are highly correlated with each other. This may be due to halo effects in that most customers that are happy with the resolution of the problem will reflect back and state that the operator was friendly. Regardless of the reason for the correlation between these two attributes, we need to find a modeling tool that is capable of determining the relative contribution of each of these attributes to overall satisfaction.

We begin our discussion by creating a data set with 5,000 observations that is typical of CSAT studies, where the properties of the dependent and independent measures are known. The simulated data set has two pairs (four) of independent variables and a dependent measure (overall satisfaction). The first pair (q1 and rq1) of independent measures is constructed to be almost perfectly correlated with each other and highly correlated with the dependent variable. The second pair (q2 and rq2) is also highly correlated with each other, but less correlated with the dependent measure. All variables are on a 10-point rating scale. The correlation matrix is shown.

	os	q1	rq1	q2	rq2
os	1.00	0.63	0.62	0.39	0.38
q1	0.63	1.00	0.98	0.26	0.26
rq1	0.62	0.98	1.00	0.25	0.25
q2	0.39	0.26	0.25	1.00	0.98
rq2	0.38	0.26	0.25	0.98	1.00

A common modeling tool is ordinary least squares regression (OLS). If we regress OS on Q1 and Q2, we find the following results.

Variable	Estimate	P-value
q1	0.57	<.0001
q2	0.23	<.0001

Both variables are significant and have a positive impact on overall satisfaction. These results are consistent with the correlation matrix shown in the previous table. To demonstrate the impact of multicollinearity on OLS, we will re-run the model adding rq1 and rq2.

Variable	Estimate	P-value
q1	0.51	<.0001
rq1	0.06	0.19
q2	0.23	<.0001
rq2	0.00	0.97

Because q1 and rq1 are almost perfectly correlated, the client will wonder why one of the two variables has more than eight times the impact on overall satisfaction while the other is not significant. If the model were run using only rq1 and rq2, we would see results very similar to the model with q1 and q2 above.

The second challenge multicollinearity presents in CSAT studies is in analyzing the results in subsequent periods of time. Most CSAT studies are in fact, tracking studies. We execute the first wave to understand where we are, then we track our key measures over time to understand how our efforts to improve overall satisfaction are working. To simulate the impact of multicollinearity on the ability to model subsequent waves of data we create a second data set. This data set is created using the same rules used to create the first data set. The correlation matrix looks like this.

	os	q1	rq1	q2	rq2
os	1.00	0.64	0.63	0.37	0.38
q1	0.64	1.00	0.98	0.25	0.25
rq1	0.63	0.98	1.00	0.24	0.24
q2	0.37	0.25	0.24	1.00	0.98
rq2	0.38	0.25	0.24	0.98	1.00

In this wave of data, rq2 is slightly more correlated with overall satisfaction than q2. When we run the OLS, we see the following results.

Variable	Estimate	Pr > |t|
q1	0.58	<.0001
rq1	0.01	0.78
q2	0.01	0.89
rq2	0.21	<.0001

If these were the results of an actual tracking study, we would have advised the client to focus on q2 in wave 1, but in this wave, we would encourage the client to focus on rq2. It would be nice to know that the change of focus is due to actual events that have happened in the market and not due to a small change in the overall correlation matrix. Clearly OLS may not be the most appropriate tool to use in modeling CSAT data.

OTHER TOOLS IN THE TOOLBOX:

There is a family of approaches that considers all possible combinations of explanatory variables. These include: Kruskal's Relative Importance (KRI), Shapley Value (SV) Regression and Penalty & Reward Analysis. Unlike traditional regression tools, these techniques are not used for forecasting. In OLS, we predict the change in overall satisfaction for any given change in the independent variables. The tools addressed in this section are used to determine how much better the model is if we include any specific independent variable versus models that do not include that measure. The conclusions we draw from these models refer to the usefulness of including any measure in the model and not its specific impact on improving measures like overall satisfaction.

For each of these modeling techniques, the impact of a particular variable on the dependent variable is determined by its contribution to the model, measured by the difference in "XXX" between including it in the model and not including in the model. The "XXX" varies by technique.

For Kruskal's Relative Importance, OLS regression is used for all possible combinations of explanatory variables. The contribution of each attribute is measured by squared partial correlations. KRI can be run using SAS with PROC REG with selection=rsquare & PROC IML.

$$KRI_j = \sum_k \sum_i \frac{k!(n-k-1)!}{n!}\left[(SSE_{i|j} - SSE_{i|j(-j)})/SSE_{i|j}\right]$$

where

$SSE_{i|j}$ is the SSE of a model i containing predictor j

$SSE_{i|j(-j)}$ is the R^2 of a model i without j

Including q1 and q2 in the model, we see that the results are very consistent with the results obtained using OLS.

	Data 1	Data 2
q1	0.37	0.39
q2	0.12	0.11

When we add the highly correlated independent measures to the model, KRI essentially splits the importance for each pair of independent measures. Both q1 and rq1 are equally important and more than three times more important than q2 and rq2. This result often has more face validity with clients than the results obtained using OLS.

	Data 1	Data 2
q1	0.20	0.21
rq1	0.18	0.19
q2	0.06	0.06
rq2	0.06	0.06

Like KRI, Shapley Value Regression uses OLS regression for all possible combinations of explanatory variables. The contribution of each attribute is measured by the improvement in R-square. This can be executed with SAS using PROC REG with selection=rsquare & PROC IML.

$$SV_j = \sum_k \sum_i \frac{k!(n-k-1)!}{n!}\left[\upsilon(M_{i|j}) - \upsilon(M_{i|j(-j)})\right]$$

where

$\upsilon(M_{i|j})$ is the R^2 of a model i containing predictor j

$\upsilon(M_{i|j(-j)})$ is the R^2 of a model i without j

Again, a model with just q1 and q2 is similar to the results obtained using OLS. The results when we add the highly correlated variables to the model are almost identical to those seen with KRI.

	Data 1	Data 2
q1	0.35	0.36
q2	0.10	0.10

	Data 1	Data 2
q1	0.18	0.19
rq1	0.17	0.17
q2	0.05	0.05
rq2	0.05	0.05

Penalty & Reward Analysis is unique in that instead of a single linear model, we build two models. The overall sample is divided into the delighted group (e.g. top boxes in overall satisfaction) and the less than satisfied group (e.g. low boxes in overall satisfaction). Dissatisfaction with explanatory variables is a potential source of penalty or barrier to satisfaction. Delight with explanatory variables is a potential source of reward or driver of delight. This approach is useful when you do not have a suitable normal distribution of the overall dependent variable, and suspect the drivers of satisfaction may be different from the drivers of dissatisfaction. The unique contribution of each attribute is measured by the difference in overall total unduplicated reach by each combination of the explanatory variables between the delighted group and the less than satisfied group. The importance of a variable is measured as the sum of its penalty and reward. This analysis is done using SAS IML.

The results for the penalty reward analysis are very similar to those obtained with KRI and SV regression.

	Data 1	Data 2
q1	0.70	0.70
q2	0.44	0.43

	Data 1	Data 2
q1	0.36	0.36
rq1	0.36	0.35
q2	0.23	0.22
rq2	0.22	0.21

Tools using all possible combinations of attributes seem to do a better job of dealing with issues of multicollinearity. OLS tends to understate the impact of highly correlated attributes and can in fact suggest a negative coefficient in some models for attributes that are expected to have a positive impact on the dependent measure. Conversely, KRI, SV regression and penalty/reward analysis assign equal weights to attributes that are highly correlated with each other. When applied to tracking data, these tools are very stable in predicting the impact of attributes between waves. We have more confidence that changes in the impact of independent measures are in fact true differences in the market and not due to small changes in the correlations with the dependent measure.

EXAMPLES FROM A REAL DATA SET:

The dataset used is from an IPSOS Loyalty Optimizer Study. The dependent measure for the models presented is the security index. This is a composite measure of customer loyalty and is a continuous variable. For some models (Penalty/Reward and logistic regression, the data can be divided into the high vs. low loyalty groups. Each respondent only evaluates one company (single observation). There are five constructs each with 2 attributes: Relationship, (Feel like a friend, Trust and treat fairly), Experience (Quality, Satisfaction), Offer (Relevance / Differentiation), Brand (Familiarity / Popularity), Price (Willingness to pay / Comparison to competitors). All explanatory variables are expected to have a positive impact on the dependent variable.

Objectives:

The goal is to assess the stability of the driver analysis from repeated samples for different sample sizes and to compare results in Regressions vs. KRI, SV regression and P/R analysis.

Methodology:

We create several datasets by pulling repeated bootstrap samples from the original loyalty optimizer data – e.g. 100 bootstrap samples (i.e. sampling with replacement) for each sample sizes of: N=2,407 (original sample size), N=1,000, N=500, N=300 and N=100. Using the Security Index variable as the dependent variable we run OLS regression, STEPWISE OLS, Partial Least Square Regression (PLS), KRI, SV regression and compare the results. For OLS, all parameters are entered into the model. In the stepwise method of OLS, parameters that are not significant in the model are assumed to be zero. For Logistic regression and P&R analysis we compare results using High vs. Low loyalty category as the dependent measure.

Criterion for Comparison:

First, we examine the gaps in the importance calculation for each attribute between the first 2 bootstrap samples in each sample size. Note, as all attributes are assumed to have a positive impact, negative betas are set to “0”. The resulting betas are re-proportioned, such that the sum of importance=100%. The following table shows the results of the OLS run on the “full” sample size of 2,407 respondents. The average gap in importance between wave 1 and wave 2 is 4.0%, the largest difference is at 8.7%

n=2,407		**Sample 1**			**Sample 2**		
OLS Regression	**Beta**	**"zeroed"**	**Importance**	**Beta**	**"zeroed"**	**Importance**	**Gap**
Feels like a friend	0.00	0.00	0%	0.00	0.00	2%	2%
Trust and treat fairly	0.01	0.01	10%	0.01	0.01	6%	4%
Quality	0.00	0.00	0%	0.01	0.01	5%	5%
Satisfaction	0.04	0.04	34%	0.03	0.03	31%	4%
Relevance	0.03	0.03	33%	0.03	0.03	24%	9%
Differentiation	0.00	0.00	3%	0.00	0.00	1%	1%
Familiarity	0.01	0.01	5%	0.01	0.01	13%	8%
Popularity	0.00	0.00	0%	-0.01	0.00	0%	0%
Willing to pay	0.01	0.01	12%	0.01	0.01	9%	3%
Comparison to competitor's price	0.00	0.00	3%	0.01	0.01	7%	4%
Average Gap between Sample 1 & 2							**4.0%**
Maximum Gap between Sample 1 & 2							**8.7%**

The second criterion for comparison is to calculate Coefficient of Variation by dividing the standard deviation by the mean across 100 bootstrap samples.

n=2,407 **OLS Regression**	Mean	STD DEV	C of V
Feels like a friend	0.004	0.004	1.094
Trust and treat fairly	0.004	0.004	1.068
Quality	-0.001	0.005	4.127
Satisfaction	0.038	0.005	0.136
Relevance	0.031	0.004	0.140
Differentiation	0.005	0.003	0.699
Familiarity	0.015	0.004	0.275
Popularity	-0.007	0.003	0.484
Willingness to pay	0.015	0.002	0.148
Comparison to competitor's price	0.003	0.003	1.278
		Median C of V.	**0.592**

The following table shows the results of all methods.

	Dependent variable: Security Index					Dependent variable: High vs. Low loyalty		
Sample size	**OLS**	**OLS (stepwise)**	**PLS**	**KRI**	**SVReg**	**Logistic Reg**	**Logistic Reg (stepwise)**	**P&R**
Average Gap Between Sample 1 & 2								
2,407	4.0%	5.5%	3.8%	0.7%	0.6%	3.9%	6.1%	0.6%
1,000	2.7%	2.2%	2.4%	1.5%	1.4%	3.0%	4.2%	0.8%
500	5.6%	6.4%	5.3%	3.2%	2.8%	4.9%	11.5%	1.6%
300	6.7%	8.0%	7.6%	3.8%	2.9%	5.4%	6.7%	2.3%
100	9.9%	20.0%	9.5%	3.6%	3.4%	-	-	-
Maximum Gap Between Sample 1 & 2								
2,407	8.7%	13.8%	8.0%	2.3%	2.1%	10.2%	16.4%	1.5%
1,000	6.4%	7.7%	5.6%	3.5%	3.3%	7.1%	13.5%	1.3%
500	13.6%	13.6%	12.2%	6.8%	5.3%	13.1%	26.8%	3.4%
300	30.5%	33.8%	32.5%	11.1%	7.3%	11.3%	33.5%	6.7%
100	21.1%	44.9%	18.7%	10.2%	9.6%	-	-	-
Median Coefficient of Variation								
2,407	59%	97%	57%	13%	12%	73%	193%	7%
1,000	75%	115%	73%	19%	17%	60%	143%	11%
500	67%	97%	64%	22%	20%	110%	236%	13%
300	112%	140%	111%	32%	33%	184%	341%	17%
100	172%	168%	166%	38%	39%	-	-	-

CONCLUSIONS:

With large sample sizes, all the methods perform fairly well. If the objective is to establish relative importance, rather than forecasting, methods that take into consideration all the possible combinations of the explanatory variables (i.e. KRI/ SV regression/ P&R) are worthwhile since they have much smaller sample-to-sample variability. These methods also do not collapse as quickly as sample size decreases. This is more suitable where we don't have large sample.

REFERENCES:

Lipovetsky, S., and Conklin, M. (2001) "Analysis of Regression in Game Theory Approach", *Applied Stochastic Models in Business and Industry*, 17, 319-330

Kruskal, W. (1987). "Relative Importance by Averaging over Orderings", *The American Statistician, 41(1)*

CONJOINT ANALYSIS CASE STUDIES

MODELING CONCEPTUALLY COMPLEX SERVICES: THE APPLICATION OF DISCRETE CHOICE CONJOINT, LATENT CLASS, AND HIERARCHICAL BAYES ANALYSES TO MEDICAL SCHOOL CURRICULUM REDESIGN

[1]*CHARLES E. CUNNINGHAM, KEN DEAL,*
ALAN NEVILLE, AND HEATHER MILLER
MCMASTER UNIVERSITY

Since its introduction at McMaster University in the late 1960s (Neufeld, Woodward, & MacLeod, 1989), small group, problem-based learning has been adopted widely as an alternative to more traditional approaches to medical education (Antepohl, Domeij, Forsberg, Ludvigisson, 2003; Kaufman & Mann, 1998; Mennin, Kalishman, Friedman, Pathak, & Snyder, 1996; Khoo, 2003; Peters, Greenberger-Rosovsky, Crowder, Block, & Moore, 2000; Schmidt & Molen, 2001). Problem-based medical education has several components. Educational activities are conducted in small tutorial groups rather than large lecture formats. Students acquire content by solving a strategically designed series of health care problems rather than simply listening to lectures. Students adopt an active, self-directed approach to the learning process, establish learning objectives, collaborate in the solution of health care problems, and debate alternative approaches to complex issues. As a part of the tutorial process, students support balanced participation by all members, resolve conflicts, evaluate their progress, and provide feedback on the learning process. Faculty tutors facilitate the tutorial group learning process by helping students set learning objectives, encouraging participation, posing questions, and providing evaluative feedback (Maudsley, 1999).

Although advocates contend that problem-based learning enhances the intrinsic motivation to learn, activates prior knowledge, improves recall, helps students apply basic science to the solution of clinical problems, improves communication and problem solving skills, and contributes to life-long learning (Norman & Schmidt, 1992; 2000; Schmidt, 1993), efforts to study the effectiveness of problem-based medical education have yielded mixed results (Colliver, 2000; Farrow & Norman, 2003; Newman & The Pilot Review Group, 2003; Norman & Schmidt, 2000).

In 2002, McMaster University's Faculty of Health Sciences initiated a process to redesign its problem-based approach to undergraduate medical education. Student-faculty collaboration is a central tenant in problem-based educational models. Students enter medical school with different educational backgrounds, career objectives, and learning preferences. Their educational experiences represent a source of tacit knowledge that could make a valuable contribution to the design of health science

T1 Author Notes: Charles E. Cunningham, Department of Psychiatry and Behavioural Neurosciences, McMaster University. This project was supported by the Jack Laidlaw Chair in Patient-Centred Health at McMaster University Faculty of Health Sciences. We acknowledge the MBA Strategic Marketing Analysis team comprised of Dave Finlay, Jason, Rinne, Kevin Robb, Bob Clark, and Jean-Marie Fiala. Address correspondence to Charles E. Cunningham, Chedoke Campus, Hamilton Health Sciences, Hamilton, Ontario, Canada, L8N 3Z5. Electronic mail may be sent via Internet to Hcunningh@mcmaster.caHT.

educational programs. Moreover, the faculty who design medical education programs and the students who are the users of these educational services may have different preferences regarding key features of the program. In the absence of strong evidence regarding the relative effectiveness of different approaches to medical education (Colliver, 2000; Norman & Schmidt, 2000), it is reasonable to assume that students may respond better to programs that are consistent with their learning process preferences.

This study used a combination of qualitative and quantitative methods (Barbour, 1999) to involve students in the revision of McMaster University's undergraduate medical curriculum. Focus groups and key informant interviews were conducted to generate curriculum redesign themes. Next, we used a quantitative consumer preference modeling strategy, discrete choice conjoint analysis (Ryan et al., 2001: Ryan & Farrar, 2000; Ryan & Gerrard, 2003), to determine the relative importance of different educational attributes to medical students. In discrete choice conjoint analytic surveys, informants make a series of choices among product or service options comprised of several attributes. It has been suggested that choice tasks prompt a more thoughtful consideration of the tradeoffs associated with competing alternatives and contribute to a more in-depth understanding of underlying preferences (Phillips, Johnson, & Maddala, 2002). In comparison to simple rating scales, choice tasks limit the superficial processing of individual items, reduce halo effects or social desirability biases, and better reflect actual behavior (Phillips et al., 2002). The complexity of multi attribute choice tasks, and the simplifying heuristics they evoke (Payne, Bettman, & Johnson, 1993), mimics real world decision making (Ryan et al., 2001). Conjoint analysis is, therefore, considered to be a more powerful approach to the analysis of preferences than simple ranking or rating approaches (Phillips et al., 2002; Ryan et al., 2001).

Although conjoint analysis has been used extensively by marketing researchers to involve consumers in product and service design, it has, more recently, been applied to health care preference modeling (Ryan, 2000; Ryan et al., 2001). Conjoint analysis has been used to design health services (Morgan, Shackley, Pickin, & Brazier, 2000), set service priorities (Farrar, Ryan, Ross, & Ludbrook, 2000), study symptom impact (Osman et al., 2001), determine treatment preferences (Fraenkel, Bodardus, Wittink, & Wittink, 2001; Maas & Stalpers, 1992; Ratcliffe, Buxton, McGarry, Sheldon, & Chancellor, 2003; Ryan, 1999; Singh, Cuttler, Shin, Silvers, & Neuhauser, 1998), and explore health outcome choices (Ryan & Farrar, 2000; Stanek, Oates, McGhan, Denofrio, & Loh, 2000).

In this study, we applied conjoint analysis to several questions. First, what curriculum and learning process attributes of the undergraduate medical educational program are most important to students? Second, what is the relative value of different curriculum and learning process redesign options? Third, are there segments of students with different educational preferences? Finally, what combination of redesign options would best match the educational preferences of current medical students?

METHOD

Participants

A sample of 256 undergraduate medical students at McMaster University completed the on-line survey. This constituted approximately 56% of available students. Participation was reduced by a SARS protocol in Ontario Hospitals which prevented MD students from entering the Medical School and accessing computer terminals and by the number of students enrolled in out of town clerkship placements who did not have access to the University's Learn Link Internet service.

Qualitative Procedures

We composed a conjoint analytic survey according to a two stage process. First, an informative sample of students, faculty, and administrators participated in an electronic focus group using the Decision Support Lab in the DeGroote School of Business at McMaster University. The Decision Support Lab is equipped with 25 individual terminals linked to a central server using GroupSystems MeetingRoom software with two screens posing questions, displaying the entries of all participants, and summarizing data. This group was conducted by an expert large group facilitator according to an interview guide. After an introduction to the operation of the decision support lab, participants listed opportunities for improvement in McMaster University's undergraduate MD program, reviewed the full range of redesign options identified by the group, and voted (yes or no) whether each was appropriate for student input. We selected those dimensions of the program that 75% of the participants agreed were appropriate areas for student input. After reviewing the complete list of appropriate redesign options generated by the group, participants designated their top five choices by assigning a rank of 1 to 5 to the remaining opportunities. Next, participants entered potential solutions to the top rated opportunity for improvement and ranked these in order of preference. This process was repeated for the five top ranked problems. At the completion of this session, the full list of alternative solutions was reviewed. Similar opportunities were grouped and their cumulative ranks pooled. The alternative solutions to each opportunity for improvement were re-ranked based on these composite rankings. Table 1 shows the ranked list of opportunities for improvement.

Table 1
Thematic opportunities for improvement and associated ranks

Theme	Total Rank
Curriculum Design and Organization	39
Presentation of Fundamental Medical Concepts	38
Tutorial Group Leader Development	31
Evaluation of Student Progress	23
Teaching Communication Skills	15
Teaching Clinical Skills	13
Design of Clerkship Placements	10

Quantitative Procedures

Using the output of focus group sessions and themes emerging from key informant interviews we composed 14 educational attributes. Each attribute was operationalized by four levels: one level describing current educational practice and three describing redesign options.

Using SSI Web from Sawtooth Software, we composed a partial profile choice-based conjoint (CBC) survey. We selected the short-cut option that builds an experimental design using three criteria: (1) Minimal Overlap (each attribute level appears as few times as possible in a choice task. (2) Level Balance (each attribute level appears as close to an equal number of times as possible), and (3) Orthogonality (to ensure that each attribute level's utility can be estimated independently, attribute levels are chosen independently of each other).

Each participant completed 15 choice tasks. To reduce choice task complexity and increase observer efficiency (Patterson, 2003), each undergraduate program was described by three attribute levels. Each choice task presented 3 optional medical school undergraduate programs described by 3 attribute levels.

The undergraduate education office e-mailed a URL link to all undergraduate medical students. An optional draw of $100.00 was offered as an incentive.

RESULTS

We segmented data from the conjoint experiment using version 3 of Sawtooth Software's Latent Class Module (Sawtooth Software, 2004a). We assumed convergence when the Log-likelihood decreased by less than .01. We replicated the solution 5 times beginning at random starting points and accepted the solution with the best fit, $X^2 =$ 1989.12. A two segment solution converged in 36 iterations with 84.9% of market weight assigned to segment 1 and 15.1% assigned to segment 2. The average maximum probability of group membership was .96. Despite differences in the preferences of the two segments, there were no statistically significant differences in the demographics, educational background, or career goals of the two segments.

Next, we used Sawtooth Software's Hierarchical Bayes (CBC/HB) Version 3.1 to compute individual partworths for the students in each segment. See papers by Sawtooth Software, (2004b), Allenby, Arora, & Ginter (1995); and Lenk, DeSarbo, Green, & Young (1996) for a detailed discussion of Hierarchical Bayes analysis. Hierarchical Bayes estimates the vector of mean population betas, the matrix of population beta covariances, and the vector of betas for each respondent using two distributions: (1) a multivariate normal distribution and (2) a multinomial logit model based on actual responses. Hierarchical Bayes samples from these distributions using a Gibbs Sampling strategy. A Metropolis-Hastings Algorithm randomly "perturbs" estimates and repeats the estimation process iteratively (Sawtooth Software, 2004b).

For this analysis, we began with default CBC/HB "smart start" estimates of beta and alpha and computed 2000 burn in iterations before convergence was assumed. We accepted 1000 draws per respondent with a skip factor of 10 to increase the independence of draws and improve the precision of our estimates (Sawtooth Software, 2004b). A total of 12000 iterations were computed.

Using individual partworths from the Hierarchical Bayes analysis, we computed importance scores for each segment (Orme, 2002). Importance scores, which are standardized to sum to 100 across the 14 attributes, reflect the relative contribution of each educational attribute to student choices. To permit comparisons, we computed standardized zero-summed utility values with the average range within attributes equal to 100 (Orme, 2002). Utility values, which reflect the relative contribution of each attribute level to hospital choices, approach zero when preferences for different levels of an attribute vary among members of a segment. Higher utility values are observed when members of a segment show similar preferences for individual attribute levels.

Program Structure Preferences

The standardized zero-summed utility values presented in Figure 1 show that, although segment 1 preferred smaller tutorial groups, group size exerted less of an influence on the choices of segment 2. Segment 1 preferred a balance of two small group problem based learning tutorials and three large group lectures per week.

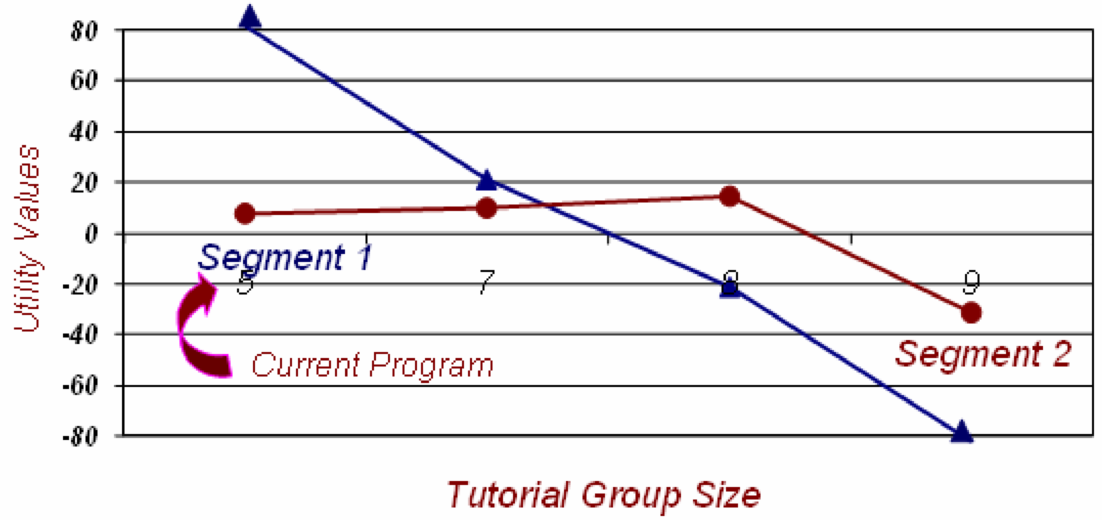


Figure 1. Utility values for tutorial group sizes of 5, 7, 8, and 9 students for segments 1 and 2.

Students in segment 1 were most attracted to the current 3 year program model without a streaming option. Students in segment 2, in contrast, favored a program with streaming based on career objectives and learning process preferences.

Education Process Preferences

The utility values in Figure 2 show that students in segment 1 were most attracted to McMaster's problems-based tutorial process; expert faculty tutors who facilitated tutorial group learning processes without teaching didactically. Segment 2's preferences seemed to be driven by an emphasis on content expertise with less concern regarding the role faculty played (optimizing process versus didactic teaching) in the teaching process.

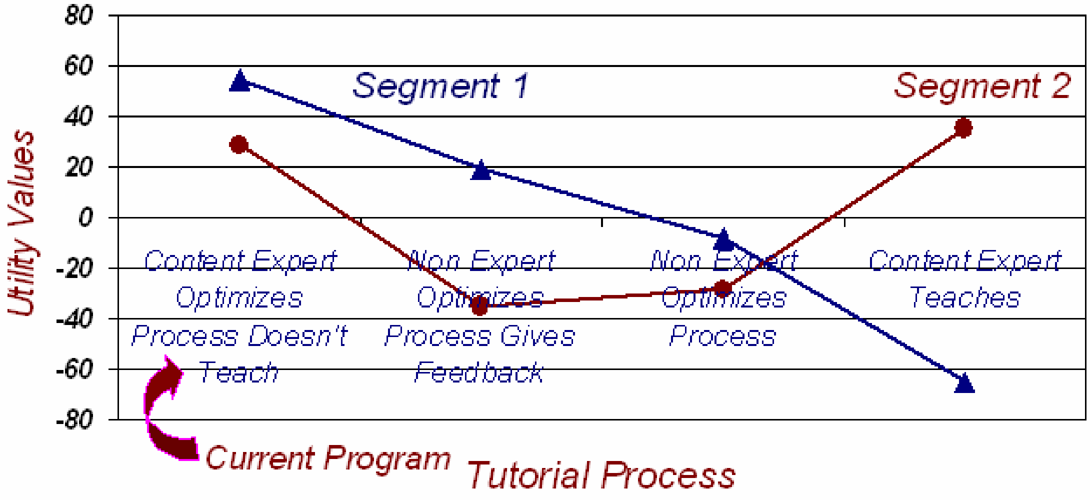


Figure 2. Utility values for tutorial group process for segments 1 and 2.

Curriculum Design

The utility values presented in Figure 3 show that students in segment 1 preferred health care problems focusing on core curriculum concepts while touching on diagnosis and treatment. This constituted a shift from the current MD program that tends to emphasize the process of diagnoses and treatment.

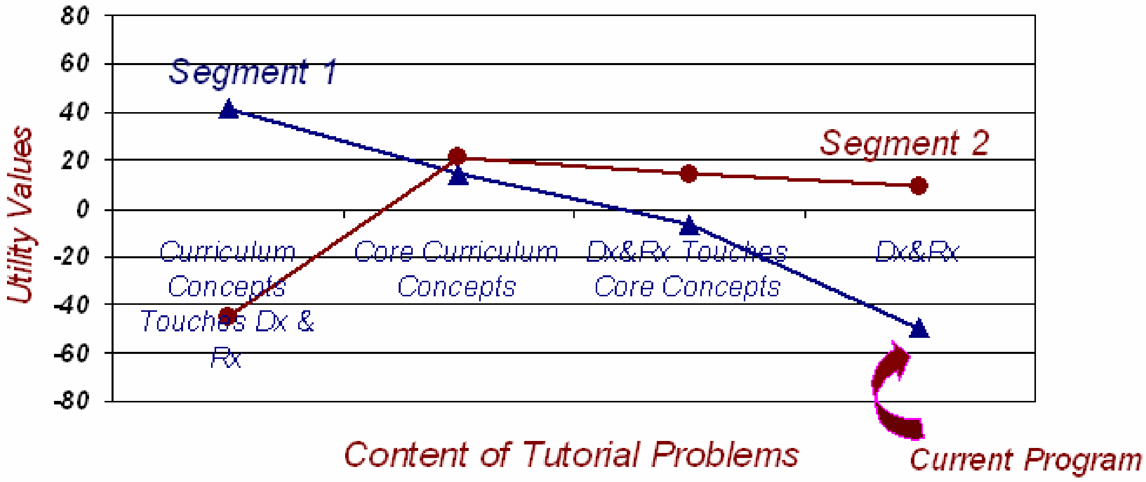


Figure 3. Utility values for the content of tutorial group problems for segments 1 and 2.

Students in segment 1 preferred an MD program in which different educational activities were more closely linked to the program's curriculum. For example, segment 1 preferred that the problems dealt with in clinical skill sessions, where students acquire history taking and physical examination skills, be more closely linked to the program's core curriculum concepts. Segment 1 also preferred that the patients they encountered while working in clerkship settings were supplemented by patients selected to illustrate key curriculum concepts.

Evaluation

Both segments preferred the current pass-fail approach rather than a graded evaluation. Segment 1 preferred an existing approach to evaluation which uses a series of interim Objective Structured Clinical Examinations (OSCEs) coupled with individual feedback and a pass-fail format. Interesting, segment 2, a group of students favoring a more traditional approach to medical education, chose an approach to evaluation requiring that clinical skills must be demonstrated before graduation.

Electronic Enhancements

In contrast to the written health care problems currently used in small group tutorial problem solving discussions, both segments 1 and 2 preferred a combination of computer simulated and live simulated heath care problems. In addition, both segments preferred that existing small group tutorials be enhanced by web-based processes.

Simulating a Redesigned Undergraduate MD Program

Medical schools in Ontario have faced an increase in MD program enrollments. Given the substantial costs of small tutorial groups lead by expert faculty tutors, we simulated student response to a program with larger tutorial group sizes. As a trade off, we retained attributes of the current program with high utility values and rejected program attributes with low utility values in favor of redesign options with higher utility values. Because a majority of students did not support the development of a streamed program with alternative curriculum options, we simulated a single program consistent with the educational preferences of a majority of students.

Utility values suggested several ways of improving the design of the undergraduate program. First, students preferred a shift to health care problems emphasizing core medical concepts rather than more practical problems focusing on diagnoses and treatment. Second, students preferred that the content of the health care problems solved in small group tutorial sessions, the content of clinical skills sessions, and the patients encountered in clerkship placements be linked to the undergraduate curriculum's learning objectives. Finally, students preferred an undergraduate curriculum based on more clearly stated learning objectives. The randomized first choice simulations presented in Figure 4 show the percentage of preference shares assigned to this more conceptually focused program. With 5 students per tutorial group in both the current and proposed conceptually focused program, 97% of the students would prefer the conceptually enhanced program. If the number of students per tutorial in a conceptually enhanced program was increased to 9, simulations predict that 62% of the students would still prefer the conceptually enhanced program. Even with 5 students per tutorial, only 38% would prefer the current program to the conceptually enhanced program.

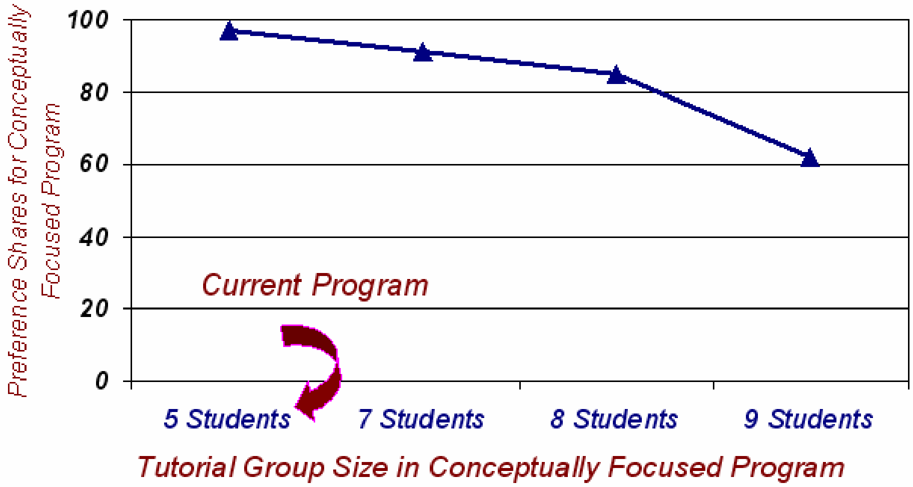
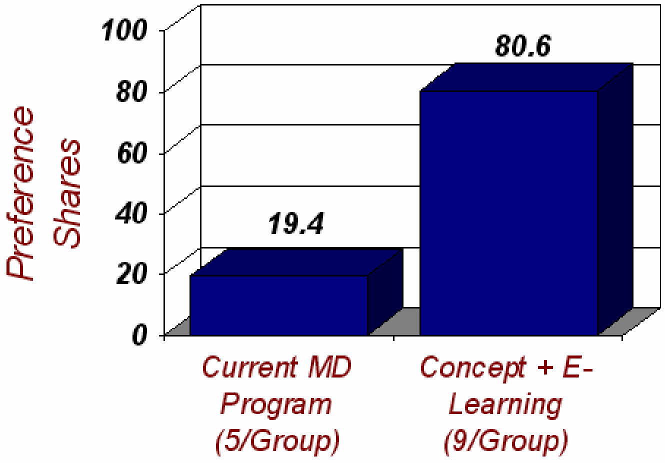


Figure 4. Simulation predicting preference shares for a conceptually enhanced MD curriculum at 5, 7, 8, and 9 students per tutorial group in comparison to the existing program with 5 students per group.

Next, we simulated a reinvestment of the savings associated with increasing tutorial group size in a series of electronic enhancements. These included an electronically enhanced tutorial process and computer simulated health care problems. The randomized first choice simulations summarized in Figure 5 predict that, even with 9 students per tutorial group, 81% would prefer an electronically and conceptually enhanced alternative to the current MD program with 5 students per tutorial group.



Electronic Enhancements

- *Web enhanced Tutorial Process*
- *50% Electronic 50% Live Simulated Tutorial Problems*

Concept Enhancements

- *Learning Objectives Explicit*
- *Tutorial problems emphasize core concepts + Dx & Rx*
- *Clinical skills linked to tutorial curriculum*
- *Curriculum-linked clerkship experiences*

Figure 5. Randomized first choice simulation predicting preference shares for current MD program at 5 students per tutorial group versus a conceptually and electronically enhanced program at 9 students per group.

DISCUSSION

This study is, to our knowledge, the first application of choice-based conjoint analysis to the design of an undergraduate medical education program.

Most students preferred a problem-based approach to medical education in small group tutorials supported by tutors with content expertise who facilitated group process but did not teach didactically. The importance of tutors with a combination of facilitative skills and content expertise is generally supported by the evidence in this area (Dolmans, Wolfhagen, Scherpbier, & van der Vleuten, 2001; Maudsley, 1999; Schmidt, 1993; Schmidt, 1994). The results of our conjoint analysis are consistent with the themes emerging from our focus groups; tutor development is an issue of considerable importance.

A majority of students preferred a closely related series of curriculum enhancements. Choices revealed a preference for tutorial problems emphasizing the curriculum's core concepts, while touching on diagnosis and treatment. This constituted a significant shift from current practice in which the design and discussion of tutorial problems tends to focus on diagnoses and treatment.

Students also preferred greater integration of learning experiences. This included a preference that the selection of tutorial problems, clinical skills training exercises, and the patients students encountered in clerkship settings be more closely linked to the program's core curriculum concepts.

Choice tasks allowed an analysis of tradeoffs in educational design. Although increasing tutorial group size would be associated with a significant cost savings, sensitivity analyses found this to be an unpopular redesign option for segment 1 students. This may reflect concern that larger tutorial groups reduce the time available for each student to participate in the solution of health care problems, increase the number of domineering students who monopolize conversational opportunities, or reduce the ability of tutors to accurately evaluate student progress (Lohfeld, Neville, & Norman, 2004). Simulations suggest that, despite these concerns, students would trade increases in tutorial group size for conceptual enhancements to the program.

Segment 2 constituted a group of students whose choices reflected a preference for a more traditional approach to undergraduate medical education. In contrast to segment 1, who preferred a problem-based approach to medical education, students in segment 2 were less sensitive to the prospect of increasing tutorial group size, preferred an increase in the number of more didactic large group sessions, and responded favorably to tutors who taught more didactically. In contrast to segment 1, who favored an unstreamed 3 year program, segment 2's students preferred a program streamed on the basis of career choices and learning process preferences. Not surprisingly, significantly fewer of these students were satisfied that McMaster's problem-based approach to learning met their educational needs.

The emergence of segment 2 is consistent with experience in educational settings in which a significant percentage of students afforded a choice opt for a more traditional approach (Colliver, 2000). A preference for a more traditional undergraduate medical program may reflect attitudes regarding the relative effectiveness of different approaches to medical education (Lohfeld, Neville, & Norman, 2004), cultural factors (Khoo, 2003), or the stresses associated with problem-based learning approaches (Moffat, McConnachie, Ross, & Morrison, 2004). For example, concerns about the availability of learning resources, individual study, progress, ability, and assessment can be significant sources of stress for students enrolled in the first year of a problem-based medical school (Moffat et al., 2004).

There were, however, a number of program design features on which segments 1 and 2 agreed. Both preferred tutorial problems that placed a greater emphasis on core curriculum concepts. Moreover, students in both segments evidenced a preference for a web-enhanced tutorial process and the addition of computer simulated tutorial problems. Implementation of the redesign suggestions emerging from our simulations yielded a program that fit the educational preferences of both segments.

The preference of students for an electronically enhanced undergraduate curriculum is consistent with a series of reports supporting the utility of computer simulated patients (Bearman, Cesnik, & Liddell, 2001; Hubal, Kizakevich, Guinn, Merino, & West, 2000; Parvati, et al., 2002), internet supported distributed problem-based learning (Stromso, Grottum, & Lycke, 2004), and on line clinical reasoning guides (Ryan, Dolling, & Barnet, 2004). While promising and of considerable interest to students, the cost and educational impact of these technologies requires careful study (Hudson, 2004).

Issues in the Application of Conjoint Analysis to the Design of Complex Services

This study raises several issues regarding the application of conjoint analysis to the design of complex educational services. First, the medical students participating in this choice study were familiar with the curriculum attributes of interest. It would be more difficult to assess the preferences of medical school applicants who are less familiar with the concepts of problem-based small tutorial group medical education. Although, familiarity may create a status quo bias in choice tasks (Salkeld, Ryan, & Short, 2000), students in the present study chose a significant number of redesign options.

Although the participants were familiar with the educational concepts on which our attributes were based, this study emphasized the importance of carefully piloting attribute level descriptions. For example, key informant interviews and an electronic focus group yielded several attribute descriptions that, though obvious to faculty, were poorly understood by students. These attributes required revision prior to fielding the survey.

Second, medical students participating in this study had the cognitive ability to process complex multidimensional choice tasks. Nonetheless, to improve informant efficiency, we limited choices to three alternative medical education programs described by three attribute levels per alternative (Patterson, 2003).

Third, the choice tasks we used were of considerable relevance to the participants. The motivation to understand the concepts presented and to consider the implications of different alternatives was relatively high.

Finally, although many students reacted negatively to the possibility of increasing the number of students participating in each tutorial group, choice tasks prompted students to consider tutorial group size in the context of other redesign options. The importance assigned to a conceptually and electronically enhanced curriculum suggests the use of discrete choice tasks contributed to a less biased consideration of the relative benefits of competing redesign options.

The results of the conjoint exercise are generally consistent with an independently conducted series of focus group with post graduate students who had completed medical school (Lohfeld et al., 2004). They are also consistent with educational theory emphasizing the more integrated focus on concept acquisition. Finally, the results were perceived to have great utility in supporting the development of McMaster University Medical School's revised undergraduate curriculum and considering admission practices. In the fall of 2004, McMaster University medical school announced its revised Compass Curriculum. While an increase in tutorial group size is inevitable, the Compass Curriculum includes a greater emphasis on fundamental medical concepts. Core curriculum concepts are integrated into tutorial problems, clinical and communication skills training, and clerkship patient selection and revisited in a spiral manner throughout the 3 year program. A task force to develop electronic enhancements has been established and a new admission process has been adopted (Kulatunga-Moruzi & Norman, 2002).

References

Allenby, G. M., Arora, N., & Ginter, J, L. (1995). Incorporating prior knowledge in the analysis of conjoint studies. *Journal of Marketing Research, 32*, 152-162.

Antepohl, W., Domeij, E., Forsberg, P., & Ludvigisson, J. (2003). A follow-up of medical graduates of a problem-based learning curriculum. *Medical Education, 37,* 155-162.

Barbour, R. S. (1999). The case for combining qualitative and quantitative approaches in health services research. *Journal of Health Services & Research Policy. 4*, 39-43.

Bearman, M., Cesnik, B., & Liddell, M. (2001). Random comparison of 'virtual patient' models in the context of teaching clinical communication skills. *Medical Education, 35,* 824-832.

Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine, 75,* 259-266.

Dolmans, D. H. J. M., Wolfhagen I. H. A. P., Scherpbier, A. J. J. A. & van der Vleuten, C. P. M. (2001). Relationship of tutors' group-dynamics skills to their performance ratings in problem-based learning. *Academic Medicine, 76,* 473-476.

Farrar, S., Ryan, M., Ross, D., & Ludbrook, A. (2000). Using discrete choice modeling in priority setting: an application to clinical service developments. *Social Science and Medicine, 50,* 63-75.

Farrow, R. & Norman, G. R. (2003). The effectiveness of PBL: the debate continues. Is meta-analysis helpful? *Medical Education, 37,* 1161-1132.

Fraenkel, L., Bodardus, S., Wittink, D. R. (2001). Understanding patient preferences for the treatment of lupus nephritis with adaptive conjoint analysis. *Medical Care, 39*, 1203-1216.

Harris, K. M. (2002). Can high quality overcome consumer resistance to restricted provider access? Evidence from a health plan choice experiment. *Health Services Research, 37*, 551-571.

Hubal, R. C., Kizakevich, P, N, Guinn, C. I., Merino, K.D., West, S. L. (2000). The virtual standardized patient. Simulated patient-practitioner dialog for patient interview training. *Studies in Health Information Technology, 70,* 133-138.

Hudson, J. N. (2004). Computer-aided learning in the real world of medical education: does the quality of interaction with the computer affect student learning? *Medical Education, 38,* 887-895.

Kaufman, D. M. & Mann, K. V. (1998). Comparing achievement on the Medical Council of Canada Qualifying Examination Part 1 of students in conventional and problem-base learning curricula. *Academic Medicine, 73,* 1211-1213.

Khoo, H. E. (2003). Implementation of problem-based learning in Asian medical schools and student's perceptions of their experience. *Medical Education, 37,* 401-409.

Kulatunga-Moruzi, C., & Norman, G. R. (2002). Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. *Teaching and Learning Medicine, 14*, 34-42.

Lenk, P. K., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity in reduced experimental designs. *Marketing Science, 15,* 173-191

Lohfeld, L., Neville, A., & Norman, G. (2004). Undergraduate medical training in Canada, Or, What do medical residents really think? Unpublished manuscript.

Maas, A. & Staplers, L. (1992). Assessing utilities by means of conjoint measurement: an application in medical decision analysis. *Medical Decision Making, 12*, 288-297.

Maudsley, G. (1999). Roles and responsibilities of the problem based learning tutor in the undergraduate medical curriculum. *British Medical Journal, 318,* 657-661.

Moffat, K. J., McConnachie, A., Ross, S. & Morrison, J. M. (2004). First year medical student stress and coping in a problem-based learning medical curriculum. *Medical Education, 38*, 482-491.

Mennin, S. P., Kalishman, S. Friedman, M. Pathak, D., & Snyder, J. (1996). A survey of graduates in practice from the University of New Mexico's conventional and community oriented, problem-based tracks. *Academic Medicine, 71,* 1079-1089.

Morgan, A., Shackley, P., Pickin, M., & Brazier, J. (2000). Quantifying patient preferences for out-of-hours primary care. *Journal of Health Services Research Policy, 5*, 214-218.

Neufeld, V. R., Woodward, C. A., MacLeod, & S. M. (1989). The McMaster MD program: a case study of renewal in medical education. *Academic Medicine, 64,* 423-432.

Newman, M. & The Pilot Review Group. (2003) *A Pilot Systematic Review and Meta-analysis on the Effectiveness of Problem-based Learning.* Newcastle: Learning and Teaching Subject Network for Medicine, Dentistry, and Veterinary Medicine. *http://www.ltsn-0.ac.uk/resources/features/pbl.*

Norman, G. R., & Schmidt, H. G. (1992). The psychological basis of problem- based learning: A review of the evidence. *Academic Medicine, 76,* 215-216.

Norman, G. R., & Schmidt, H. G. (2000). Effectiveness of problem-based learning curricula: Theory, practice and paper darts. *Medical Education, 34,* 721-728.

Orme, B. (2002). Interpreting conjoint analysis data. Sawtooth Software Research Paper Series. Sawtooth Software, Inc. Retrieved November 1, 2004, from http://www.sawtoothsoftware.com/download/techpap/interpca.pdf

Osman, L. M., Mckenzie, L., Cairns, J., Friend, J. A., Godden, D. J., Legge, J. S., & Douglas, J. G. (2001). P*atient weighting of importance of asthma symptoms, 56,* 138-142.

Patterson, M. (2003). Partial profile discrete choice: What's the optimal number of attributes. Sawtooth Software Conference Proceedings. Sequim: Sawtooth Software.

Dev, P., Montgomery, K., Senger, S., Heinrichs, W. L., Srivastava, S., & Waldron, K. (2002). Simulated medical learning environments on the internet. *Journal of the American Medical Informatics Association, 9,* 437-447.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The use of multiple strategies in judgment and choice*. Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Peters, A. S., Greenberger-Rosovsky, R., Crowder, C., Block, S. D., & Moore G.T. (2000). Long-term outcomes of the New Pathway Program at Harvard Medical School: a randomized controlled trial. *Academic Medicine. 75,* 470-479.

Phillips, K. A., Johnson, R., & Maddala, T. (2002). Measuring what people value: A comparison of "attitude" and "preference" surveys. *Journal of Health Services Research, 37,* 1659-1679.

Ratcliffe, J., Buxton, M., McGarry, T., Sheldon, R., & Chancellor, J. (2003). Patients' preferences for characteristics associated with treatments for osteoarthritis. *Rheumatology*, *43,* 337-345.

Ryan, G., Dolling, T., & Stewart, B. (2004). Supporting the problem-based learning process in the clinical years: evaluation of an online Clinical Reasoning Guide. *Medical Education, 38,* 638-645.

Ryan, M. (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal, 320*, 1530-1533.

Ryan, M. (1999). Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilization. *Social Science and Medicine, 48,* 535-546.

Ryan, M., & Gerard, K. (2003). Using discrete choice experiments to value health care: current practice and future prospects. *Applied Health Economics and Policy Analysis, 2,* 55-64.

Ryan, M., & Farrar, S. (2000) Using conjoint analysis to elicit preferences for health care. *BMJ, 320*, 1530-1533.

Ryan, M., Scott, D. A., Reeves, C., Bate, A., van Teijlingen, E. R., Russell, E. M., Napper, M., & Robb, C. M. (2001). Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technology Assessment, 5,* 1-186.

Salkeld, G., Ryan, M., & Short, L. (2000). The veil of experience: Do consumers prefer what they know best? *Health Economics. 9,* 267-270.

Sawtooth Software Inc (2004a). The CBC latent class technical paper (version 3) Retrieved November 1, 2004, from http://www.sawtoothsoftware.com/download/techpap/lctech.pdf

Sawtooth Software Inc (2004b). The CBC/HB System for Hierarchical Bayes estimation (version 3.2). Sawtooth Software Technical Paper Series. Retrieved November 1, 2004, from http://www.sawtoothsoftware.com/download/techpap/hbtech.pdf

Schmidt, H. G. (1993). Foundations of problem-based learning: Some explanatory notes. *Medical Education, 27,* 422-432.

Schmidt, H. G. (1994). Resolving inconsistencies in tutor expertise research: Does lack of structure cause students to seek tutor guidance? *Academic Medicine, 69,* 652-662.

Schmidt, H. G., & Molen, H. T. (2001). Self-reported competency ratings of graduates of a problem-based medical curriculum. Academic Medicine, 76, 466-468.

Schmidt, H. G. van der Arend, A., Moust, J. H. C., Kokx, I., & Boon, I. Influence of tutor's subject-matter expertise on student effort and achievement in problem-based learning. *Academic Medicine, 68,* 784-791.

Singh, J., Cuttler, L., Shin, M., Silvers, J. B., & Neuhauser, D. (1998). Medical decision-making and the patient: understanding preference patterns for growth hormone therapy using conjoint analysis. *Medical Care, 36*, 31-45.

Stanek, E. J., Oates, M. B., McGhan, W. F., Denofrio, D., & Loh, E. (2000). Preferences for treatment outcomes in patients with heart failure: Symptoms versus survival. *Journal of Cardiac Failure, 6,* 225-32.

Stromso, H., Grottum, P., & Lycke, K. H. (2004). Changes in student approaches to learning with the introduction of computer-supported problem-based learning. *Medical Education, 38,* 390-400.

APPLIED CHOICE MODELS

Over-Estimation in Market Simulations: An Explanation and Solution to the Problem with Particular Reference to the Pharmaceutical Industry

Adrian Vickers, Phil Melleor and Roger Brice
Adelphi International Research, UK

Introduction

The nature of the pharmaceutical product development means that ability to change product design to reflect customer preferences is extremely limited. Clinical trials can be powered to provide statistical significance for properties that a drug may be expected to exhibit, aiding FDA approval for their inclusion in product promotion. This means increasing sample sizes and extending time lines, each with associated cost and competitive risk. Time and money can also be invested in developing new formulations and drug delivery methods; again with cost and time implications. Equally, if the market demands combinations of product properties that cannot be delivered, either at all or within a cost-effective competitive landscape, the development process may be truncated or even abandoned.

Outputs from discrete choice models are increasingly becoming important inputs to development investment decisions within the pharmaceutical industry. Utility values identify relative importance of alternative product features, but the risk, time and cost analysis associated with investment decisions must include uptake forecasts for alternative eventual new product profiles. The focus is therefore much more on the market simulation output than on utility values. With both discrete choice and traditional conjoint models, a common problem of over-estimation of new product uptake has been identified (Ayland & Brice, 1999). Even when we can replicate the current market well, we find that new product predictions frequently over-estimate the eventual reality. This has occupied much discussion at industry market research conferences. Discount factors of up to three have been mentioned with no associated basis in economic theory or models of customer behaviour.

We have taken steps to improve both the choice question and its setting. Partial profiles (usually using CBC) have addressed the problems of dominant attributes and the number of attributes affecting utility scores (Sawtooth, 1999). The choice question can reflect whether the chosen profile would be used, as well as how it would be used, accommodating changes in co-prescribing of two or more products. Our preferred setting is usually a recent patient, rather than the alternative ‘out of x patients’ setting, with separate choice exercises for a patient within each of a series of patient segments. Respondents have commented on their preference for this approach and it reflects what physicians do (making treatment decisions for individual patients). This is also considered to be the current best practice. Louviere *et al*. (2000) states that the objective should be to simulate real markets as closely as possible. Unfortunately the individual patient setting may exacerbate the over-estimation problem we are trying to resolve. The physician is responding logically to the choices posed for each of a series of patients and, if he/she were the sole decision-maker, the resulting share predictions would not be unreasonable. Marketplace choices are influenced by many factors (Wittink, 2000). If these complexities have a sufficiently large effect then the extent to which a market simulation can be seen as a useful predictive tool could be seriously undermined. External effects impact the percentage of patients that might get a new drug and market simulations need to accommodate this. Hall *et al*. (2004) state that health care decision making and spending are often characterised by a complex interaction between the consumer, the physicians and other providers. They argue that in order to tackle contemporary health systems, understanding better the decision making of all factors in the system is imperative. The degree to which external factors influence physicians varies greatly (Rosenthal *et al*., 2001; Landon *et al*. 2001, Landon *et al*., 1998; Reschovsky *et al*., 2001; Conrad & Christian, 2002). Landon (2004) argues that organisational characteristics only explain a small amount of the variance observed in clinical practice. Factors related to individual patients' preferences and characteristics of the physician treating the patient all have important effects. Our paper addresses this issue. This paper tackles the problems of the complex physician patient interaction by conducting the choice tasks at this level. The assumption is that physicians can factor in the patient's response when we ask them to imagine that they are choosing a product for a named patient. We have also used the physicians' own perception of the external factors as a basis for changing the assumptions used in patient share simulations. In order to completely represent the healthcare decision making process for a new drug, patient and third party payer opinions also need to be consulted. This is not addressed in this research.

CURRENT BEST PRACTICE

Discrete choice designs are generated using Sawtooth Software's CBC System for Choice-Based Conjoint which normally comprise 16-20 tasks where respondents pick their preferred profile out of three alternatives. Physicians are asked to consider a particular recent patient when making this choice. The exercise may be repeated for different sets of cards for up to three different patient types. Partial profiles are used if the number of attributes in a study is greater than six (see Sawtooth Software, 1999). Physicians are asked to assume that the cost of the treatment is either same as market leader or comparable to the last new drug of this type if the product is at too early a stage of development for a price to be stated. The data collected are analysed by Sawtooth Software's CBC/HB for Hierarchical Bayes estimation and the individual parameter estimates generated.

THE THRESHOLD QUESTION

A number of further questions are asked where physicians are shown full profile descriptions for new products and asked whether they would have definitely prescribed the new product instead of the current treatment or added it to the patient's current treatment. This type of question can also be used to find the current and new treatment combination to understand fully the market change when co-prescribing occurs. Once the parameter estimates have been determined, the sum of utility scores for these additional full profile products can be calculated. These values can then be compared to the physician's choice of whether they would have actually prescribed the new treatment or not. The sum of utilities that will give a 50% chance of being prescribed is then estimated (the 'cut-off') for each individual physician/patient consultation. This is similar to the methodology of Brice *et al*. (2000). The 'cut-off' value is typically much larger than the value of the current treatment in terms of purely its sum of utilities, from the attribute and level scores. Once the 'cut-off' value has been determined it is included in the market share simulation using a logistic function to generate the probability of uptake for the new treatment for each individual physician/patient consultation. The predicted share produced using this method means that the current treatments do not need to be included in the simulation. This would not be recommended because the physicians do not behave in a way that is typical of non-pharmaceutical markets because they make many repeat prescription decisions but are not brand loyal. Physicians may be reluctant to change the patient's current therapy due to risk associated with doing so or the level of satisfaction with the patient's current therapy.

OSERVATIONS ON PHARMACEUTICAL SHARE PREDICTIONS

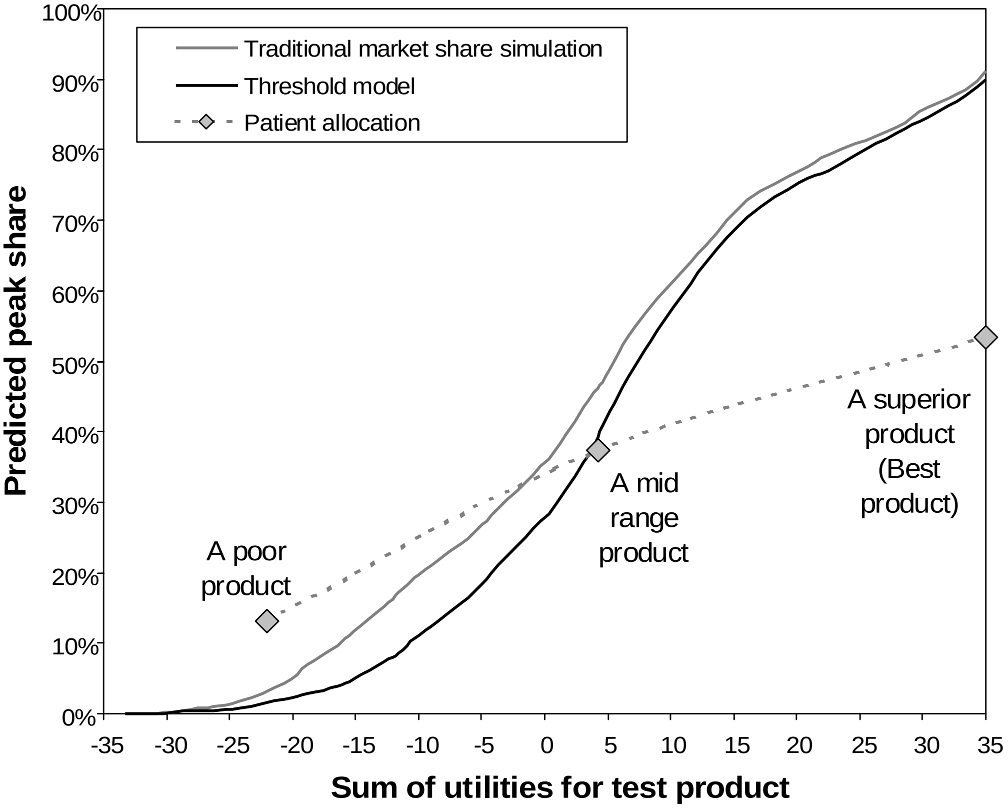
Despite the use of the 'cut-off' values in the simulation, the model can still under certain situations greatly over-estimate market share. Typically we find that this model gives predicted uptake values that are close to what market knowledge would suggest for new products that can only offer a small improvement over current treatments. This is the most common situation. However, based on our past experience, for new products that offer significant improvement over current treatments, the predicted share can be much higher than what it is likely to be in reality.

CASE STUDY (1)
A NEW DRUG FOR THE TREATMENT OF HIGH BLOOD PRESSURE

The design of the study followed the methodology outlined above and included threshold questions and calculation. In addition physicians were asked to profile a number of current types of treatment. Simulations of market share could then be determined for a model based purely on the sum of utilities of different products and a model based on the threshold calculation.

In one particular example the expected profile for the new blood pressure lowering drug had a mean sum of utilties of –14. However, high blood pressure can be treated satisfactory with a number of very cheap drugs called diuretics. The current market leader only has a market share of 18% with diuretics making up over 50% of the market. Figure 1 shows that if the sum of utilities is high enough, the predicted share could be close to 90%. Considering the cost implications of treating 90% of patients with this condition with a new drug such as this, and the fact that formularies (procedures which restrict use of drugs in say a hospital) and other decision makers will restrict the physician's budget, a much lower share would be expected. Physicians were also shown three set profiles and asked—"What percentage of patients would receive each product?" The results are also shown in Figure 1.

Figure 1. The effect of threshold question. Data is from a study for a new blood pressure lowering drug. Sample size: 264 physicians (792 consultations).

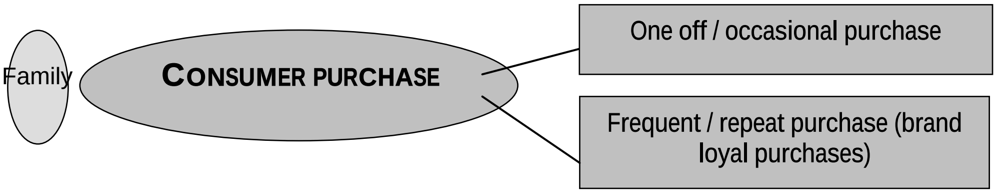


External Effects

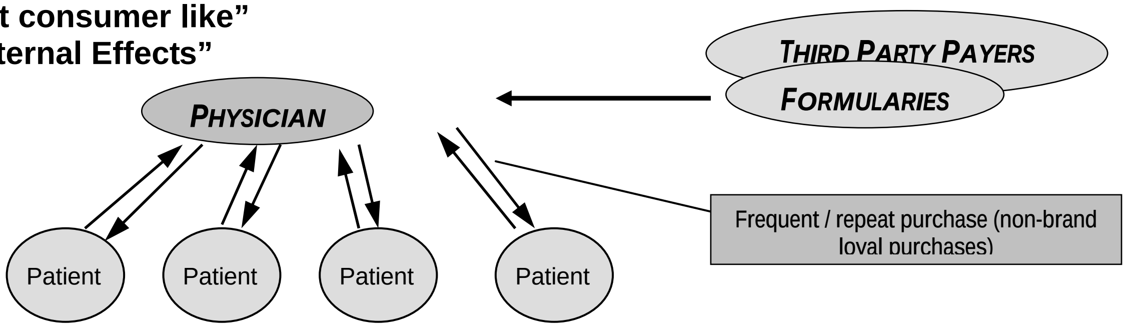
From the results of the study outlined above it would appear that physicians do not consider all the cost implications when they asked to only consider up to three patients. They make the choice on whether to prescribe or not, predominantly on whether the clinical data provided are sufficiently good enough to warrant changing the patient's treatment.

The models below help to emphasise the difference between the pharmaceutical market and the non-pharmaceutical market.

Non pharmaceutical decision-making


Family
CONSUMER PURCHASE
One off / occasional purchase
Frequent / repeat purchase (brand loyal purchases)

Pharmaceutical decision-making
"Not consumer like"
"External Effects"



The extent to which the external effects determine final market share, are dependent on the number of patients with a particular condition (market size), how serious the disease is, and what budget constraints exist. For a common disease which is not life threatening, there will be a greater third party focus on costs. This will be the opposite where a disease is normally fatal and where there are only a small number of patients with the condition. The competitive nature of the market will also have a part to play. If there is a large number of different drugs available and if physicians are not sure which product is the best or which one is most appropriate for different patients, then their prescribing behaviour may become close to a random decision process. Under these circumstances the physicians might allocate the drugs equally across patients. It may be only physicians that have specialised in that particular disease area that have a clear idea which patients should receive which drug.

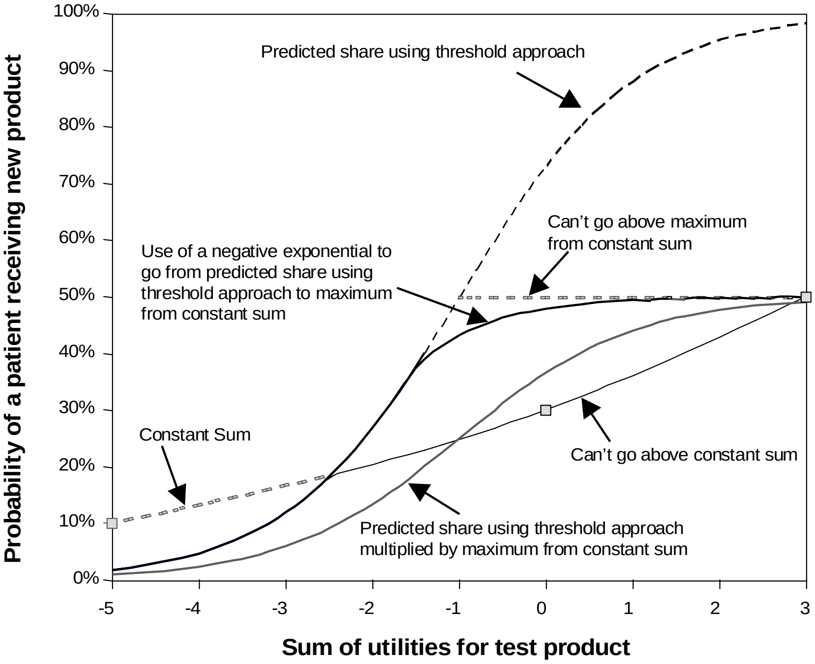
DEALING WITH THE EXTERNAL EFFECTS

The patient allocation data in Figure 1 suggests that physicians are taking the external effects into account when asked—"What percentage of your patients would realistically receive the new treatment (after a period of time for which peak share can be assumed)?" From our experience physicians tend to over-estimate the uptake of a poor product. Maybe they feel obliged to give it some share. This might also be exacerbated by the fact they tend to go up and down in 10% increments so that if they do give a poor product some share it is likely to be estimate at 10% or 20%. Many products in a pharmaceutical market are likely to obtain less than 5% market share.

We assume generally that, while physicians have difficulty estimating volumes, they are able to give reliable estimates of upper limits, if only by reference to their current pattern of prescribing. From experience it is also likely that these values are also over-estimated, but to a smaller extent. We therefore ask physicians to give their upper limit of prescribing, given the current constraints *etc*. that are imposed upon them, if the best possible product was available from the attribute and levels grid. We can have some confidence that predicted estimates should not exceed this value.

The next question that needs answering is how to apply this information to each individual. Figure 2 shows a number of possible ways to apply the maximum from the patient allocation question to create a new model.

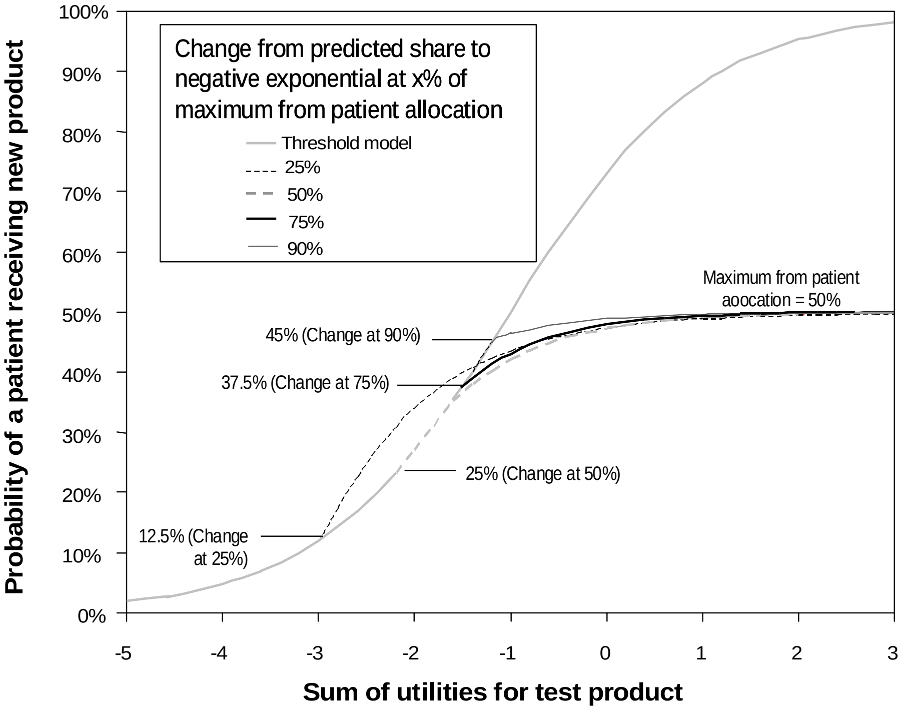
Figure 2. Possible ways of incorporating the maximum value from the patient allocation exercise to limit the predicted uptake.



As mentioned before, so far the predicted share for relatively poor products appears to be fairly well predicted by including the threshold calculation. We need a method that can leave this area of the curve unchanged. Our preferred choice is therefore to use the curve with the threshold calculation until it reaches a certain percentage of the maximum from the patient allocation exercise, and then use a negative exponential to go from this point to the maximum value. This assumes that that the physician has a predefined maximum percentage for the patients that can receive a new and relatively expensive product for a particular disease area, and that the physician is only concerned about this value as this maximum value is approached. If we were to simply multiply the probability value from the logistic equation by the maximum value from the patient allocation question then we would be assuming that the physician feels under pressure not to prescribe the new product before he/she has even started prescribing it. In practice, however, it may take a phone call from a third party to notify the physician of a high number of prescriptions for a new drug. If this is the case then our original assumption holds *i.e.*, the probability of prescribing will follow the logistic curve until a certain percentage of the maximum is reached and then the form of the curve will change to reach the new maximum.

The next question that needed answering was at what point should the predicted uptake change from the logistic curve to the negative exponential. Figure 3 shows that this decision actually makes very little difference to the shape of the curve. Even going from 10% of the maximum and constraining the new predicted value to be no greater than that from the threshold model still gives a very similar result to the rest. The value that we have chosen to use is 75%. This analysis does show that our model can be expected to be fairly robust and should produce better results than the use of a simpler model.

Figure 3. Point of change from the logistic equation for the threshold model to the maximum from the patient allocation exercise using a negative exponential curve.



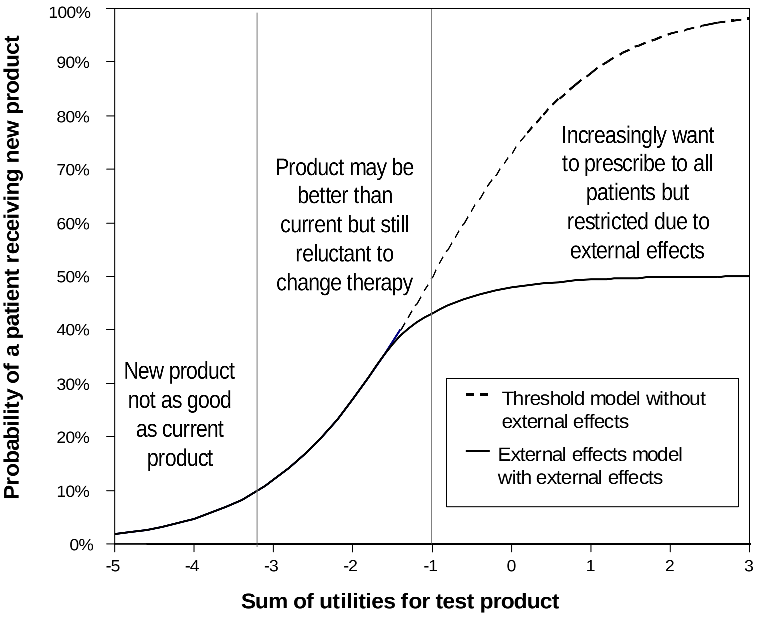
The Assumptions of the New Model

The assumptions behind our new model are now radically different from the traditional market simulator where probability is derived from just the sums of utilities for different products. Figure 4 summarises these differences. The predicted uptake now has three different components depending upon how good the product is, the physician's reluctance to change the patient's therapy and to what extent external effects limit the number of patients eligible to receive the new treatment. This gives three stages to the probabilistic curve. The first is where the new product is considered to be poorer than the current treatment so there is a very low probability of uptake. The second is where the new product is perceived to be equal or slightly better than the current treatment but the risks or work involved to change the current treatment do not outweigh the benefits to the patient. The third is where the physician is convinced of the benefits of the new treatment and wants to prescribe it to all patients with the particular condition but cannot do so due to budget limitations *etc*.

This new model would imply that no matter how much improvement a new drug might offer compared to existing products it will not continue to increase in share beyond a certain point. This is likely to occur where there is a limit to the money available to treat the condition and once this is reached then the new product will not increase in share any more. However, it may not mean that gaining extra clinical trial data *etc.* is no longer worth while once this maximum has been reached, since the more the product can offer above other treatments, the more likely the product is to get reimbursement from the third party payers and formularies.

With the advance of Hierarchical Bayes it means that the external effects can be applied to the level of the individual physician patient consultation.

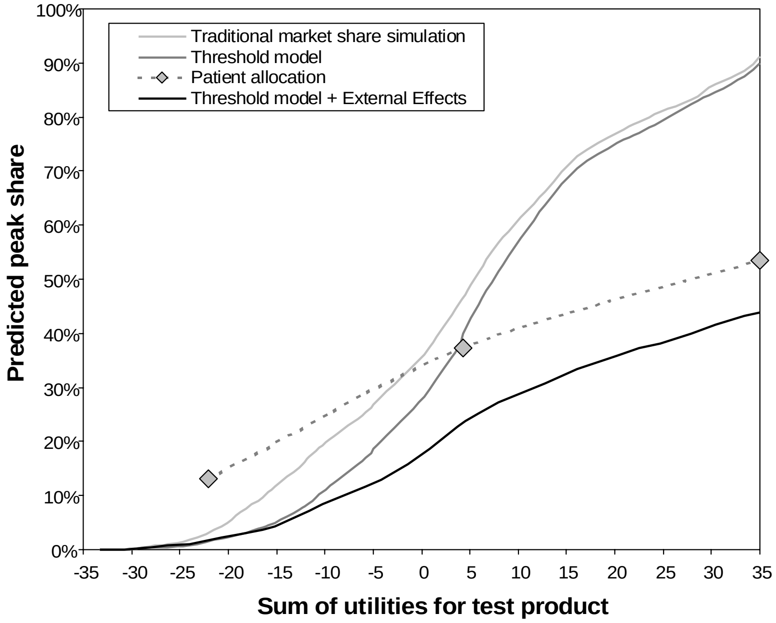
Figure 4. The new assumptions of the model that incorporated the threshold calculation and the external effects model.



CASE STUDY (1) REVISITED

The external effects model was combined with the predicted uptake from the threshold calculation to give a new predicted uptake calculation for each consultation. The new share estimate can be seen to be much lower than that predicted by the other methods. The predicted values do now lie in the range suggested by market knowledge. Considering this is a market that is dominated by readily available very cheap alternatives that many physicians consider to give satisfactory results, this large drop in share is in line with what market knowledge would predict.

Figure 5. The external effects model applied to the high blood pressure drug data.

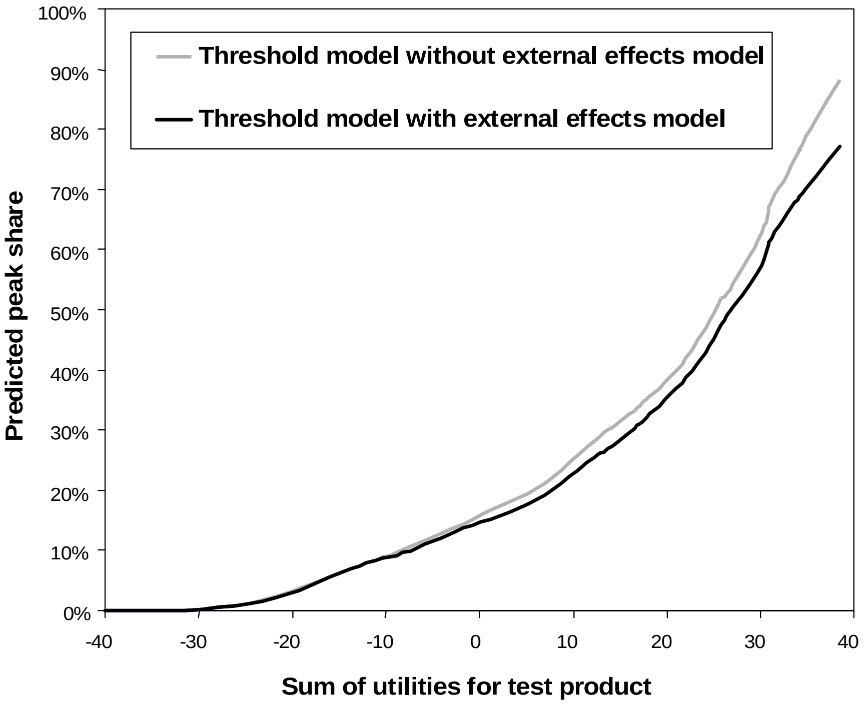


CASE STUDY (2) A NEW ANTI-CANCER TREATMENT

Data were collected and analysed using our standard approach and including the external effects model. The new product on offer is for the treatment of early stage breast cancer. The results of the simulations are shown in Figure 6.

According to our assumptions above the external effects calculation should have little impact on predicted market share because of the serious nature of this disease. This can be seen with the resulting share estimates. In this example the external effects model only starts to have an impact for a new product with a sum of utilities of greater than zero. The maximum effect it has is to decrease the predicted share by 10%. In this case the manufacturer’s product had a sum of utility score of close to five which meant that the external effects calculation only dropped the share by two percent.

Figure 6. The external effects model applied to data collected from a market simulation for a new product for the treatment of early stage breast cancer.

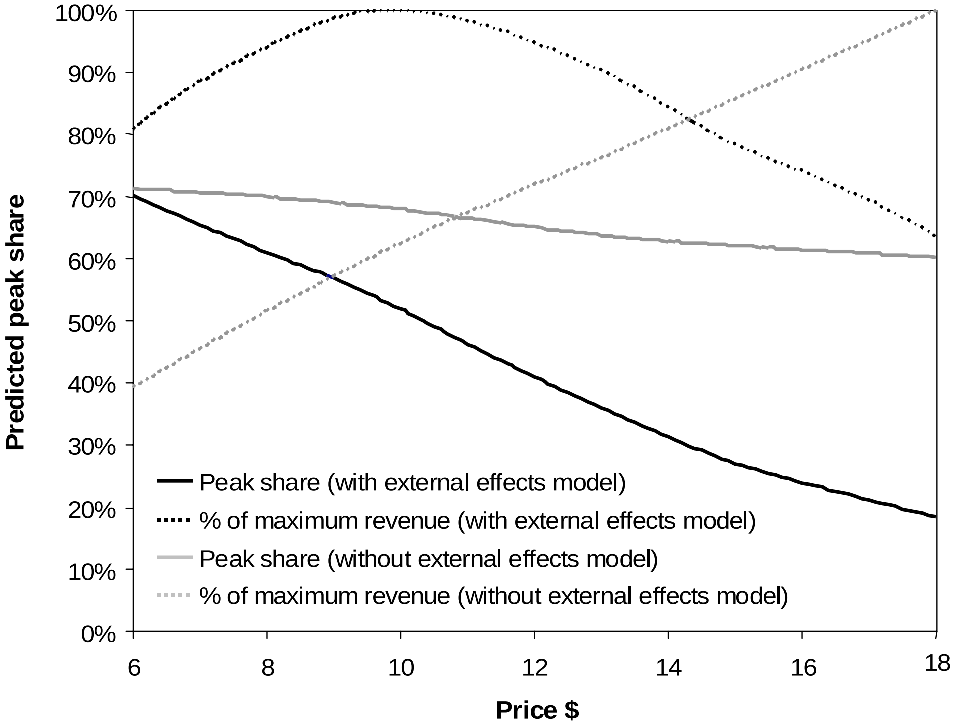


Case Study (3) A Pricing Study for a 'First in Class' Product for the Treatment of Osteoporosis

This study was carried out in a different way to the ones above. The choice-based design and the threshold method were used in the same way but the external effects model was built in a more complex way since price was one of the attributes. Physicians were asked to circle levels they considered to be the best on the attributes and levels grid and were then asked to give what percentage of their patients would receive the new product. This was done for five different prices ($6, $9, $12, $15, $18). Constrained cubic spline equations were then used to estimate the values in between these levels. These new estimated values went into the external effects calculation. The resulting market share predictions are shown in Figure 7. The simulation is based on the target profile for the product which was about two thirds of the maximum value from the attribute and levels grid.

The predicted share for the model without the external effects calculation appears to be very price insensitive. The share only drops from 72% to 62% for the whole of the price range tested. This results in the index of maximum revenue continuing to increase for the full length of the given price range. This again suggests that the physicians do not consider budget constraints *etc.* when asked to consider a single patient. They only judge the products offered in terms of the clinical attributes and many probably ignore price. We often ask physicians to consider the last patient they considered for treatment when they do the discrete choice exercise. This may have a bias since the patient they saw last is also more likely to be the patient the physician sees most frequently. Patients that are seen more frequently are likely to have the most severe symptoms or are less satisfied with the treatment they are currently receiving. This may mean the patient we ask the physician to consider is the one that is most likely to receive the new product. However, through the use of the external effects calculation we now ask the physician to consider all of their patients and so this bias is largely removed.

Figure 7. The external effects model applied to data collected from a market simulation for a new product for the treatment of early stage breast cancer. The % maximum revenue refers to the predicted share multiplied by the price to give a revenue value. This value has then been indexed so that the maximum value for the given price range has been given a value of 100% and then all other values are relative to this.



In contrast to the model without the external effects calculation, the simulation with the external effects shows a large drop in share from 70% to 20% for the most expensive product profile. This gives an optimum price well within the range tested of $10. This price of $10 is only slightly more than the cost of current treatments. However, it is possible that by having a set of questions that only refer to price that we may have made physicians overly sensitive to price. In this particular situation market knowledge suggests that formularies and third party payers would have considerably limited the number of patients receiving this product if it had been priced at the top end of the range tested. This suggests that the model is behaving in a realistic way.

CONCLUSIONS

Our new model has achieved the following

1. It addresses issues relating to the reluctance to change the current treatment of a patient.
2. The extent to which external effects limit the extent to which a physician can give all their patients a new treatment.
3. The external effects method also gets round the problem of whether to use the ‘choose none’ option in the discrete choice design. The uptake might be zero but there is no loss in information in order to calculate utility scores.
4. There is also no need to use additional parameters in the logistic equation to reduce the steepness of the probability curve.

The use of the ‘choose none’ option is known to increase during the exercise (Johnson & Orme, 1996; Sattler *et al*., 2003). From a review of the studies we have done so far there appears to be much less bias in this question in terms of changes taking place during the exercise.

The fact that we are only modelling a few new competitors against the threshold value in the logistic equation also greatly reduces the problem of the IIA (Independence from Irrelevant Alternatives) that still occur at the individual level even with the use of Hierarchical Bayes. Instead of modelling a vast number of different products with many having similar properties we are only modelling up to five new products against the current treatment.

The drop in share due to external effects varies considerably according to the particular situation—the seriousness of the disease, the scarcity of the disease, the competitive nature of the market. All these factors mean that it is not possible to come up with a single correction factor, that will reduce the predicted share to a realistic level, and that will work in all situations. Applying different correction factors under different scenarios might be practical, but there would be the risk of making the wrong assumptions about the market. Also such an approach would not be at the individual level. The view of market, priority of treatment *etc*. may vary according to physician, hospital, country *etc*. By being able to include the external effects calculation at the level of individual physician/patient consultations, the simulation should behave in a very realistic way and so produce much more reliable market share predictions.

Modelling external effects in the way presented may not lend itself particularly well to other market scenarios outside the pharmaceutical industry. However, there may be a number of specific scenarios where the external effects model could be used.

1. Any situation where a person with specialist knowledge makes a decision on behalf of others e.g. financial advisor
2. For repeat purchases where a more expensive product might be bought on special occasions

The threshold calculation may be of particular interest where respondents might be reluctant to change from their current product. This may be particularly true for service industries where there maybe considerable effort involved to use a different bank, supermarket, electricity provider *etc*.

This paper has so far mainly had a theoretical basis for the method we have presented. The best tests will be how accurately it is able to predict changes to future markets once the new products are launched. It is in the light of this type of information that the main developments of CAPMOD from the paper by Brice *et al*. (2000) to the paper presented here have occurred.

The particular characteristics of the pharmaceutical industry make this an even greater challenge.

1. The time between market research studies and the launch of the new product/time to physician awareness is very long
2. Policies and guidelines change—budgets, recommendations etc.
3. Other competitors are launched which are beyond the ranges tested in the original attribute and levels grid
4. The new product does better or worse than expected—performs outside of the levels tested
5. The new product may not be launched—poor performance in clinical trials, new side effects emerge etc.

As the number of simulations using this new approach increases so our knowledge base will improve and we should gradually gain more experience as to how close our simulations get to reality. As this is achieved we will be able to continue the further development of CAPMOD (now termed Choice-Adapted Predictive Modelling).

What we have achieved is a step closer to converting preference to market share. Conjoint used on its own only predicts preference, not market share (Orme, 1996). However, in the pharmaceutical market much knowledge already exists about preference but not how this preference can be translated into market share. Other external factors have not been addressed by this piece of research such as the time differences between physicians to become aware of new products and also different launch sequences and promotional efforts for new products. These still need to be factored in through market knowledge to arrive at a final market prediction. However, through being able to model two very important properties of this particular market the need for guesswork has been greatly reduced. This should result in much greater precision in share predictions in a very complex market.

References

Ayland, C. & Brice, R. (1999) From Preference Share to Market Share: One Giant Leap. *Marketing and Research Today*, 28, (3).

Brice, R., Mellor, P. and Kay, S. (2000) Choice-Adapted Preference Modelling. *Sawtooth Software Conference Proceedings*, 59-69.

Conrad, D. A., & Christianson, J.B. (2004). Penetrating the 'Black Box': Financial Incentives for Enhancing the Quality of Physician Services. *Medical Care Research and Review*, 61, (3), 37S-68S.

Hall, J., Viney, R., Haas, M. and Louviere, J.J. (2004) Using stated preference discrete choice modelling to evaluate health care programs. *Journal of Business Research,* 57, (9), 1026-1032.

Johnson, R.M. & Orme, B. (1996). How many questions should you ask in choice-based conjoint studies? *Conference Proceedings of the ART Forum, Beaver Creek.*

Landon, B.E. (2004) Commentary on "Penetrating the 'Black Box': Financial Incentives for Enhancing the Quality of Physician Services," by Douglas A. Conrad and Jon B. Christianson. *Medical Care Research and Review,* 61, (3), 69S-75S.

Landon, B.E, Reschovsky, J.D., Reed, M.C., Blumenthal, D. (2001) Personal, Organizational Market Level Influences on Physicians' Practice Patterns: Results of a National Survey of Primary Care Physicians. *Medical Care*, 39, (8), 889-905.

Landon, B.E., Wilson, I.B., Cleary, P.D. (1998) A Conceptual Model of the Effects of Health Care Organizations on the Quality of Medical Care. *Journal of the American Medical Association,* 279, (17), 1377-1382.

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

Orme, B. (1996). Helping Managers Understand the Value of Conjoint. *Sawtooth Software Research Paper Series. Sawtooth Software, Inc.*

Reschovsky, J.D., Reed, M.C., Blumenthal, D. and Landan, B. (2001) Physicians' Assessments of Their Ability to Provide High Quality Care in a Changing Health Care System. *Medical Care*, 39, (3), 254-269.

Rosenthal, M.B., Landon, B.E. and Huskamp, H.A. (2001) Managed Care and Market Power: Physician Organizations in Four Markets. *Health Affairs*, 20, (5), 187-193.

Sattler, H., Hartmann, A. and Kroger, S. (2003) Number of Tasks in Choice-Based Conjoint Analysis. *Research Papers on Marketing and Retailing No. 13, University of Hamburg.*

Sawtooth Software, Inc. (1999) *CBC User Manual Version 2.0.* Sawtooth Software, Inc.

Wittink, D.R. (2000) Predictive Validation of Conjoint Analysis. *Sawtooth Software Conference Proceedings*, 221-237.

COMMENT ON VICKERS, MELLOR AND BRICE

LUIZ SÁ LUCAS
IDS-INTERACTIVE DATA SYSTEMS

Most of market research work deals with DISPOSITIONS. Following Rosenberg (1979) we could say a disposition refers not to a state or a condition of the individual, but instead to a tendency to react in a certain way to a certain kind of stimulus.

Rosenberg gives two examples:

- Consider the term 'elastic'. When we say a piece of material is elastic, we do not mean it is in process of being stretched; we mean that it is *able* to be stretched
- Similarly, we say a piece of copper is a good electrical conductor, even when electricity is not induced in it.

Hempel (1952, quoted by Rosemberg (1979)) speaks of dispositions as potentials or tendencies to respond in certain ways under specifiable circumstances.

Most of major dependent variables of survey research are essentially dispositional concepts. But again following Rosenberg dispositions:

- Are not directly observable
- Are poor predictors of behavior

Well, Conjoint Analysis essentially estimates a PREFERENCE SYSTEM for a given respondent. We can look at it as a DISPOSITIONAL SYSTEM.

The question is that in real world MR problems client needs Market Share much more than they need Share of Preference. So we go back to the problem of bridging as best as we can the gap between attitude (disposition) and behavior.

I would accept dispositional systems can be reasonable accurate predictors provided we can specify carefully circumstances that would trigger specific behavior. And I think this is exactly the direction of Vickers, Mellor and Brice's work (2004). Using devices such as Cut-off Threshold and the "External Effects" Resource Allocation techniques, they try as best as they can to define the circumstances that can trigger a specific set of prescriptions by a physician.

This is done at individual level, as Wittink (2003) suggested last year in our Sawtooth Conference, when commenting on two papers that dealt with accurate prediction of real outcome from changes in product and services or their price.

This is really a very interesting contribution and we should encourage other works in similar areas: we need work that can fill, as best as we can, the gap between attitude and behavior.

REFERENCES

Hempel, C. (1952). *Fundamentals of Concept Formation in Empirical Science*. International Encyclopedia of Unified Science. Vol. II No. 7. University of Chicago Press.

Rosemberg, M. (1979). *Dispositional Concepts in Behavioral Science, in Qualitative and Quantitative Social Research – Papers in Honor of Paul Lazarsfeld, ed. by Merton, R., Coleman, J. and Rossi, P.* Free Press.

Vickers, A., Mellor, P. and Brice, R.(2004). *Over-estimation in Market Simulations: an Explanation and Solution to the Problem with Particular Reference to the Pharmaceutical Industry*. Sawtooth Software Conference 2004 Proceedings

Wittink, D. (2003). *Comment on Arenoe and Rogers/Renken* . Sawtooth Software Conference 2003 Proceedings, 233-236

ESTIMATING PREFERENCES FOR PRODUCT BUNDLES VS. *A LA CARTE* CHOICES

DAVID BAKKEN, MEGAN KAISER BOND
HARRIS INTERACTIVE

Conjoint analysis has long been used to identify optimal combinations of features to include in a product. This is accomplished by estimating the utility or "part-worth" of every feature that might be included in the product *bundle* and then comparing the aggregate utilities for different bundles. Conjoint analysis is a decompositional method, and as long as we are interested only in finding the best bundle of features, works well. However, there are many product and service situations where consumers are offered features or benefits both in bundles and as "a la carte" choices. This situation is known as *mixed* bundling, in contrast to *pure* bundling, where the features or benefits are available only in predetermined (by the firm) bundles.

THEORY OF BUNDLING

Bundling, or the combining of multiple "benefits" that are separately available into a single package, is typically viewed from the perspective of the firm rather than the consumer. The theory of bundling, such as it is, therefore focuses primarily on the ways in which bundling benefits the firm. For example, we might ask if a bundled offer will reach more consumers or enable the firm to realize higher margins. Bundling may confer competitive advantage by concealing information about either the costs or prices for the component benefits. Most importantly, as noted by Lilien, Kotler and Moorthy (1992), bundling is a way to implement *discriminatory pricing*. Discriminatory pricing takes advantage of reservation price heterogeneity by setting different prices for different customers based on their different reservation prices. However, many forms of discriminatory pricing are illegal. In theory, bundling exploits a particular form of reservation price heterogeneity in a way that does not directly discriminate. If segments exist for which reservation prices for different benefits are negatively correlated, then bundling can be an effective strategy for increasing sales for each of the separate benefits. We can see this in the following example. A quick service restaurant offers three different items: hamburgers, French fries, and soft drinks. Imagine that there are two distinct segments, A and B, of customers for this restaurant. Segment A, comprised of 50 people, is willing to pay up to $3.00 for a hamburger, $.50 for fries, and $1.00 for a soft drink. The 75 consumers in segment B will pay only $1.00 for a burger, but they will pay up to $1.50 for fries and $2.00 for a soft drink. Each segment is willing to pay $4.50 in total for the three items. Assume that the restaurant decides to charge $3.00 for the hamburger, $1.50 for fries, and $2.00 for a soft drink (i.e., the highest reservation prices across the two segments). In the best-case scenario, where every one of the 125 customers buys the items that are equal to or less than their reservation prices, the restaurant will realize $412.50 in revenue (selling 50 burgers @ $3.00, 75 orders of fries @ $1.50, and 75 soft drinks @ $2.00). The restaurant might decide to charge the lowest of the two reservation prices for each item, in which case total revenues will fall to

$312.50 as every consumer purchases all three items at a total price of $2.50 (which is lower than the total reservation price of $4.50 for each segment). However, if the restaurant bundles all three items into a "combo" meal and charges $4.50, assuming that all 125 consumers would prefer to buy all three items, the total revenue will be $562.50. In this case bundling is inclusive, since the strategy exploits the negative correlation in reservation prices (consumers who are willing to pay the high price for a burger will only pay a low price for French fries, and vice versa). This strategy works only if consumers are unaware of the separate item prices and the items are complementary, so that all consumers would prefer to purchase all three items in the bundle.

In this economic theory of bundling, the firm benefits as long as it is not forced to sell one or more components of the bundle to some consumers at below the marginal cost of that benefit. In our quick service restaurant example, assume that the cost of each item is 50% of the *highest* of the two reservation prices. Selling the individual items at the highest reservation price for each item results in a profit of $206.25. Consumers in segment B are excluded from buying the burger, and consumers in segment A are excluded from buying the fries and soft drink. With pure bundling (at a price of $4.50), the marginal cost of the bundle of three items is $3.25 ($1.50 + $.75 + $1.00) and the total profit realized is $156.25.

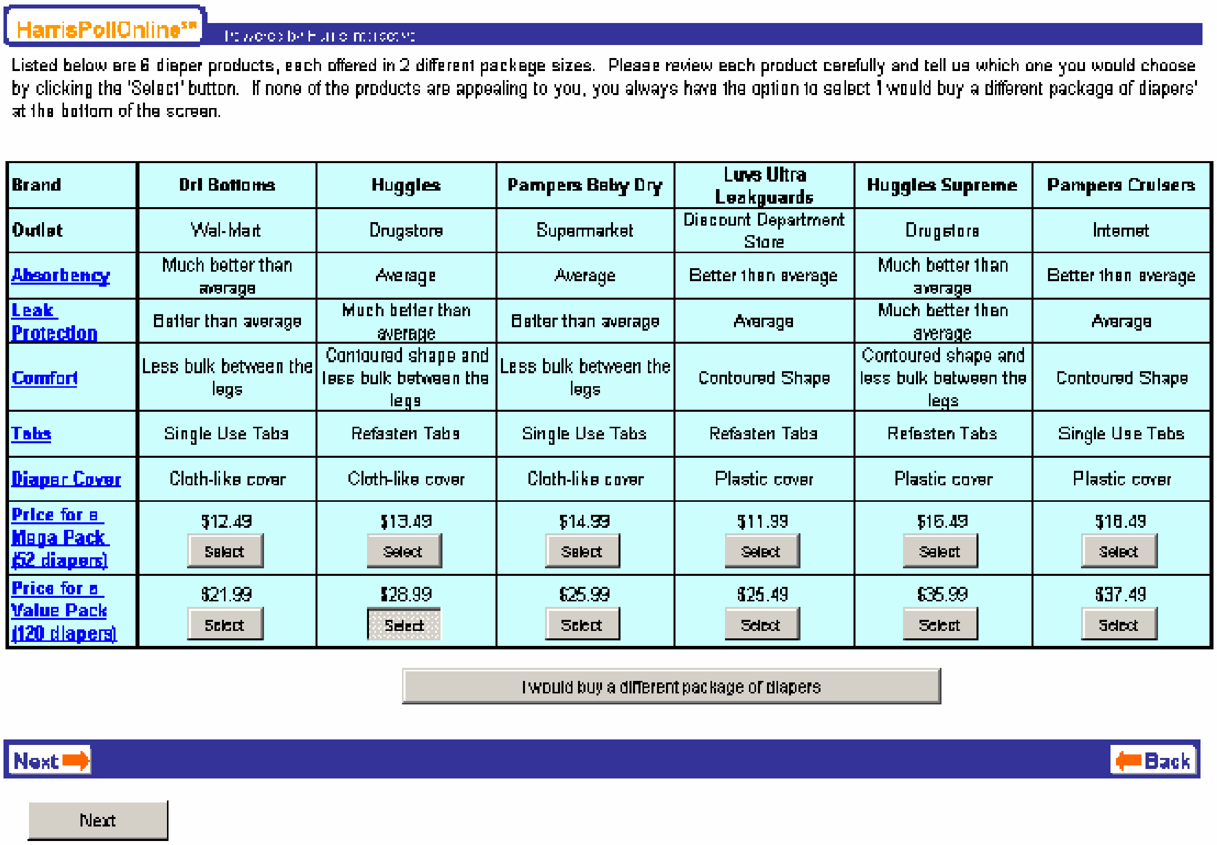
Mixed bundling is a strategy that can maximize profit under these circumstances by combining *exclusion* and *inclusion.* With mixed bundling, the price of the bundled options can be set so that all components are priced above their marginal cost. In addition, offering the components as individual items (again, at a price above marginal cost) captures purchasing from individuals for whom the bundle price exceeds the sum of the reservation prices for the components.

We can, of course, look at bundling from the *consumer's* perspective. While bundling satisfies the consumer's desire to pay no more than his or her reservation price for each of the component benefits, bundling may offer additional benefits as well. For example, bundling may simplify the consumer decision process. This is particularly likely to be true when the bundled features are part of a complex product solution and some degree of knowledge or expertise is required to assemble the right components. Thus, a consumer may be more likely to buy a personal computer bundled with a printer and digital camera because of the perception that the bundled components will be compatible with one another, reducing the risk and search effort of choosing the components separately.

MODELING CHOICES FOR MIXED BUNDLING

Standard choice-based modeling approaches may not capture the choice process for mixed bundling strategies. In a typical choice-based conjoint exercise, each alternative represents a fixed bundle, as illustrated in the example found in Figure 1. Respondents evaluate the different bundles and make a selection. Utilities for each of the components are estimated (typically in an additive, compensatory model) so that the total utility for any bundle constructed from the component features can be determined. However, we cannot use the component utilities to estimate the probability of choosing a particular feature by itself unless those choices were presented in the experiment. This also makes it difficult to determine the reservation prices for the individual components of the bundles and to estimate the impact of mixed bundling on total potential sales.

Figure 1
Typical Choice Task

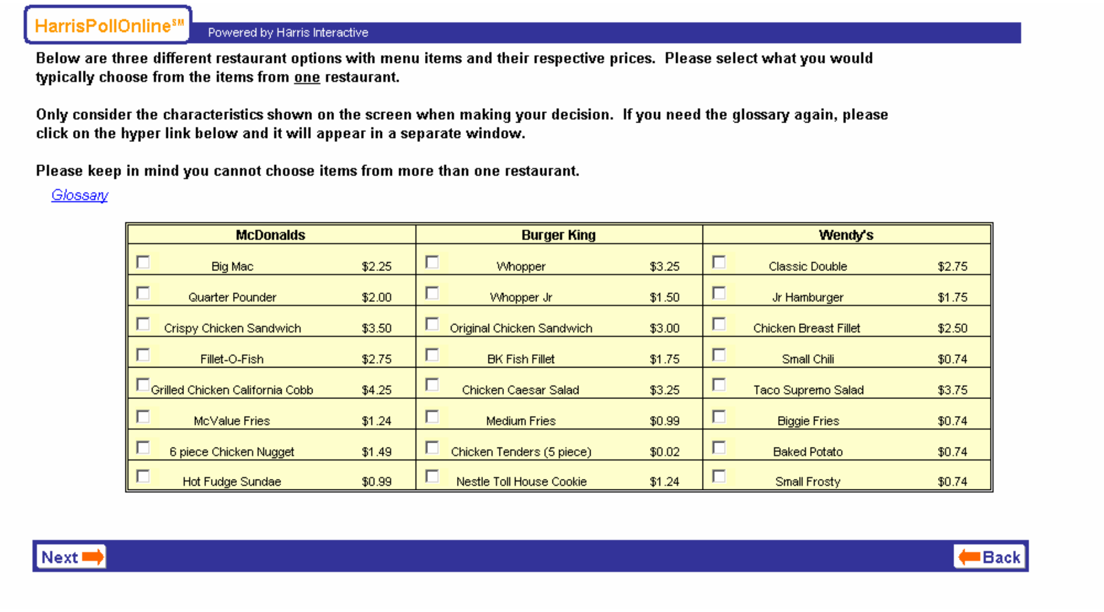


Listed below are 6 diaper products, each offered in 2 different package sizes. Please review each product carefully and tell us which one you would choose by clicking the 'Select' button. If none of the products are appealing to you, you always have the option to select 'I would buy a different package of diapers' at the bottom of the screen.

Brand	Dri Bottoms	Huggles	Pampers Baby Dry	Luvs Ultra Leakguards	Huggles Supreme	Pampers Cruisers
Outlet	Wal-Mart	Drugstore	Supermarket	Discount Department Store	Drugstore	Internet
Absorbency	Much better than average	Average	Average	Better than average	Much better than average	Better than average
Leak Protection	Better than average	Much better than average	Better than average	Average	Much better than average	Average
Comfort	Less bulk between the legs	Contoured shape and less bulk between the legs	Less bulk between the legs	Contoured Shape	Contoured shape and less bulk between the legs	Contoured Shape
Tabs	Single Use Tabs	Refasten Tabs	Single Use Tabs	Refasten Tabs	Refasten Tabs	Single Use Tabs
Diaper Cover	Cloth-like cover	Cloth-like cover	Cloth-like cover	Plastic cover	Plastic cover	Plastic cover
Price for a Mega Pack (52 diapers)	$12.49 Select	$13.49 Select	$14.99 Select	$11.99 Select	$16.49 Select	$18.49 Select
Price for a Value Pack (120 diapers)	$21.99 Select	$28.99 Select	$25.99 Select	$25.49 Select	$35.99 Select	$37.49 Select

As an alternative, we might construct choice tasks where the respondent selects individual components from a *menu* of features. Such a task is illustrated in Figure 2. In this case, we effectively allow the respondents to define their own bundles. A mixed bundling variation of this choice task is shown in Figure 3.

Figure 2
Menu Choice Task



Below are three different restaurant options with menu items and their respective prices. Please select what you would typically choose from the items from one restaurant.

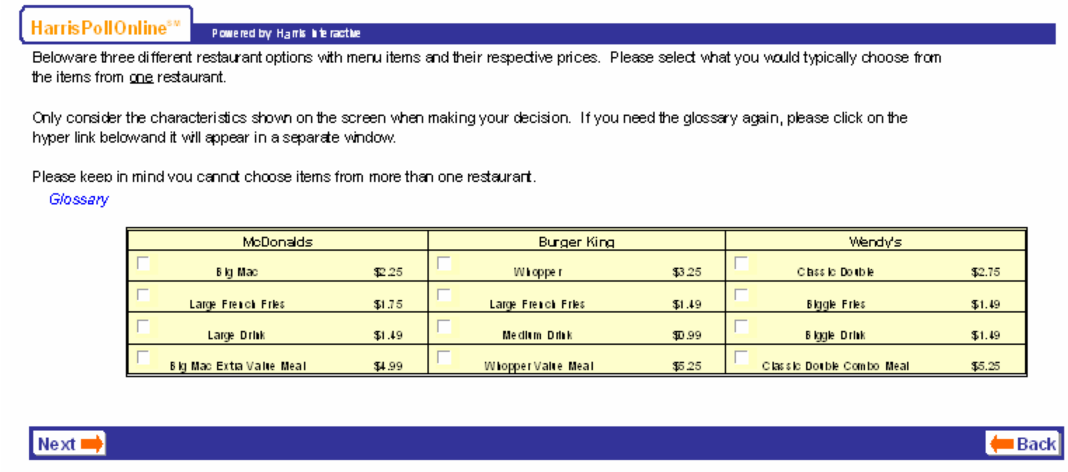
Only consider the characteristics shown on the screen when making your decision. If you need the glossary again, please click on the hyper link below and it will appear in a separate window.

Please keep in mind you cannot choose items from more than one restaurant.

Glossary

McDonalds			Burger King			Wendy's		
☐	Big Mac	$2.25	☐	Whopper	$3.25	☐	Classic Double	$2.75
☐	Quarter Pounder	$2.00	☐	Whopper Jr	$1.50	☐	Jr Hamburger	$1.75
☐	Crispy Chicken Sandwich	$3.50	☐	Original Chicken Sandwich	$3.00	☐	Chicken Breast Fillet	$2.50
☐	Fillet-O-Fish	$2.75	☐	BK Fish Fillet	$1.75	☐	Small Chili	$0.74
☐	Grilled Chicken California Cobb	$4.25	☐	Chicken Caesar Salad	$3.25	☐	Taco Supremo Salad	$3.75
☐	McValue Fries	$1.24	☐	Medium Fries	$0.99	☐	Biggie Fries	$0.74
☐	6 piece Chicken Nugget	$1.49	☐	Chicken Tenders (5 piece)	$0.02	☐	Baked Potato	$0.74
☐	Hot Fudge Sundae	$0.99	☐	Nestle Toll House Cookie	$1.24	☐	Small Frosty	$0.74

Figure 3
Mixed Bundle Choice Task



Beloware three different restaurant options with menu items and their respective prices. Please select what you would typically choose from the items from one restaurant.

Only consider the characteristics shown on the screen when making your decision. If you need the glossary again, please click on the hyper link belowand it will appear in a separate window.

Please keep in mind you cannot choose items from more than one restaurant.

Glossary

McDonalds			Burger King			Wendy's		
☐	Big Mac	$2.25	☐	Whopper	$3.25	☐	Classic Double	$2.75
☐	Large French Fries	$1.75	☐	Large French Fries	$1.49	☐	Biggie Fries	$1.49
☐	Large Drink	$1.49	☐	Medium Drink	$0.99	☐	Biggie Drink	$1.49
☐	Big Mac Extra Value Meal	$4.99	☐	Whopper Value Meal	$5.25	☐	Classic Double Combo Meal	$5.25

Both of these alternatives present challenges. For one thing, menu-based approaches can generate a very large alternative space. If each component is treated as a binary condition (included or not included), there are 2^N possible alternative bundles, where N = the total number of components that could be bundled. Additionally, combining predetermined bundles with individual component choices complicates the task for respondents. In Figure 3, for example, respondents can choose either the bundle or one or more individual items, but not a combination of the bundle and individual items (although this may be a realistic choice for some occasions, and we can design the task to allow choice of a bundle plus optional features not included in any bundle, for example).

At least three approaches for modeling *menu*-based choices have been described. Ben-Akiva and Gershenfield (1998) presented consumers with choices that included predetermined bundles and *a la carte* options. Each combination of chosen features is treated as a separate alternative in an aggregate multinomial logit model. For example, for telecommunications services, we might have choices that offer local service alone or bundled with other services (e.g., long distance, internet access, voice mail), along with *a la carte* options, so that a consumer might choose local and long distance service from a single provider, and internet access from a different company, and so forth. As noted, each combination represents a different choice alternative, with the included options comprising the "attributes" or predictors in the model.

Liechty, Ramaswamy, and Cohen (2001) took a different approach to the same basic problem. Rather than treat each possible combination as an alternative in the choice model, resulting in a large (and sparsely populated) multinomial data set, they treat each option in a choice task as a binomial choice. The probability of choice is a function of each option's intrinsic utility, plus option-specific covariates (e.g., price) and option-covariate cross-effects (e.g., the prices of other options on the menu). Respondents evaluate multiple menus (similar to Figure 3) in a stated preference experiment where the covariates of the options vary across menus).

Both of these approaches rely on repeated measures of consumer preference. In contrast, a "design your own product" approach, where consumers select options from a single menu (Bakken and Bremer, 2003) may elicit only one observation per respondent for each optional feature or component. In this model, component level price sensitivity was estimated using *relative* price (the incremental price of the feature divided by the total price paid for the configured product). This measure introduced component-level variability in the price. Appeal ratings for each component were used as the basis for the prior distribution of the intrinsic utilities for each component.

We take a different approach to modeling menu-based choices in this paper. We conceive of the mixed-bundle decision process as consisting of two steps. The buyer first chooses any one of the bundled offers or rejects all bundles. If all of the bundles are rejected, the buyer then chooses one or more of the components from an *a la carte* menu. The probability of choosing a bundle is a function of attributes of the bundle components as well as any unique bundle attributes, as the following equation illustrates:

$$\text{Prob(Bundle)}=f(\text{Att}_{\text{Brand}}+\text{Att}_{\text{Products}}+\text{Price}_{\text{Bundle}}+\text{Price}_{\text{Products}})$$

Where:

Att_{Brand} = the brand level attributes of the bundle

$Att_{Products}$=the attributes of the individual products comprising the bundle

$Price_{Bundle}$=the price of the bundle

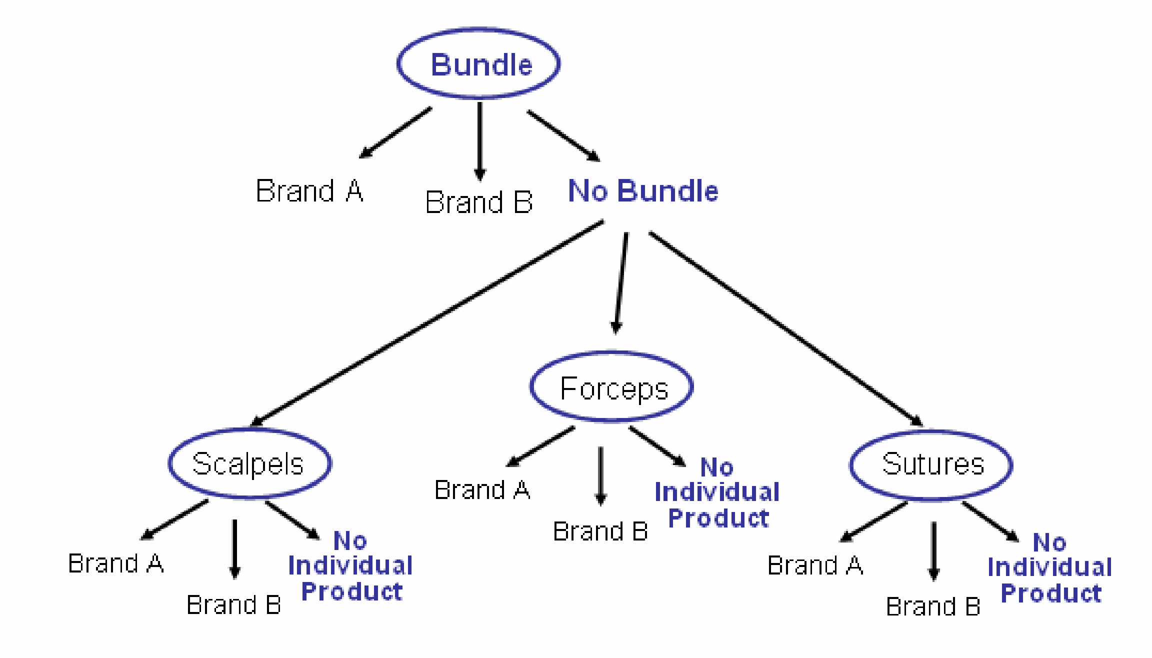
$Price_{Products}$=the prices of the individual products if purchased separately.

Similarly, the probability of choosing each individual product is a function of the attributes of the product choice, the brand attributes (shared by all products from the same brand), the price of the product, and the price of the bundle:

$$\text{Prob(Product)}= f(\text{Att}_{\text{Brand}}+\text{Att}_{\text{Product}}+\text{Price}_{\text{Bundle}}+\text{Price}_{\text{Product}}).$$

We estimate models for bundle choice and for choice of each of the individual products, as illustrated in Figure 4. In practice, respondents are presented with choice tasks in which they can choose bundles of products from different brands or each individual product from any of the competing brands.

Figure 4
Model Framework



Case Study

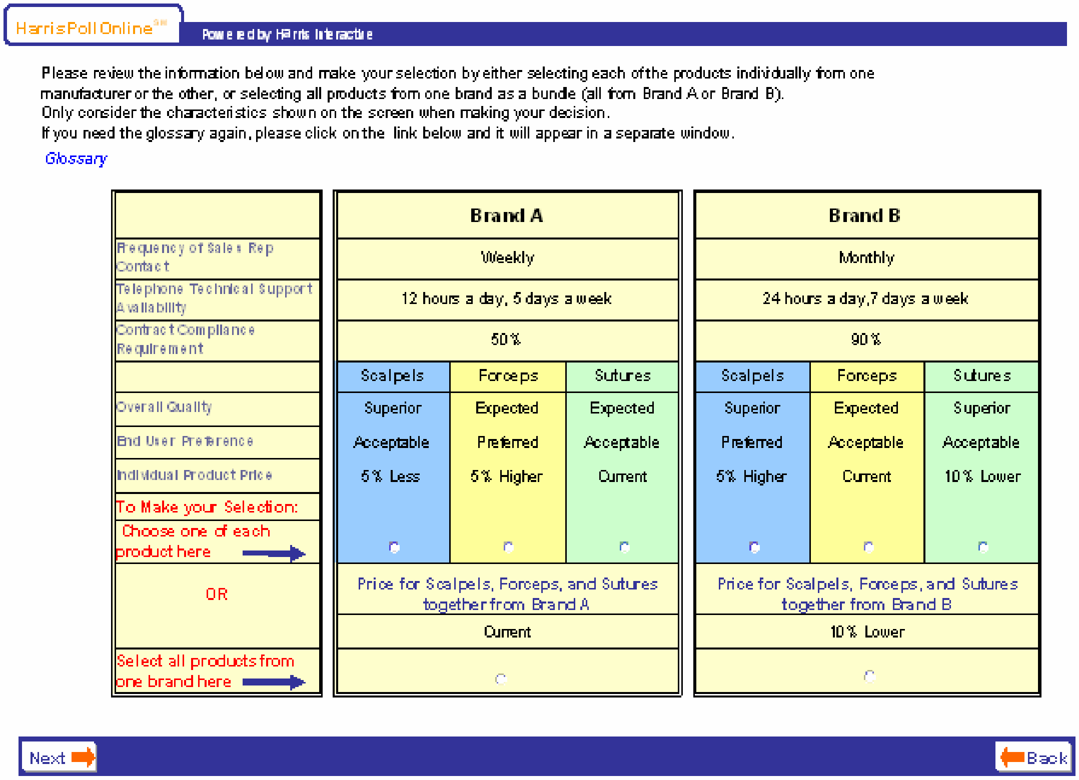
Our case study (disguised) involves the marketing of sutures, scalpels, and forceps to hospitals and stand-alone surgical centers. There are two dominant suppliers (brands) for these products, and, in many cases, the suppliers seek to establish contracts with their customers that will result in high utilization of their products. Contracts usually have discounts linked to specific compliance levels (e.g., 75% of eligible products purchased under the contract). Buyers are free to buy "off contract."

We replicated these market conditions by constructing choice tasks where respondents could choose to buy all three products from a single supplier (bundled options) or choose each product individually. In theory, customers will tend to choose individual products when there are perceived performance differences between brands in the individual products (for example, supplier A's scalpels are superior to supplier B's scalpels, and supplier B's sutures are superior to supplier A's) and these differences are not offset by the benefits (including potential savings) of choosing a bundled offer from either supplier.

Study Method

The survey was conducted over the Internet; 227 interviews were completed with qualified decision makers for the three products. As part of the development of the survey, we conducted a "think aloud" pre-test. A small number of individuals who matched the screening criteria for the study were invited to a central location to take the survey and "think aloud" as they did. We used this concurrent protocol method to evaluate different ways of presenting the choice tasks to the respondents. A sample from the final design is shown in Figure 5 below.

Figure 5
Case Study Sample Choice Task



Please review the information below and make your selection by either selecting each of the products individually from one manufacturer or the other, or selecting all products from one brand as a bundle (all from Brand A or Brand B).
Only consider the characteristics shown on the screen when making your decision.
If you need the glossary again, please click on the link below and it will appear in a separate window.

Glossary

	Brand A			Brand B		
Frequency of Sales Rep Contact	Weekly			Monthly		
Telephone Technical Support Availability	12 hours a day, 5 days a week			24 hours a day,7 days a week		
Contract Compliance Requirement	50%			90%		
	Scalpels	Forceps	Sutures	Scalpels	Forceps	Sutures
Overall Quality	Superior	Expected	Expected	Superior	Expected	Superior
End User Preference	Acceptable	Preferred	Acceptable	Preferred	Acceptable	Acceptable
Individual Product Price	5% Less	5% Higher	Current	5% Higher	Current	10% Lower
To Make your Selection:						
Choose one of each product here ➡	○	○	○	○	○	○
OR	Price for Scalpels, Forceps, and Sutures together from Brand A			Price for Scalpels, Forceps, and Sutures together from Brand B		
	Current			10% Lower		
Select all products from one brand here ➡	○			○		

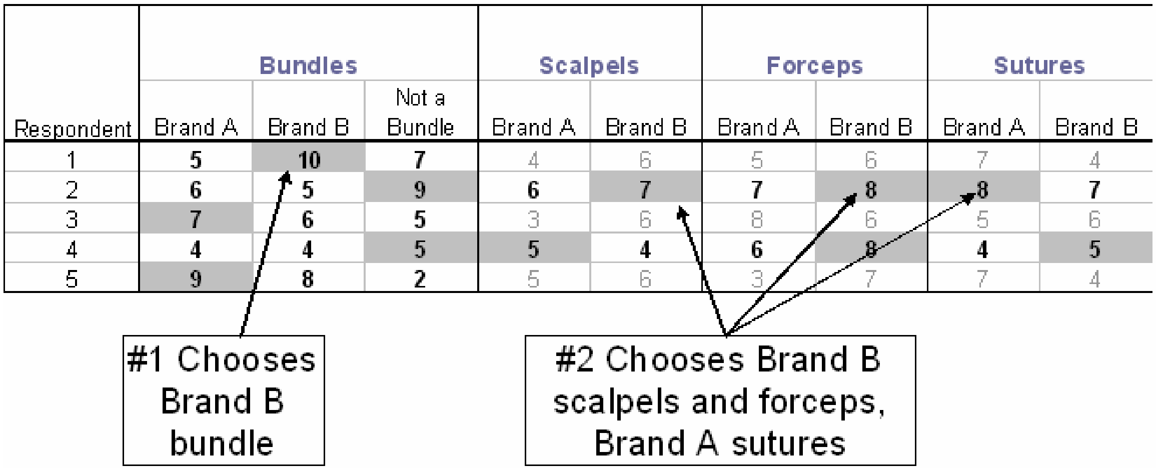
Model Estimation

The four choice models (one for bundles and one each for the individual products) were estimated using Sawtooth Software's CBC-HB. Although there was not a "none" option in the choice task, the models were estimated with a "none" parameter in order to create the linkages between bundled and individual product choices. For example, if a respondent did not choose a bundle, the response for that choice for the bundle model was coded as "none." Similarly, if the respondent chose a bundle, the response for each of the individual products was coded as "none."

Combining the Models

The models were integrated by means of a market simulator. The simulator was constructed so that the utility for each specified bundle was calculated for each respondent and compared to that respondent's utility for not choosing a bundle (i.e., the none parameter). In a first choice model (with a random component for breaking ties), the respondent was counted as choosing either bundle A or bundle B or, if the "none" utility was greater, each of the individual products with the highest utility. This is illustrated in Figure 6.

Figure 6
Respondent-Level Simulator Example



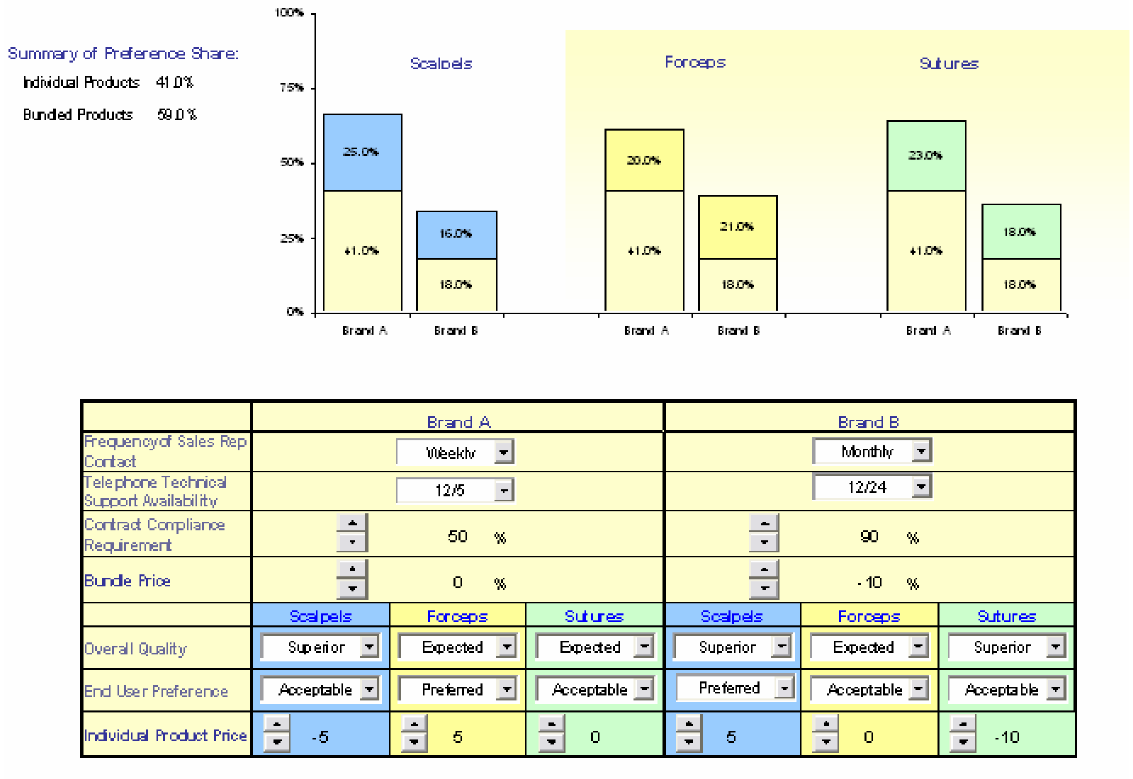
	Bundles			Scalpels		Forceps		Sutures	
Respondent	Brand A	Brand B	Not a Bundle	Brand A	Brand B	Brand A	Brand B	Brand A	Brand B
1	5	10	7	4	6	5	6	7	4
2	6	5	9	6	7	7	8	8	7
3	7	6	5	3	6	8	6	5	6
4	4	4	5	5	4	6	8	4	5
5	9	8	2	5	6	3	7	7	4

#1 Chooses Brand B bundle

#2 Chooses Brand B scalpels and forceps, Brand A sutures

The market simulator interface emulates the choice task. The user specifies the levels of the attributes for the bundles as well as the individual products. In the example presented in Figure 7, results show, for each product, the share chosen as part of a bundle for each brand and the share chosen individually.

Figure 7
Market Simulator Interface



Summary of Preference Share:
Individual Products 41.0%
Bunded Products 59.0%

	Brand A			Brand B		
Frequencyof Sales Rep Contact	Weekly			Monthly		
Telephone Technical Support Availability	12/5			12/24		
Contract Compliance Requirement	50 %			90 %		
Bundle Price	0 %			-10 %		
	Scalpels	Forceps	Sutures	Scalpels	Forceps	Sutures
Overall Quality	Superior	Expected	Expected	Superior	Expected	Superior
End User Preference	Acceptable	Preferred	Acceptable	Preferred	Acceptable	Acceptable
Individual Product Price	-5	5	0	5	0	-10

Discussion

The case study demonstrates the feasibility of combining separate models via the market simulator to capture relatively complex decision processes such as those that might occur with a mixed bundling strategy. Although we combined the bundled and individual product choices into a single choice task, we believe that this approach could be used in a two step process as well, where respondents first evaluate different bundles, then evaluate individual products. This might be somewhat less burdensome on respondents.

Given the complexity of the choice tasks for mixed bundling scenarios, we think it is important to pre-test the survey thoroughly. The think-aloud process we employed appears to be particularly effective for uncovering survey problems that might otherwise compromise the quality of the data obtained.

The modeling for our approach is more time consuming than with single product choice discrete choice models. In this case we estimated four separate models. With a longer list of individual products in the menu, that number would increase. The linkage between the models is purely through the market simulator, not model estimation. Because of the way we linked the models (via the "none" parameter for the bundled model), we did not need to be concerned about scale differences between the models. In effect, the choice probability for any one alternative (bundle or individual product) is a function of the part-worth estimates in only one model. However, we caution that, with this approach, custom market simulators end up being relatively large with fairly complex behind the scenes workings.

References

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

Ben-Akiva, M. and S. Gershenfeld, "Multi-Featured Products and Services: Analyzing Pricing and Bundling Strategies," Special Issue of the *Journal of Forecasting*., Vol. 17, Issue 3/4, June-July 1998.

Liechty, J., V. Ramaswamy and S.H. Cohen, "Choice Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand with an Application to a Web-based Information Service," *Journal of Marketing Research*, May, 2001.

Lilien, G.L., P. Kotler and K. S. Moorthy, *Marketing Models*, Prentice-Hall, 1992.

DISCRETE CHOICE DESIGN

THE IMPORTANCE OF SHELF PRESENTATION IN CHOICE-BASED CONJOINT STUDIES

GREG ROGERS
PROCTER & GAMBLE
TIM RENKEN
COULTER/RENKEN

INTRODUCTION

Market researchers have studied choice behaviour for many years and have created increasingly complex methods like choice-based conjoint. The development of choice models has been aided by advances in computer technology. One only has to think of the acceptance of hierarchical Bayes modelling to analyse choice-based conjoint data to see an example of how computing power has helped advance the discipline.

Computer technology has also provided an opportunity to improve the visual stimulus within a choice-based conjoint study. In the past decade many researchers have strived for greater realism in how they present a choice task to a respondent, believing that a more realistic interface would more likely result in the respondent behaving as they normally would when shopping. This is a reasonable hypothesis since the researcher wants the respondent to use the same heuristics in the test situation as in reality, and visual cues are known to be critical.

Burke (1992) compared a rudimentary text description with a shelf image interface for a choice model and found little difference between the methods when comparing choice share to market share. However, there was a difference in how the methods estimated sensitivity to price changes, with the shelf image interface resulting in greater sensitivity.

Beyond visual enhancements to the choice task, there have also been changes to experimental design by many researchers. Most researchers tended to limit the number of products shown in a choice task believing that respondents would be overloaded with information and not be able to complete the task responsibly if a full array of products were displayed. As store shelves have become more cluttered so too have choice tasks, with many researchers now opting for a full complement of products within a choice task.

This research compares two choice-based conjoint interfaces: One contains a simple grid of products from which consumers make their choices, and the other attempts to make the choice exercise more realistic by replicating a store shelf. We would like to determine whether the more realistic exercise yields superior results. We estimate choice models from data derived from the two interfaces, and we evaluate the interfaces on two dimensions: (1) how well the models' choice shares approximate unit market shares derived from scanner data, and (2) how well the models' price sensitivities approximate scanner data-based price sensitivity estimates. We use scanner data as a benchmark because managers in many packaged goods categories view scanner data as the gold standard—i.e. real people paying real money—measure of consumer behaviour.

We run our comparison on data for three categories—salted snacks, laundry detergent, and facial tissues—and find little evidence that the shelf exercise provides a better fit to unit market shares than the grid exercise; we do, however, find evidence that the shelf exercise more closely replicates price sensitivities.

Design and Data Collection

Figure 1 contains an example of the grid choice exercise. The grid contains pictures of the products, together with descriptions of the products and prices. We choose the most important products in the category in terms of unit shares as competitors in the exercise. A respondent sees the grid, and we tell her to choose the one item she would be most likely to purchase. We then show the respondent another grid and have her make another choice. The prices and locations of the products in the grid change from choice to choice. Note that respondents can also choose "None" if none of the products in the grid are acceptable.

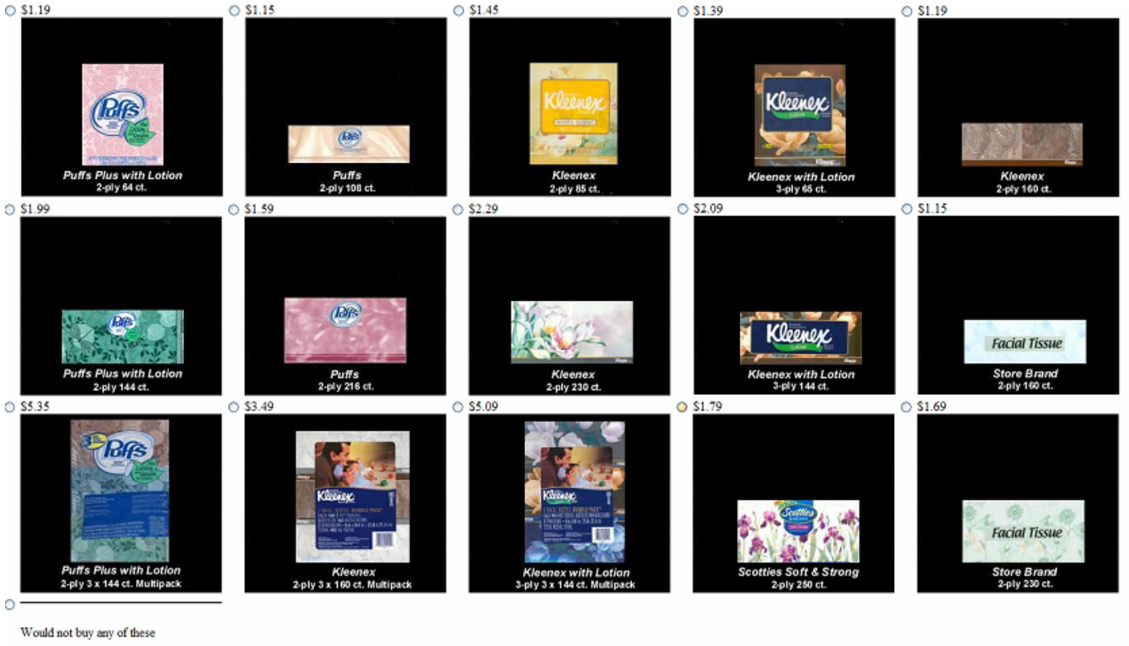


Figure 1: Grid Choice Exercise Layout

Figure 2 contains an example of the shelf choice exercise. We use planograms to make the shelf as much like a shelf in an actual store as possible, though limitations in the speed of dial-up modems prevent us from showing all of the products that would appear on an actual supermarket or mass merchant shelf. A respondent sees the shelf, and we tell her to choose the one item she would be most likely to purchase. We then show the respondent another shelf and have her make another choice. The prices of products change from shelf to shelf, but the locations of products remain the same. As with the grid exercise, respondents can choose "None" if none of the products on the shelf are acceptable.

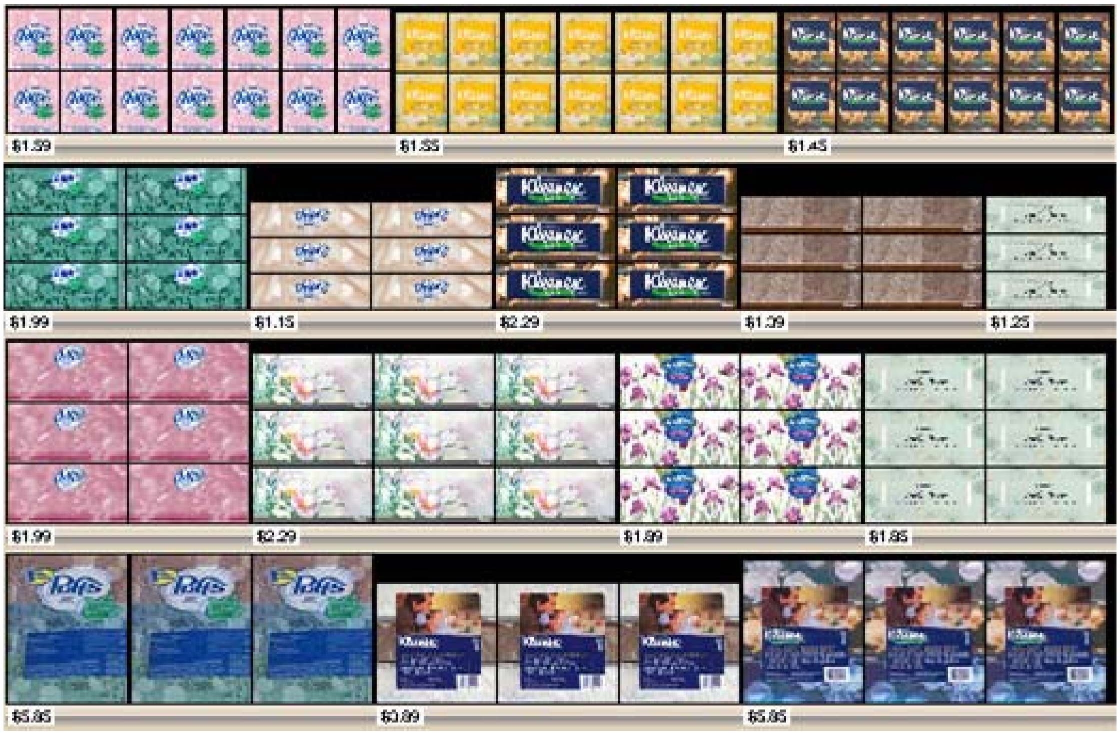


Figure 2: Shelf Choice Exercise Layout

We examine three categories of fast-moving consumer goods in this study: salted snacks, laundry detergents, and facial tissues. *Table 1* describes how the conjoint and scanner data were collected and how many products were used in the market share vs. choice share and price sensitivity comparison. As the table indicates, the conjoint data were collected over the Internet using consumers from Internet panels. In the conjoint studies we tested each product at five price points, and the price change percentages we tested are typically those of interest to a manager. We determined the base price (+0% price point in the table) using average past 52 week non-promoted price information data for each product derived from supermarket scanner data. The experimental design contains four blocks (choice exercise versions) of six choice sets per block, and the design was generated using the SAS PROC FACTEX and PROC OPTEX procedures. We employed this relatively fractionated design because we wanted to limit the number of choices respondents had to make to only six. We have found that grid and shelf exercises demand a lot of respondents, particularly over the Internet, and we wanted to minimize the negative impact of respondent fatigue on our data.

The scanner data were collected by Information Resources Incorporated, and we use their marketing mix model as the benchmark in our analysis below. The data collection dates for the scanner data differ from those of the conjoint data, a potential limitation of our analysis. Since our test analyzes well-established products in well-established categories, however, our expectation is that the price sensitivities of these products probably do not change much from one year to the next.

	SALTED SNACKS		LAUNDRY DETERGENT		FACIAL TISSUES	
	Grid	**Shelf**	**Grid**	**Shelf**	**Grid**	**Shelf**
Conjoint data						
Data collection method	Internet	Internet	Internet	Internet	Internet	Internet
Base size	1158	578	869	564	685	515
# of products in test	15	30	30	30	15	15
Price points tested	-15%, -8%, 0%, 8%, 15%		-15%, -5%, 0%, 5%, 15%		-15%, -5%, 0%, 5%, 15%	
# of questionnaire versions	4	4	4	4	4	4
# choice tasks/respondent	6	6	6	6	6	6
Data collection dates	6/02	6/02	6/02	6/02	6/02	6/02
Scanner data						
Time period	52 Weeks Ending 8/01		52 Weeks Ending 8/01		52 Weeks Ending 8/01	
Coverage	Food + Mass		Food + Mass		Food + Mass	
Number of products for Choice Share vs. Market Share Comparison	14		29		15	
Number of products for Price Sensitivity Comparison	5		7		5	

Table 1: The Data

Scanner Data Model

IRI's Marketing Mix methodology uses a multiplicative model estimated on weekly store-level scanner data. The model takes the form:

$$s_A = \beta_0 \overline{p}_A^{\beta_1} \left(\frac{\widehat{p}_A}{\overline{p}_A} \right)^{\beta_2} \ldots e^{(\beta_3 \text{Feat}_A + \beta_4 \text{Disp}_A + \ldots + \varepsilon_A)} \qquad (1)$$

where:

- s_A = volume sales of product *A* in a given store week
- $\overline{p}_A$ = base (or shelf) price for product *A*
- $\widehat{p}_A$ = promoted price for product *A*. When there is no promotion, $\widehat{p}_A = \overline{p}_A$.
- Feat_A = 1 if product *A* is on feature, 0 otherwise
- Disp_A = 1 if product *A* is on display, 0 otherwise
- $\beta_0, \ldots, \beta_4, \ldots$ = parameters
- ε_A = normally distributed error term.

Taking logs of both sides of Equation (1) enables estimation of the parameters $\beta_0,\ldots,\beta_4,\ldots$ with multiple regression.[1] IRI incorporates a number of variables in its analysis to model weekly sales—not all of which are enumerated in Equation (1) above—but our focus is the parameter β_1 that describes the impact of shelf price changes on volume. $\beta_1 = (\overline{p}_A / s_A)(\partial s_A / \partial \overline{p}_A)$ is the own-price elasticity for this demand function for shelf price changes.

Conjoint Model

The conjoint model assumes a utility function of the form

$$U_{A,h} = \alpha_{Ah} + \lambda_h p_A + \varepsilon_{A,h} \tag{2}$$

where

- $U_{A,h}$ = the total utility of product *A* for respondent *h*,
- p_A = the price of product A (in dollars),
- α_{Ah}, λ_h = parameters,
- $\varepsilon_{A,h}$ = an extreme value (Gumbel)-distributed error term.

We estimate the parameters α_{Ah}, λ_h using a hierarchical Bayes choice modelling algorithm (see Allenby and Ginter (1995)) and calculate the probability that respondent *h* will choose product *A* using the standard logit formula

$$P_h(A) = \frac{e^{U_{A,h}}}{\sum_B e^{U_{B,h}}} \tag{3}$$

While there are a variety of ways we could have specified the choice model's utility function in Equation (2), we chose the specification in Equation (2) because of its simplicity. Our goal in this paper is to compare discrete choice stimulus materials and not models.

Results

The comparison of choice share to market share is shown in *Table 2*. In order to make the fairest comparison possible, the choice shares were adjusted by multiplying the raw choice share from the model (as specified above) by the all commodity distribution for that item. All items were then re-percentaged to ensure they summed to 100% (including the 'none' option). Similarly the market share data based on unit volume was re-percentaged as the entire category was not represented in the test. Thus, an attempt

TP[1]PT Our description of the scanner data-based model is derived from IRI's 2003 "IRI Marketing Mix Technical White Paper."

was made to account for the relative distribution and transaction size differences that are known causes of mismatch between choice models and market data.

The key measure of comparison between the choice models and the market data is mean absolute error (MAE), as shown in *Table 2*. The results show that there is no significant difference (at 95% confidence) in MAE between the grid and shelf designs when compared to market share.

(re% unit share vs. re% choice share)	***Grid***	***Shelf***
	(%)	***(%)***
Salted Snacks	4.7	4.5
Laundry Detergent	1.1	1.3
Facial Tissue	2.2	3.4
Total (across 3 categories)	**2.3**	**2.6**

Table 2: Market Share vs. Choice Share MAE Comparison

We also looked at the range of choice shares estimated from the grid and shelf designs, as sometimes averaging data can mask differences in sensitivity. The standard deviation, as shown in *Table 3*, indicates that the shelf design is more sensitive than the grid design. This is an important finding in that it shows the shelf method has more potential to show differences, even though on average it did not prove to be more accurate than the grid design.

	Unit Share	*Grid Choice Share*	*Shelf Choice Share*
Salted Snacks	0.0403	0.0500	0.0325
Laundry Detergent	0.0222	0.0250	0.0300
Facial Tissue	0.0439	0.0352	0.0428
Total (across 3 categories)	**0.0511**	**0.0390**	**0.0465**

Table 3: Market Share vs. Choice Share Standard Deviation Comparison

The comparison of price sensitivity estimates from the various methods is shown in *Table 4*. The mean error shows that the estimates are centred close to zero. This is an interesting result in itself as it provides evidence that neither method is particularly biased in providing particularly high or low estimates of price sensitivity when compared with econometric models.

The mean absolute percent error (MAPE) shows that the shelf design is somewhat closer to the econometric prediction, though neither the shelf or grid designs are very close to the price sensitivity estimates from the econometric models. However, it should be noted that neither choice model was tuned, or adjusted, in a way that would likely minimise the error. The main focus here is the comparison of the choice-based conjoint designs to each other, using the default scaling of part worths from HB estimation.

	15% price increase		15% price decrease	
	Grid	Shelf	Grid	Shelf
Mean Error	0%	-1%	-1%	0%
Mean Absolute Error (MAE)	9%	8%	15%	11%
Mean Percent Error	35%	30%	38%	28%
Mean Absolute Percent Error (MAPE)	79%	67%	90%	65%

Table 4: Price Sensitivity Comparison

For researchers that were fielding choice-based conjoint studies in the mid 90's even the grid layout may be seen as a high level of finish. It would be interesting to extend this work to include text only designs. However, even though there may be little difference, it is still worth the effort (often considerable) to construct shelf layout designs as there is a greater degree of face validity attached to the technique by marketing managers. This is even more important for methods like choice-based conjoint, while they are often practiced, have relatively little to offer in the way of in-market validation.

CONCLUSIONS

Our first hypothesis was largely disproved by this data set. We hypothesized that better shelf depiction in a choice-based conjoint task would lead to a closer match of choice share to market share. While the shelf depictions can provide more variance in choice shares, they were not any better predictors of unit share.

The second hypothesis we tested was that shelf depiction in a choice-based conjoint task would lead to a closer match of estimated price sensitivity when compared with econometric models. This hypothesis was proved to be correct as the shelf layout prediction of price sensitivity was closer to econometric models, particularly for price decreases.

REFERENCES

Allenby, Greg and James Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32 (November), 392-403.

Burke, Raymond M., Harlam, Bari A., Kahn, Barbara E. & Lodish, Leonard M. (1992). "Comparing Dynamic Consumer Choice in Real and Computer-simulated Environments," *Journal of Consumer Research*, 19, June.

Information Resources Incorporated (2003) "IRI Marketing Mix Technical White Paper."

Feurstein, M., Natter, M. & Kehl, L. (1999). "Forecasting scanner data by choice-based conjoint models". In Proceedings of the Sawtooth Software Conference (No. 7, pp. 169-182). Sequim, WA: Sawtooth Software.

COMMENTS ON ROGERS AND RENKEN

ADRIAN VICKERS AND PHIL MELLOR
ADELPHI INTERNATIONAL RESEARCH

One of the most important aspects of conjoint studies is being able to create a realistic setting. Louviere (2000) states that the closer the experiment resembles the real market, the higher the face validity. Rogers and Renken (2004) address this issue by comparing a grid based approach with the more realistic setting of depicting what a shelf might actually look like. Steps were also undertaken to make the market share prediction more realistic by weighting the share of preference from the grid experiment with the distribution levels and also by making the shelf space of each product proportional to this as well. This is in accordance with Arenoe (2003). The authors found that the shelf exercise does not necessarily provide a better fit to scanner data than the grid exercise. However, there were two important findings:

1. there is greater variance from the shelf exercise
2. the shelf exercise provides closer estimates of price sensitivity to scanner data

This suggests it may be harder for respondents to comprehend the information on the shelf. This is perhaps not surprising when the screen shots are viewed. However, this is much closer to what consumers see when they walk into a shop and so should therefore be a more realistic setting which in theory will produce better utility estimates.

The ability to depict shelves in conjoint tasks is a very recent introduction to the available software. Rogers and Renken's paper helps market researchers to make a more informed decision of which approach to use and what bias there may be in using either technique. However, an important aspect of this paper was that the improvement in the predicted shares was very small considering the large difference between what their models predicted and real market share. Wittink (2003) argues that for fixed consumption levels, such as detergents, proper measurements of household-level purchases or consumption should improve the predictive validity of CBC results. Wittink (2003) also states that for categories with expandable consumption levels it may be useful to have separate approaches that allow users to predict purchase incidence and volume as a function of price. If these procedures had been followed then a comparison with the real market share data might have been more informative. Although Orme (1996) does list a number of other reasons why conjoint analysis might not reflect real market share particularly well. Wittink (2003) also argues against the use of scanner data to assess price effects (because regular prices vary little over time) and that for an analysis of regular price effects, it should be sufficient to focus exclusively on conditional brand choice in a CBC study.

In order to assess the predictive accuracy of conjoint models based on depicting shelf layout versus a grid design a more controlled test is probably needed. This could be achieved by asking respondents to do two conjoint exercises, one with the grid and the other the shelf. There might also be the possibility to ask far more conjoint tasks, than what are traditionally used in conjoint studies, so that reliable utility estimates, with very little error, can be generated for each respondent. The prediction from the conjoint output then needs to be compared with real purchase data for that respondent, whether in a controlled environment or from a shop.

Comparing the results of different types of methods and techniques is always a very interesting choice of subject and we should encourage other works on these types of comparisons. However, where the differences are likely to be very small between alternative approaches, perhaps more attention should be given to controlled experiments instead of trying to predict what the whole population will do.

REFERENCES

Arenoe, B. (2003). Determinants of External Validity in CBC. *Sawtooth Software Conference 2003 Proceedings*, 217-232.

Louviere, J.J., Hensher, D.A. and Swatt, J.D. (2000). *Stated Choice Methods: Analysis and Application.* Cambridge University Press.

Orme, B. (1996). *Helping Managers Understand the Value of Conjoint.* Sawtooth Software Research Paper Series. Sawtooth Software, Inc.

Rogers, G. and Renken, T.(2004). The Importance of Shelf Presentation in Choice-Based Conjoint Studies. *Presented at the eleventh Sawtooth Software Conference*.

Wittink, D.R. (2003). Comment on Arenoe and Rogers/Renken. *Sawtooth Software Conference 2003 Proceedings*, 233-236.

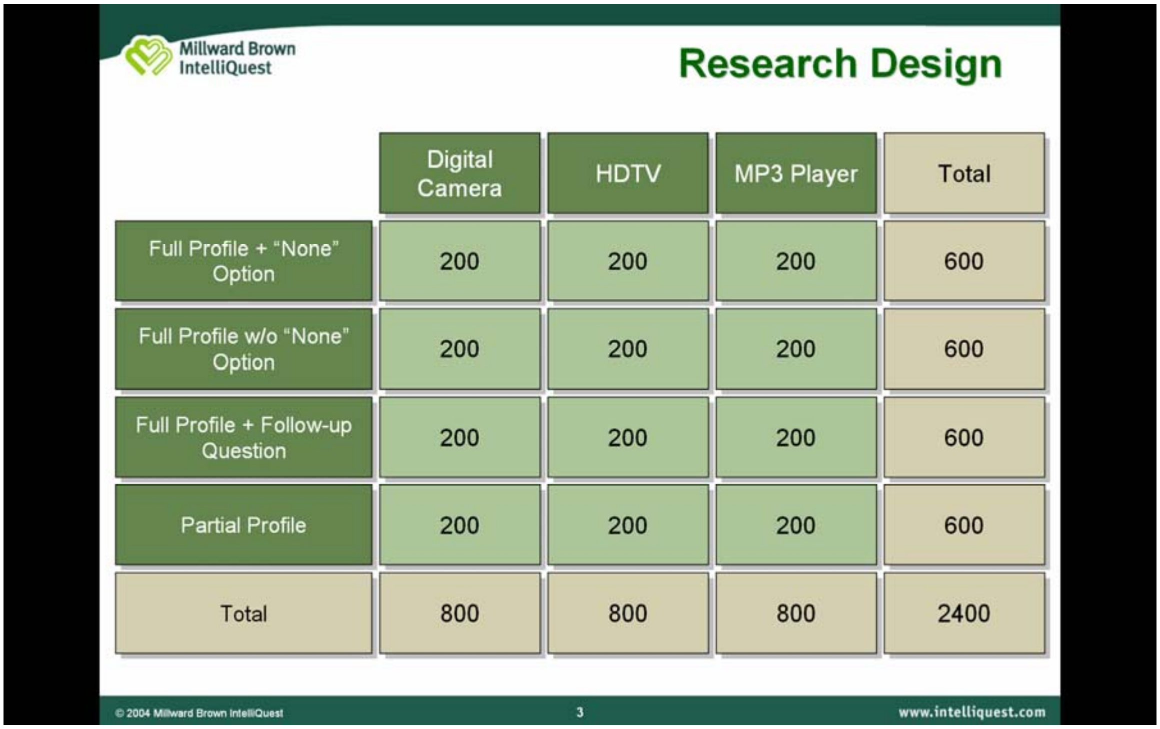
THE EFFECT OF DESIGN DECISIONS ON BUSINESS DECISION-MAKING

CURTIS FRAZIER, URSZULA JONES
MILLWARD BROWN-INTELLIQUEST

There has been a significant amount of research done on the statistical properties of various discrete choice approaches. These studies tend to devote the bulk of their attention to measures of accuracy, such as hit rates and mean absolute error. While these studies are exceptionally valuable, they are not complete. This is because these studies assume that a higher hit rate or lower MAE is sufficient information in determining the best approach for a project. We contend that this information must be augmented by an understanding of the impact of design decisions on the substantive outcomes of the research.

In order to assess the effects of design decisions on business decision-making, we have adopted an experimental cell design. The image below describes the fundamental design. We choose to study three separate consumer electronics products. For each of these products, respondents we randomly assigned to one of four variants of discrete choice (full profile with a "none of these" option, full profile without a "none of these" option, full profile with the "none of these" option asked as a follow-up question, and partial profile).

Figure 1



Millward Brown IntelliQuest

Research Design

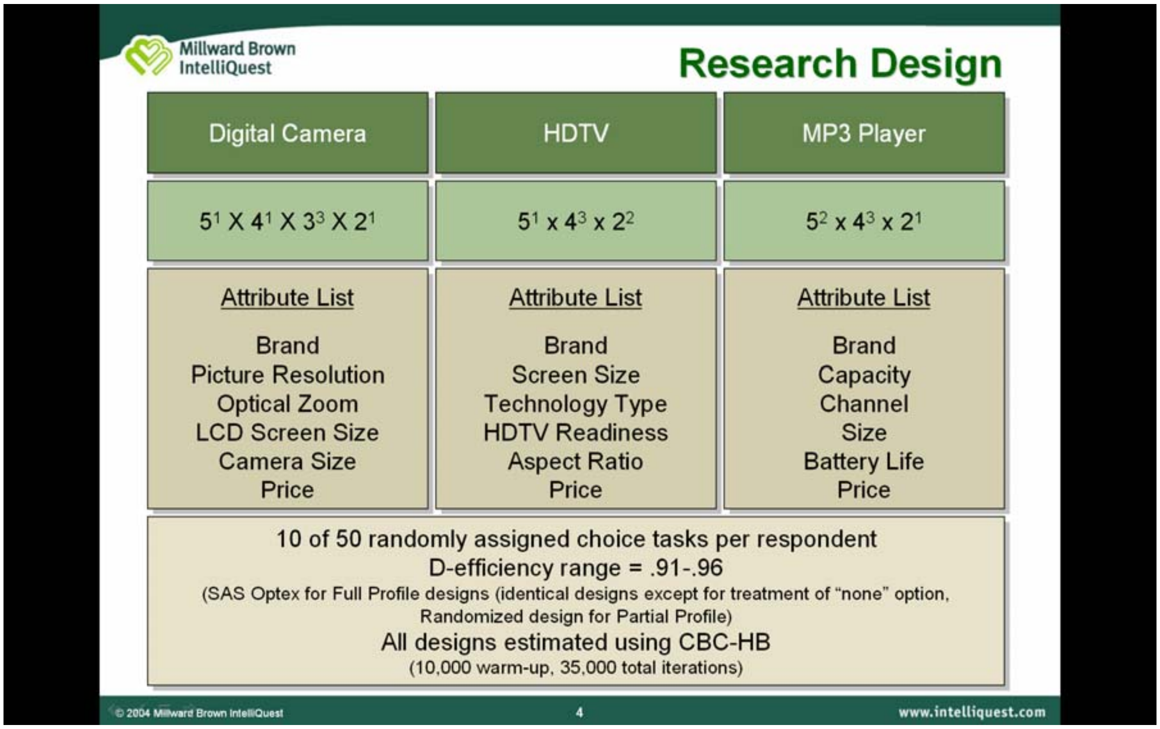
	Digital Camera	HDTV	MP3 Player	Total
Full Profile + "None" Option	200	200	200	600
Full Profile w/o "None" Option	200	200	200	600
Full Profile + Follow-up Question	200	200	200	600
Partial Profile	200	200	200	600
Total	800	800	800	2400

© 2004 Millward Brown IntelliQuest 3 www.intelliquest.com

Each of the four conjoint variants used identical attributes and levels. The three full profile variants used identical designs apart from the treatment of the "none of these" options.

Figure 2 illustrates the designs used for each of the three products.

Figure 2



Millward Brown IntelliQuest

Research Design

Digital Camera	HDTV	MP3 Player
5^{1} X 4^{1} X 3^{3} X 2^{1}	5^{1} x 4^{3} x 2^{2}	5^{2} x 4^{3} x 2^{1}
Attribute List Brand Picture Resolution Optical Zoom LCD Screen Size Camera Size Price	Attribute List Brand Screen Size Technology Type HDTV Readiness Aspect Ratio Price	Attribute List Brand Capacity Channel Size Battery Life Price

10 of 50 randomly assigned choice tasks per respondent
D-efficiency range = .91-.96
(SAS Optex for Full Profile designs (identical designs except for treatment of "none" option,
Randomized design for Partial Profile)
All designs estimated using CBC-HB
(10,000 warm-up, 35,000 total iterations)

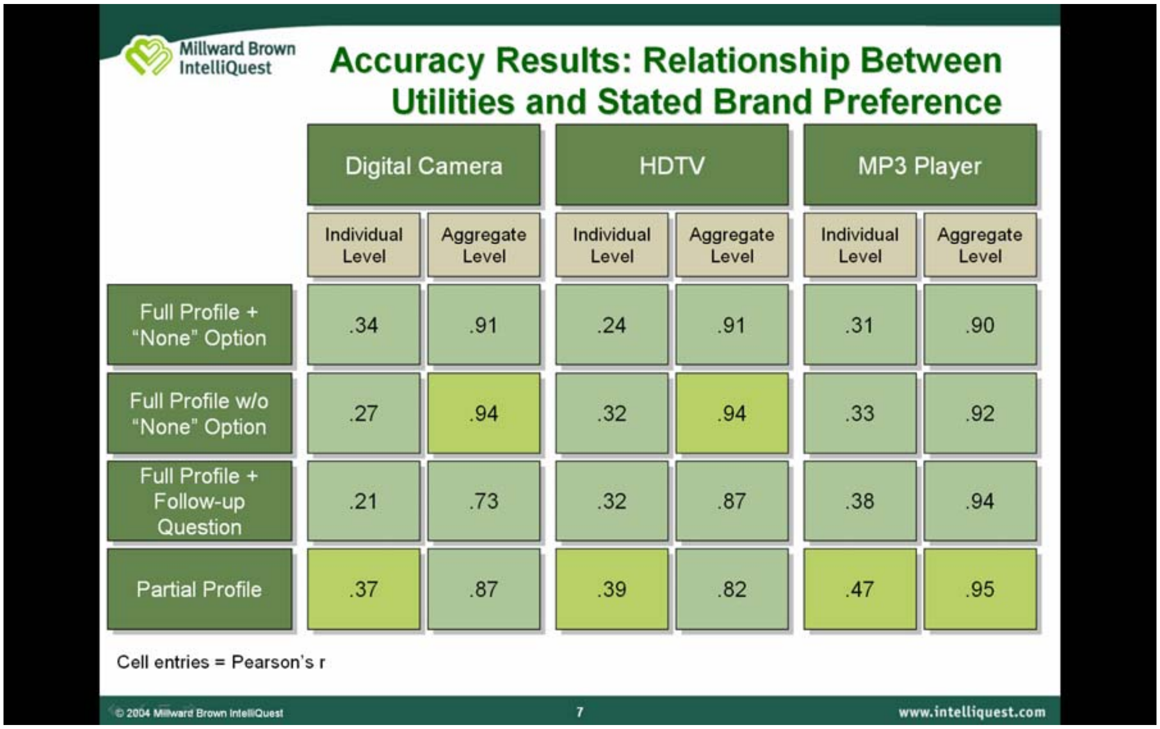
© 2004 Millward Brown IntelliQuest 4 www.intelliquest.com

As a first step, we ran the usual assessments of internal validity. The results were similar to previous studies, showing that partial profile performed better at the individual level (hit rate estimates) and full profile with the "none of these" option performed better at the aggregate level (MAE).

Evaluation of Substantive Results

We first decided to evaluate brand utilities. As a first step, we conducted a simple correlation between brand utilities and stated brand preference (Figure 3). A consistent picture emerged. The partial profile brand utilities were found to be more correlated to stated brand preferences than the other choice variants. We hypothesize that because less information is provided about the alternative products in the conjoint tasks, respondents would assign some additional variance to the partial profile brand utilities. For example, if the respondent perceives Sony to provide higher resolution pictures and picture resolution is not included in the choice task, we believe the respondent projects better picture quality onto the Sony alternative.

Figure 3



Millward Brown IntelliQuest

Accuracy Results: Relationship Between Utilities and Stated Brand Preference

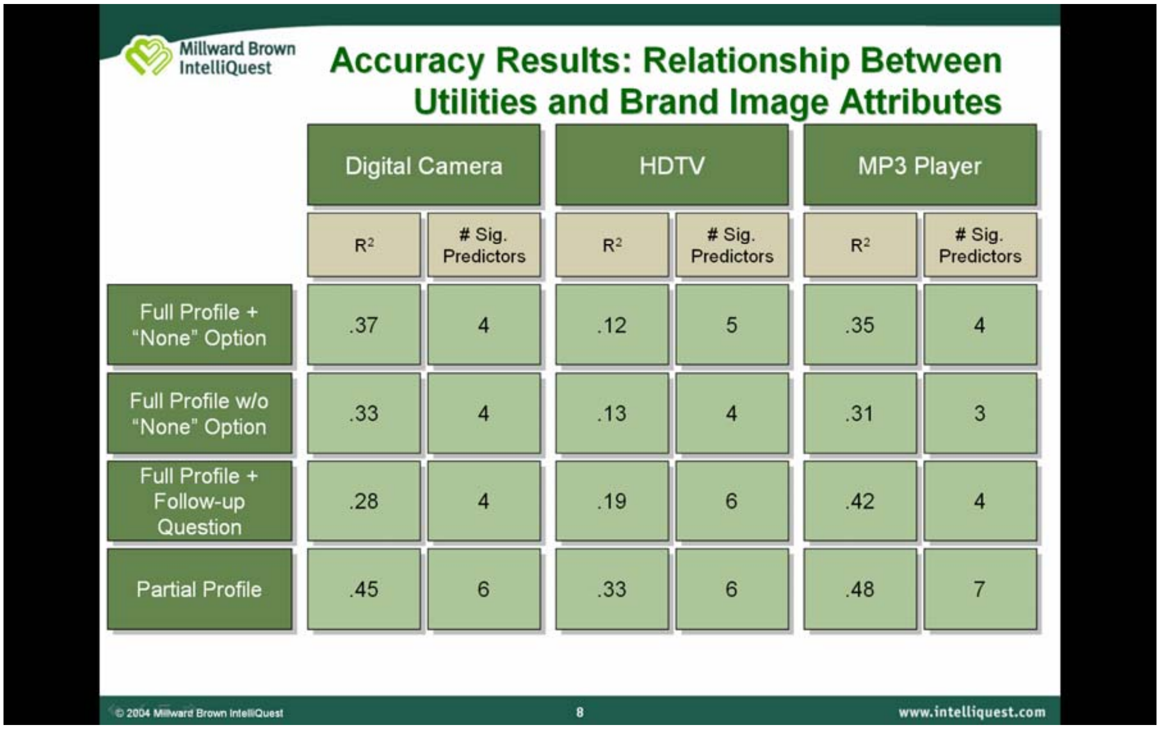
	Digital Camera		HDTV		MP3 Player	
	Individual Level	Aggregate Level	Individual Level	Aggregate Level	Individual Level	Aggregate Level
Full Profile + "None" Option	.34	.91	.24	.91	.31	.90
Full Profile w/o "None" Option	.27	.94	.32	.94	.33	.92
Full Profile + Follow-up Question	.21	.73	.32	.87	.38	.94
Partial Profile	.37	.87	.39	.82	.47	.95

Cell entries = Pearson's r

© 2004 Millward Brown IntelliQuest 7 www.intelliquest.com

The notion that brand utilities derived from a partial profile design were reflected in secondary regression models. We used the brand utilities as a dependent variable in a set of OLS models. The independent variables were brand image attributes. The results, shown in Figure 4, show that the models were stronger and that a larger number of image attributes were found to be statistically significant predictors of the partial profile utilities. Again, we believe that the projection of brand attributes is occurring with the reduction of information in the partial profile designs.

Figure 4



Millward Brown IntelliQuest

Accuracy Results: Relationship Between Utilities and Brand Image Attributes

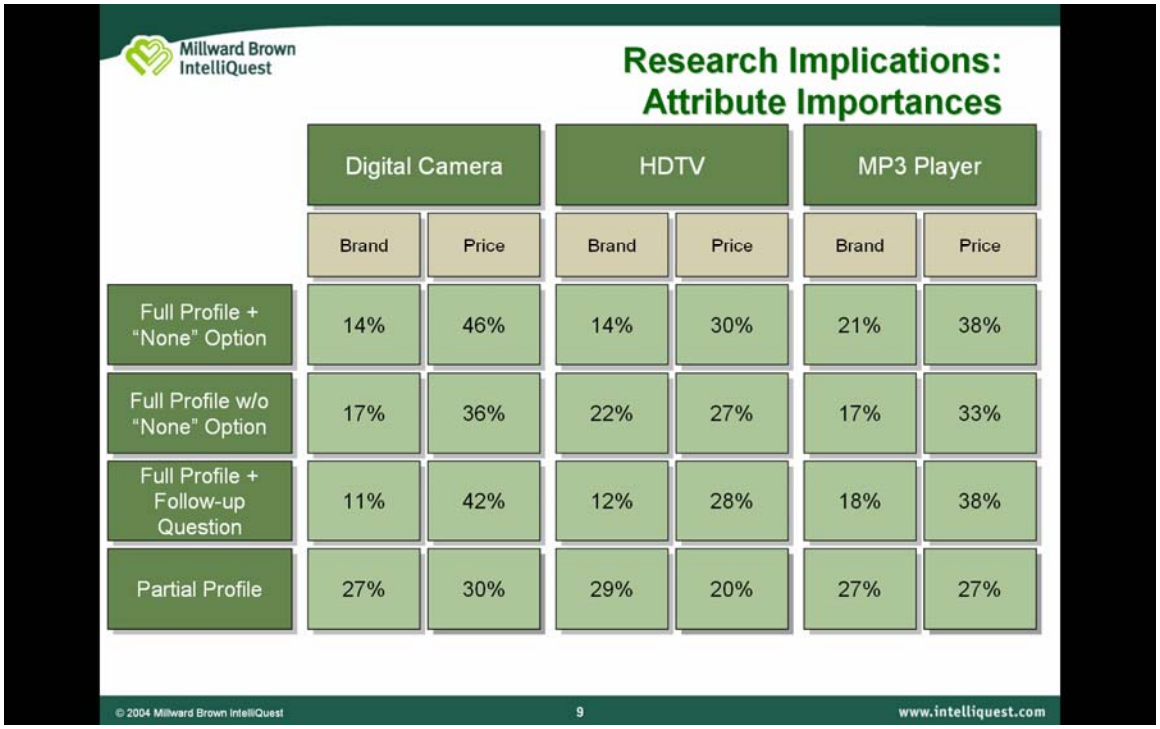
	Digital Camera		HDTV		MP3 Player	
	R^2	# Sig. Predictors	R^2	# Sig. Predictors	R^2	# Sig. Predictors
Full Profile + "None" Option	.37	4	.12	5	.35	4
Full Profile w/o "None" Option	.33	4	.13	4	.31	3
Full Profile + Follow-up Question	.28	4	.19	6	.42	4
Partial Profile	.45	6	.33	6	.48	7

© 2004 Millward Brown IntelliQuest 8 www.intelliquest.com

We also tested whether the attribute importances were significantly different between the four variants (Figure 5). We had previously found that the utilities themselves were not significantly different between the variants. However, as the image below shows, there are consistent differences between the variants in terms of the importance of brand and price. We can see that price is consistently more important for the full profile variants (particularly those that offer a "none of these" option) than the partial profile option.

The importance of brand is significantly higher in the partial profile than in the full profile alternatives.

Figure 5



Millward Brown IntelliQuest

Research Implications: Attribute Importances

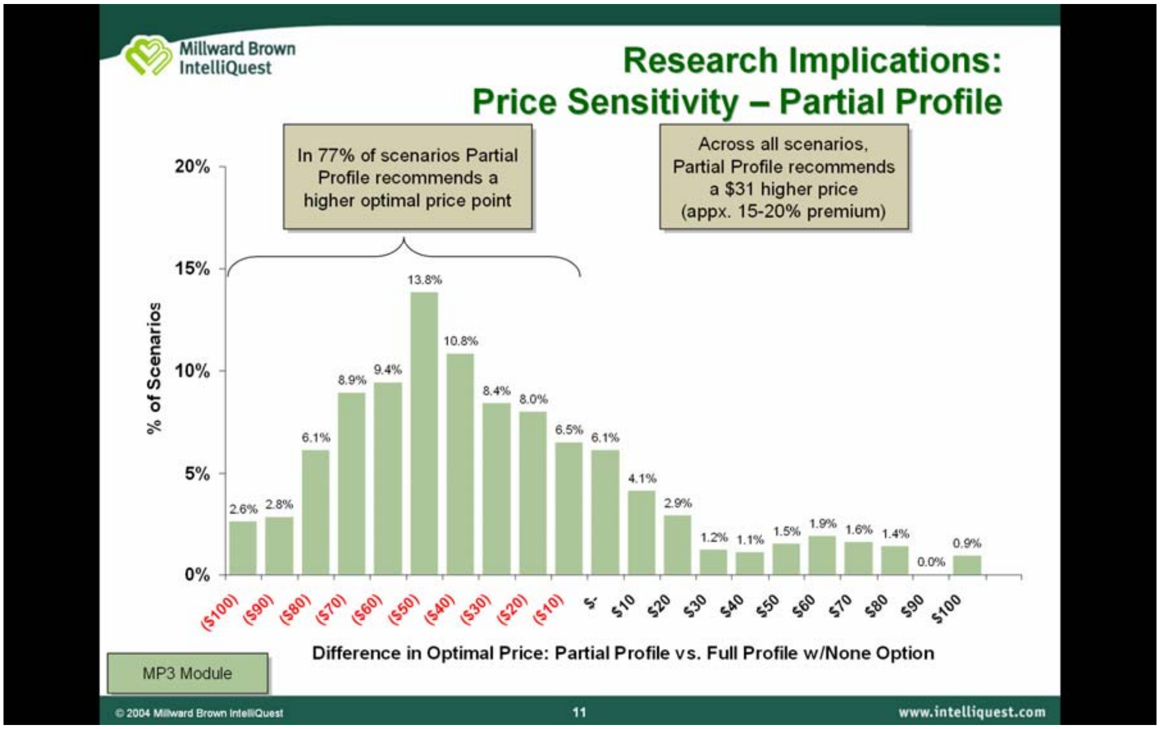
	Digital Camera		HDTV		MP3 Player	
	Brand	Price	Brand	Price	Brand	Price
Full Profile + "None" Option	14%	46%	14%	30%	21%	38%
Full Profile w/o "None" Option	17%	36%	22%	27%	17%	33%
Full Profile + Follow-up Question	11%	42%	12%	28%	18%	38%
Partial Profile	27%	30%	29%	20%	27%	27%

© 2004 Millward Brown IntelliQuest 9 www.intelliquest.com

Finally, we decided to test the price sensitivity more directly. This is because attribute importances, because they are usually represented as a proportionalized measure, can be misleading. As a remedy, we generated 500 random market scenarios. For each of these scenarios, we calculated the optimal price (operationalized as the price that maximizes expected revenue) for each of the choice variants.

Figure 6 illustrates one set of the results. In this chart, we are measuring the delta between the optimal price point for full profile with a "none" option and partial profile. The negative numbers indicate the percentage of scenarios in which the partial profile recommended a higher price. Looking across all scenarios, partial profile would lead the client to a price 15-20% higher than would the full profile approach.

Figure 6



Millward Brown IntelliQuest

Research Implications: Price Sensitivity – Partial Profile

In 77% of scenarios Partial Profile recommends a higher optimal price point

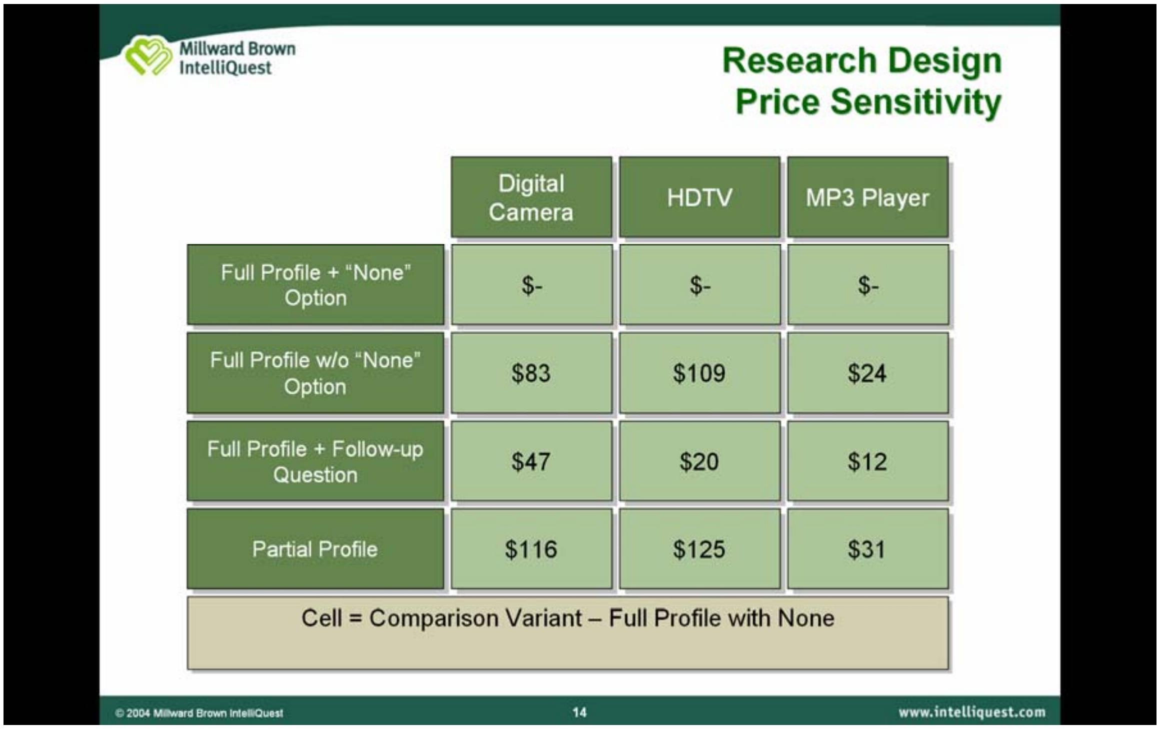
Across all scenarios, Partial Profile recommends a $31 higher price (appx. 15-20% premium)

MP3 Module

© 2004 Millward Brown IntelliQuest 11 www.intelliquest.com

Figure 7 aggregates the price sensitivity analyses across products and techniques. The results are remarkably consistent.

Figure 7



Millward Brown IntelliQuest

Research Design Price Sensitivity

	Digital Camera	HDTV	MP3 Player
Full Profile + "None" Option	$-	$-	$-
Full Profile w/o "None" Option	$83	$109	$24
Full Profile + Follow-up Question	$47	$20	$12
Partial Profile	$116	$125	$31

Cell = Comparison Variant – Full Profile with None

© 2004 Millward Brown IntelliQuest 14 www.intelliquest.com

CONCLUSIONS

This study has focused on assessing the impact of discrete choice design decisions on business decision-making. The results have shown that design decisions can have a substantial impact on the substantive outcomes. The most consistent and interesting result is the difference in the impact of brand and price. Partial profile, and to a lesser extent full profile without a "none" option, showed significantly less price sensitivity and significantly more brand value than did other techniques.

This study does not definitively show whether the partial profile *price effect* is an underestimation of price sensitivity or if the full profile techniques are overestimating price sensitivity. However, there is a compelling benefit in partial profile if we wish to model drivers of brand utility, as partial profile brand utilities appear to represent more than do the brand utilities from a full profile design.

QUANTITATIVE APPROACHES WITH SOFT ATTRIBUTES

APPLICATION OF LATENT CLASS MODELS TO FOOD PRODUCT DEVELOPMENT: A CASE STUDY

RICHARD POPPER, JEFF KROLL
PERYAM & KROLL RESEARCH CORPORATION
JAY MAGIDSON
STATISTICAL INNOVATIONS

INTRODUCTION

Food manufacturers need to understand the taste preferences of their consumers in order to develop successful new products. The existence of consumer segments that differ in systematic ways in their taste preferences can have important implications for product development. Rather than developing a product to please all potential consumers, the manufacturer may decide to optimize the product for the most important segment (perhaps the largest or most profitable). Alternatively, the manufacturer may opt for developing a number of products with different sensory profiles, with the goal of satisfying two or more segments.

A number of analytical methods exist for conducting consumer segmentations, including such traditional methods as hierarchical clustering and k-means clustering. Recently, latent class (LC) models have gained recognition as a method of segmentation with several advantages over traditional methods (see, for example, [1], [2], [3]), but so far almost no applications of these models to food product development have been reported in the literature.

In some types of latent class models (namely, LC regression), segments are formed on the basis of predictive relationships between a dependent variable and a set of independent variables. As a result, segments are comprised of people who have similar regression coefficients. These models can be of particular utility to food developers who need to relate a segment's product preferences to the underlying sensory attributes (taste, texture, etc.) of the products. By including sensory attributes as predictors, LC regression models promise to identify the segments and their sensory drivers in one step and provide highly actionable results.

This paper presents a case study involving the consumer evaluation of crackers. The objectives of the research are a) to determine if consumers can be segmented on the basis of their liking ratings of the crackers b) to estimate and compare a number of LC models, as well as some non-LC alternatives and c) to identify and interpret segments in terms of the sensory attributes that drive liking for that segment (in the case of the regression models).

Many consumer segmentation studies show evidence of individual differences in response style (i.e. consumers differ systematically in how they use the response scale). A question of considerable practical importance is how to deal with such response style differences in the analysis. When such differences are ignored, the resulting segments often display a response level effect—one segment comprising individuals who rate all items using the upper end of the response scale, another segment comprising individuals who consistently use the lower end of the scale. Usually, the differences in response style are of little substantive interest; instead, the researcher is interested in how segments differ in their *relative* ratings of the items involved.

In addition to the objectives mentioned above, this paper also illustrates two approaches for separating an overall response level effect from differences in relative preferences for one cracker over another.

DESCRIPTION OF CASE STUDY

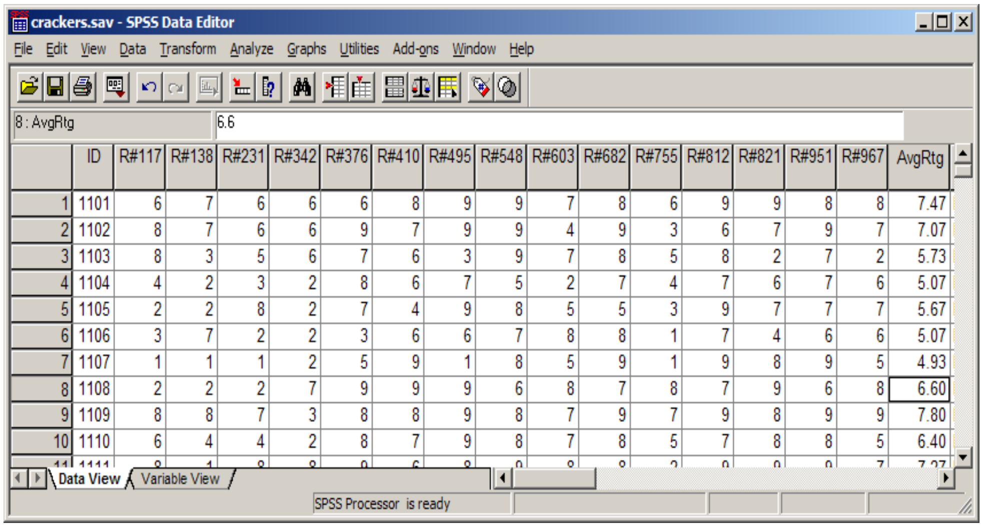
In this case study, consumers (N=157) rated their liking of 15 crackers on a nine-point liking scale that ranged from "Dislike Extremely" to "Like Extremely." Consumers tasted the crackers over the course of three sessions, conducted on separate days. The serving order of the crackers was balanced to account for the effects of day, serving position, and carry-over.

An independent trained sensory panel (N=8) evaluated the same crackers in terms of their sensory attributes (e.g. saltiness, crispness, thickness, etc.). The panel rated the crackers on 18 flavor, 20 texture, and 14 appearance attributes, using a 15-point intensity scale ranging from "low" to "high." These attribute ratings were subsequently reduced using principal component analysis to four appearance, four flavor, and four texture factors. The factors are referred to generically as APP1-4, FLAV1-4, and TEX1-4.

SEGMENTATION ANALYSES

Several types of models were used to obtain segments that differed with respect to their liking of crackers and to relate these differences to the sensory attributes. In all models, the liking data were treated as ordinal. Model fit was assessed using the Bayesian Information Criterion (BIC). Mathematical formulations of the models can be found in the Appendix.

<u>LC Cluster Model</u>. This is the traditional latent class model, which imposes no special structure to distinguish between variation due to differences in overall response level and those due to *relative* differences in the liking of crackers. That is, the latent classes simply represent unordered levels (i.e., a nominal factor). The data layout required for this model is shown in Figure 1. In this layout, each respondent occupies one row (record) and the ratings for the 15 products are arranged in successive columns. No adjustment was made to the data to account for individual differences in overall response level (i.e., the original raw ratings data were analyzed). The analysis was carried out using Latent Gold 3.0.


crackers.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
SPSS Processor is ready

8 : AvgRtg 6.6

	ID	R#117	R#138	R#231	R#342	R#376	R#410	R#495	R#548	R#603	R#682	R#755	R#812	R#821	R#951	R#967	AvgRtg
1	1101	6	7	6	6	6	8	9	9	7	8	6	9	9	8	8	7.47
2	1102	8	7	6	6	9	7	9	9	4	9	3	6	7	9	7	7.07
3	1103	8	3	5	6	7	6	3	9	7	8	5	8	2	7	2	5.73
4	1104	4	2	3	2	8	6	7	5	2	7	4	7	6	7	6	5.07
5	1105	2	2	8	2	7	4	9	8	5	5	3	9	7	7	7	5.67
6	1106	3	7	2	2	3	6	6	7	8	8	1	7	4	6	6	5.07
7	1107	1	1	1	2	5	9	1	8	5	9	1	9	8	9	5	4.93
8	1108	2	2	2	7	9	9	9	6	8	7	8	7	9	6	8	6.60
9	1109	8	8	7	3	8	8	9	8	7	9	7	9	8	9	9	7.80
10	1110	6	4	4	2	8	7	9	8	7	8	5	7	8	8	5	6.40

Figure 1: Data Layout for the LC Cluster and LC Factor Models

LC Factor Model. A factor-based version of the latent class model was applied in order to try to "factor out" response level effects. Ordered levels of the first discrete factor (D-Factor #1) were used to model *overall* liking, isolating a response level effect. Additional dichotomous factors (D-Factors #2, 3, etc) were considered in order to identify segments that differ in their *relative* liking of the products. The data layout required for this analysis is the same as for the LC Cluster model. Latent Gold 3.0 was used for the analysis.

Regression Models. Four types of regression models were explored. All of these models used a continuous random intercept to account for individual differences in average liking across all products. Use of a continuous factor (C-Factor) rather than a discrete factor to account for the overall level effect was expected to result in segments that better represented pure *relative* differences in cracker liking. In addition, regression models allow for the possibility of using the attributes as predictors, thus allowing the segments to be defined in terms of differences in the attribute effects. This cannot be done using the cluster or factor models.

The data layout required for these analyses is shown in Figure 2. In this layout, there are 15 rows (records) per respondent. The consumer overall liking ratings of the products are contained in the column labeled "Rating", the sensory attribute information is in the succeeding columns.

The regression models differ in the predictors. Two use only PRODUCT (15 nominal categories for the 15 crackers) as the sole predictor, while two others use the quantitative attributes as predictors. The models also differ in their approach to modeling the respondent heterogeneity in the effect of the products. The key differences among the four types of models are summarized in Table 1.

Table 1: Four Types of Regression Models

		Segments Defined Using	
		Latent Classes	Continuous Factors
Predictors	PRODUCT (15 nominal product levels)	Model 1	Model 2
	Twelve quantitative sensory attributes	Model 3	Model 4

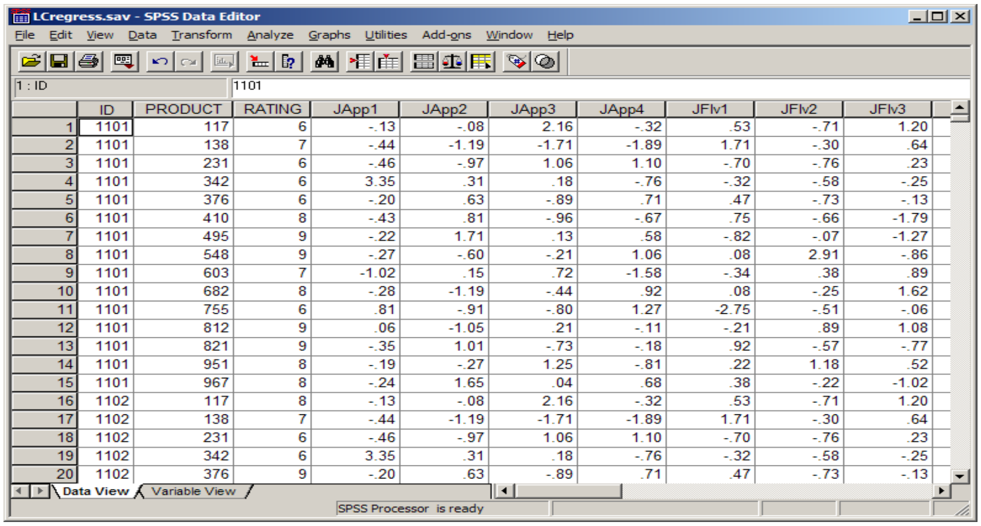
Model 1 used the nominal variable PRODUCT as the sole predictor. It included a class-independent continuous random intercept (C-Factor #1) to capture respondent differences in average liking across all products, and latent classes as a nominal factor to define the segments in terms of the heterogeneity in this PRODUCT effect.

Model 2 also used the nominal variable PRODUCT as the predictor and included a class-independent continuous random intercept (C-Factor #1) to capture response levels effect. However, in contrast to Model 1, one or more additional continuous factors (C-Factors #2, etc.) were considered in order to account for the heterogeneity in the PRODUCT effect.

Model 3 is the same as Model 1, except that it used the 12 sensory attributes as predictors.

Model 4 is the same as Model 2, except that it used the 12 sensory attributes as predictors.

The four factor-regression models (containing two or more factors) were estimated using a forthcoming version of Latent Gold (4.0).


LCregress.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : ID
1101
Data View
Variable View

	ID	PRODUCT	RATING	JApp1	JApp2	JApp3	JApp4	JFlv1	JFlv2	JFlv3
1	1101	117	6	-.13	-.08	2.16	-.32	.53	-.71	1.20
2	1101	138	7	-.44	-1.19	-1.71	-1.89	1.71	-.30	.64
3	1101	231	6	-.46	-.97	1.06	1.10	-.70	-.76	.23
4	1101	342	6	3.35	.31	.18	-.76	-.32	-.58	-.25
5	1101	376	6	-.20	.63	-.89	.71	.47	-.73	-.13
6	1101	410	8	-.43	.81	-.96	-.67	.75	-.66	-1.79
7	1101	495	9	-.22	1.71	.13	.58	-.82	-.07	-1.27
8	1101	548	9	-.27	-.60	-.21	1.06	.08	2.91	-.86
9	1101	603	7	-1.02	.15	.72	-1.58	-.34	.38	.89
10	1101	682	8	-.28	-1.19	-.44	.92	.08	-.25	1.62
11	1101	755	6	.81	-.91	-.80	1.27	-2.75	-.51	-.06
12	1101	812	9	.06	-1.05	.21	-.11	-.21	.89	1.08
13	1101	821	9	-.35	1.01	-.73	-.18	.92	-.57	-.77
14	1101	951	8	-.19	-.27	1.25	-.81	.22	1.18	.52
15	1101	967	8	-.24	1.65	.04	.68	.38	-.22	-1.02
16	1102	117	8	-.13	-.08	2.16	-.32	.53	-.71	1.20
17	1102	138	7	-.44	-1.19	-1.71	-1.89	1.71	-.30	.64
18	1102	231	6	-.46	-.97	1.06	1.10	-.70	-.76	.23
19	1102	342	6	3.35	.31	.18	-.76	-.32	-.58	-.25
20	1102	376	9	-.20	.63	-.89	.71	.47	-.73	-.13

SPSS Processor is ready

Figure 2. Data Layout for the Regression Models.

MODEL RESULTS

LC Cluster Model

According to the BIC, a two-cluster solution was a better fit to the data than either a one-cluster or three-cluster solution. The two clusters (segments) were approximately equal in size (53% and 47%). Figure 3 shows each segment's average liking ratings for the products. This figure shows that the two segments are clearly and almost exclusively differentiated by their overall average liking of the crackers. Segment 2 rated almost all products higher than Segment 1. This result is not unexpected, since no attempt was made in the analysis to adjust the data for differences in response level. The figure also shows *some* relative differences in liking between the two clusters. For example, Segment 2 liked Products #495, #376, #821 and #967 more relative to the other products, whereas Segment 1 liked them less.

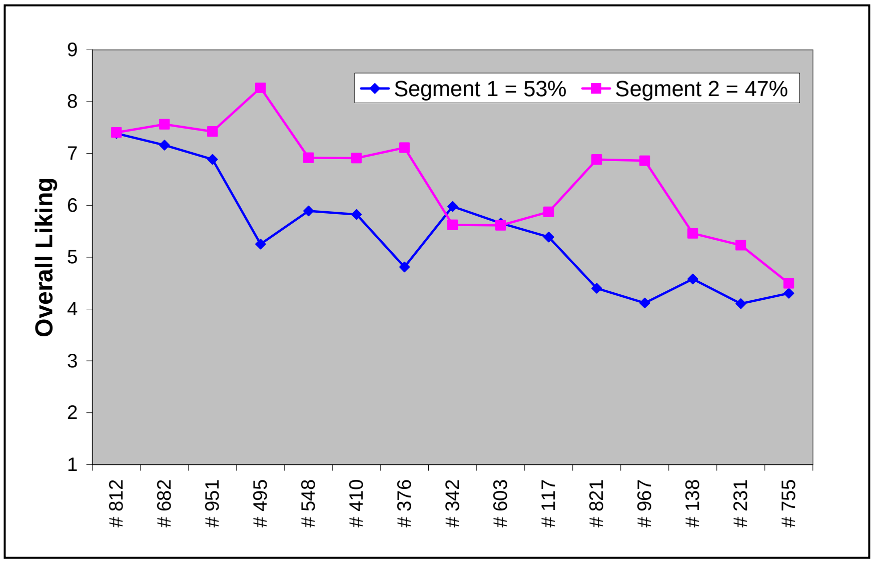


Figure 3. LC Cluster Results.

LC Factor Model

In fitting one-factor (ordered-level) models, the BIC indicated that a factor with four levels resulted in the best fit to the data. Figure 4 shows the average product liking scores for the four levels (classes) on this discrete factor (D-Factor #1). This factor corresponds largely to a level effect. Average liking scores (across products) for the four levels were 4.7 (Level 1), 4.8 (Level 2), 6.7 (Level 3) and 7.3 (Level 4). The interpretation of the factor as a level effect is further supported by the fact that the correlation between individual respondents' scores on this factor and their average liking was 0.87.

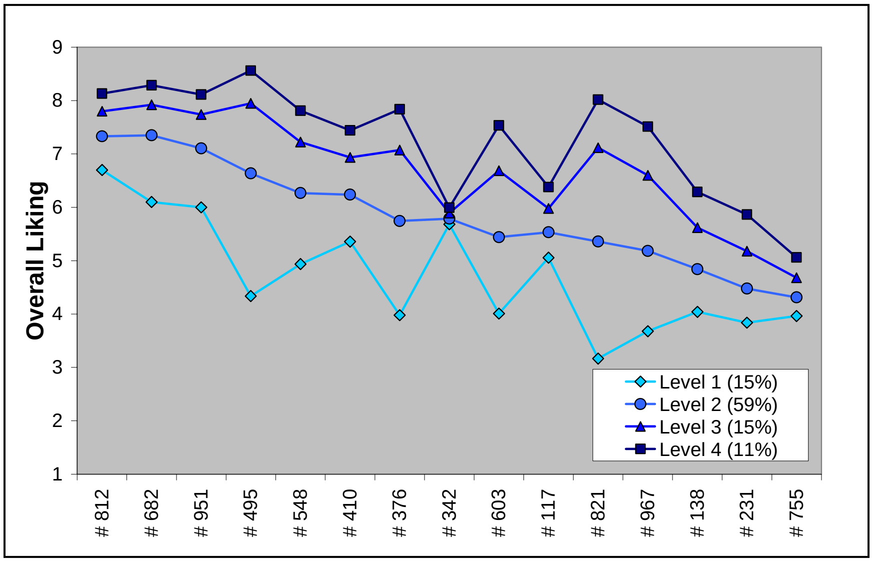


Figure 4. LC Factor Results for D-Factor #1.

According to the BIC, a second dichotomous factor further improved the fit over the one factor model, but additional dichotomous factors did not. The average product liking scores for the two levels (classes) of this second factor (D-Factor #2) are shown in Figure 5. In contrast to the first factor, the classes of D-Factor #2 are differentiated mainly in their relative liking of the products – the average liking across all products was nearly the same for both classes (5.9 and 6.0 for the two levels of D-Factor #2, respectively). Taking each level of D-Factor #2 as a segment, we see that Segment 2 liked Products #495, 376, 821, and #967 more than Segment 1, but liked Products #812, #342, and #603 less. The two factors were assumed to be uncorrelated with each other in the model.

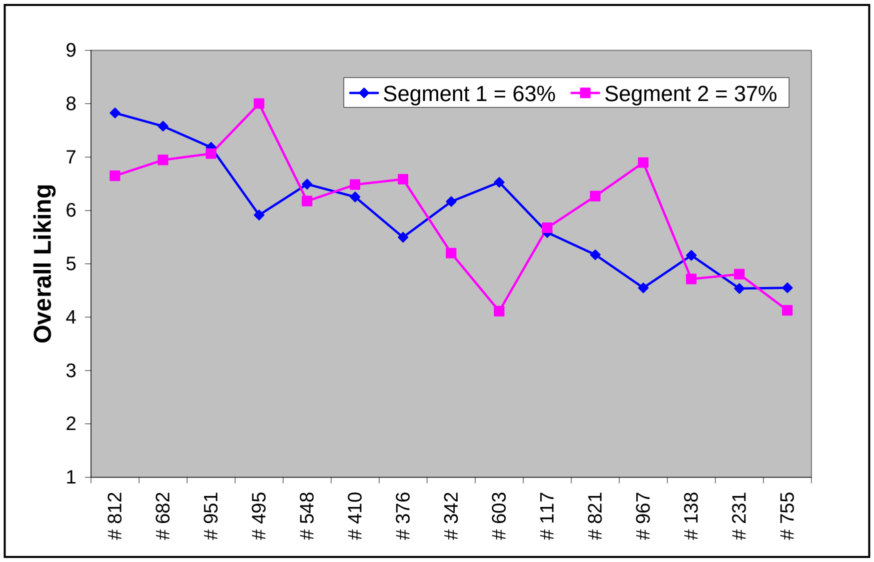


Figure 5. LC Factor Results for D-Factor #2.

Comparison of LC Cluster and LC Factor Models

The most important difference between the two models is that the LC Cluster model confounded relative differences in liking with average response level (one segment rated almost all products higher than the other). The LC Factor model, on the other hand, was able to separate out a response level effect (D-Factor # 1) from an effect that reflected the relative differences in liking (D-Factor #2). Also, the LC Factor model was preferred over the LC Cluster models according to the BIC (BIC = 9,887 for the LC Factor model, compared to 9,926 for the two-class cluster model and 9,930 for the three-class cluster model).

The same products that differentiated the segments in the LC Cluster model differentiated the segments on D-Factor #2 (#495, 376, 821, and 967). But D-Factor #2 further differentiated the segments in their response to Products #812 and #603, whereas the LC Cluster model showed no difference between these products (see Figure 6). In the LC Cluster solution, these differences were masked by the differences in response level between the clusters. In addition to these differences between the two models, the two models also differed in the relative sizes of the two segments.

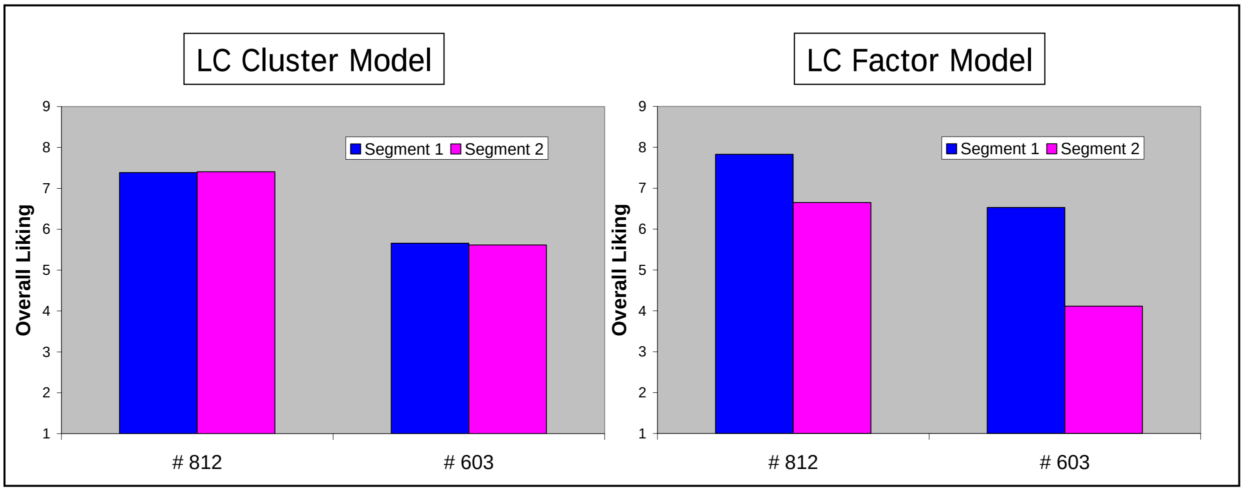


Figure 6. Comparison of LC Cluster and Factor Models for three products**.**

Regression Models

Model 1

The correlation of the random intercept with respondents' average liking was higher than 0.99 (compared to 0.87 for D-Factor #1), indicating that the random intercept was better able to capture individual differences in average response level than D-Factor #1 in the LC Factor model. A two-class solution provided the best fit to the data, with a model R^2 of 0.39.

Figure 7 shows the average product liking scores for the two segments. Segment 2 liked products #495, #376, #821, and #967 more than Segment 1, but liked #812, #342,

#603 less. Liking when averaged across all products was nearly identical for the two segments (5.9 and 6.1 for Segments 1 and 2, respectively).

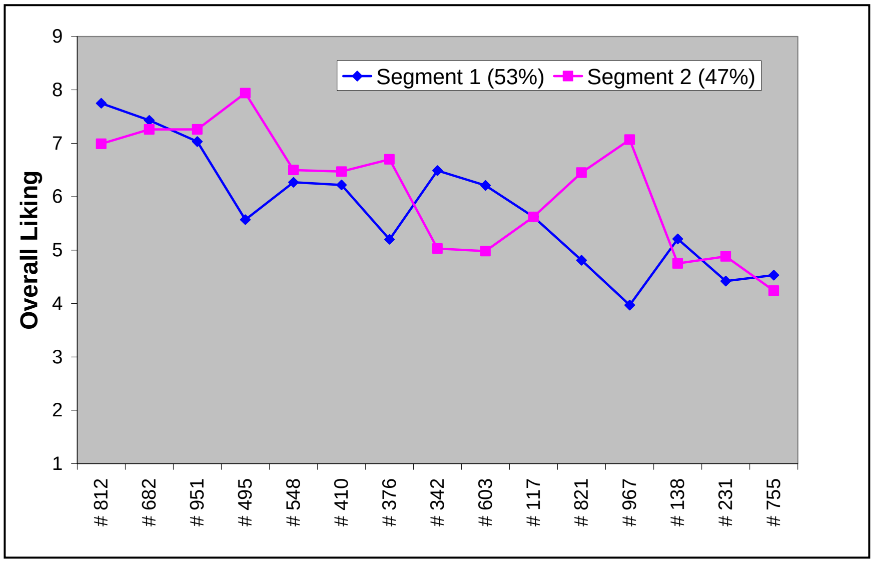


Figure 7. Regression Model 1 Results.

Model 2

As in Regression Model 1, the correlation of the random intercept in Model 2 with average liking was greater than 0.99. The addition of a second C-Factor improved the model fit according to the BIC, but a third C-Factor did not. The model with two C-Factors had an R^2 of 0.41.

Unlike Regression Model 1, Regression Model 2 is not a latent class model and thus does not provide guidance as to the number of underlying segments. For comparison to Model 1, two segments of respondents were formed on the basis of the distribution of scores on C-Factor #2 (using a cutoff point at the mean value of zero). Respondents with scores on C-Factor #2 less than zero were assigned to one segment, and those with scores greater than zero to another segment. Figure 8 shows the average product liking scores by these two groups of respondents, which were approximately equal in size and differentiated on the same products as the segments in Model 1.

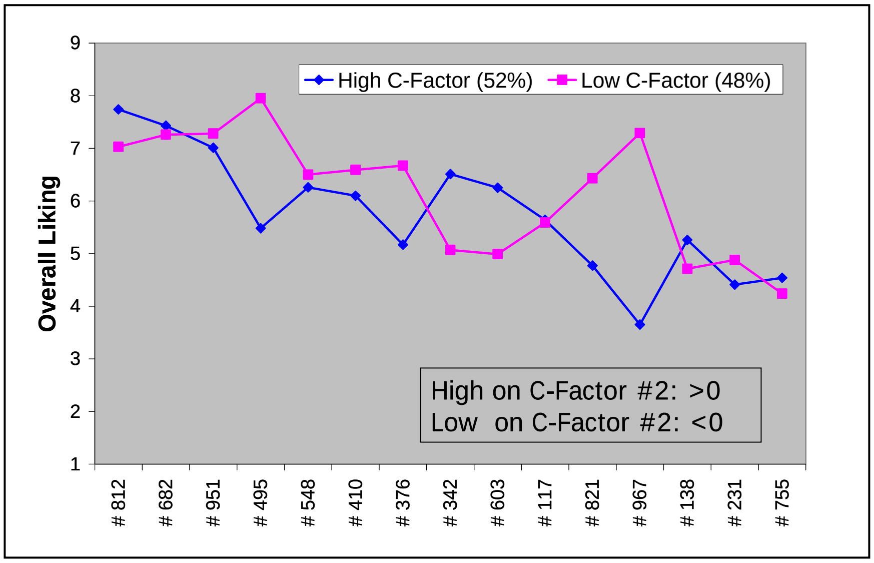


Figure 8. Regression Model 2 Results.

The respondent heterogeneity with respect to the effect of PRODUCT can be assessed more precisely without the formation of segments, by simply examining the magnitude of the interaction effects between the nominal PRODUCT effect and C-Factor #2 (see Table 2). The results represented by the interaction z-statistic are visually represented by the segment means plot in Figure 8. For example, the largest interaction effects occur for Crackers #495 and #967, which are seen in Figure 8 to yield the largest differences between the High and Low C-Factor #2 segments.

Table 2. Regression Model 2: Main Effects and Interaction Effects

Product	Main Effect	Main Effect Z-statistic	Interaction Effect	Interaction Effect Z-statistic
812	0.46	7.57	0.25	3.62
682	0.42	7.66	0.13	2.04
951	0.33	6.56	0.03	0.41
495	0.30	4.41	-0.46	-5.69
548	0.07	1.62	0.06	1.11
410	0.07	1.54	-0.03	-0.54
376	-0.03	-0.71	-0.18	-3.38
342	-0.07	-1.60	0.27	5.00
603	-0.12	-2.97	0.18	3.80
117	-0.13	-3.31	0.10	1.99
821	-0.11	-2.45	-0.23	-3.92
967	-0.13	-2.29	-0.43	-5.50

Note: Z-statistics with absolute values larger than 2 are statistically significant at the .05 level.

Comparison of Regression Models 1 and 2

The two models lead to similar conclusions regarding segment differences and are equally parsimonious (both models used 38 parameters). According to the BIC, Model 2 is preferred to Model 1 (BIC(2)=9,461 vs. BIC(1)=9,487), and the R^2 is also slightly higher. On the other hand, Model 1 provides guidance as to the number of underlying segments, Model 2 does not. Model 2 also does not offer guidance as to where to choose cut-points on the C-Factor and required the use of an arbitrary assignment rule in the formation of segments.

Model 3

The correlation of the random intercept with average liking across products was again very high (>0.99). The BIC was lower for an unrestricted two-class model (BIC=9,535) than for an unrestricted three-class model (BIC=9,560), indicating that the two-class model was preferred. However, a three-class *restricted* model that restricted the third class regression coefficients to zero for all 12 predictors had a slightly better BIC (9,531) than the two-class model. The model R^2 for the three-class restricted regression model was 0.39, the same as for Model 1 (which used the nominal PRODUCT variable as the predictor).

The interpretation of the third class is that it consists of individuals whose liking does not depend on the levels of the 12 sensory attributes. This segment was small (8%), compared to the size of the other two segments (42% and 50% for Segments 1 and 2, respectively).

Figure 9 shows the average product liking scores for the three-class restricted model. The plot of regression coefficients in Figure 10 provides a visual display of the extent of the segment differences in attribute preferences. Segment 2 prefers products high in APP2 and low in APP3. Segment 1 was not highly influenced by these two characteristics, but preferred crackers high in APP1. Both clusters agree that they prefer crackers that are high in FLAV1-3, low in FLAV4, low in TEX1 and high in TEX2-3.

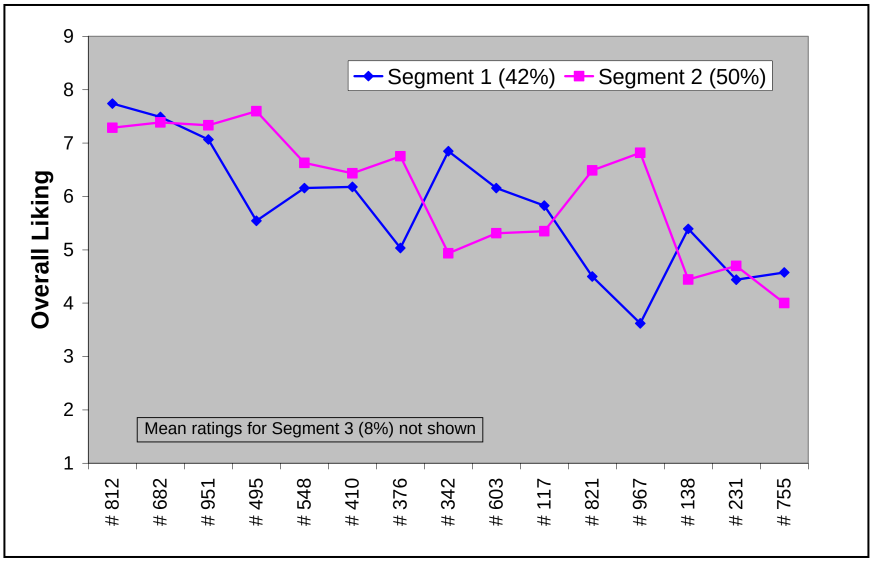


Figure 9. Regression Model 3 Results.

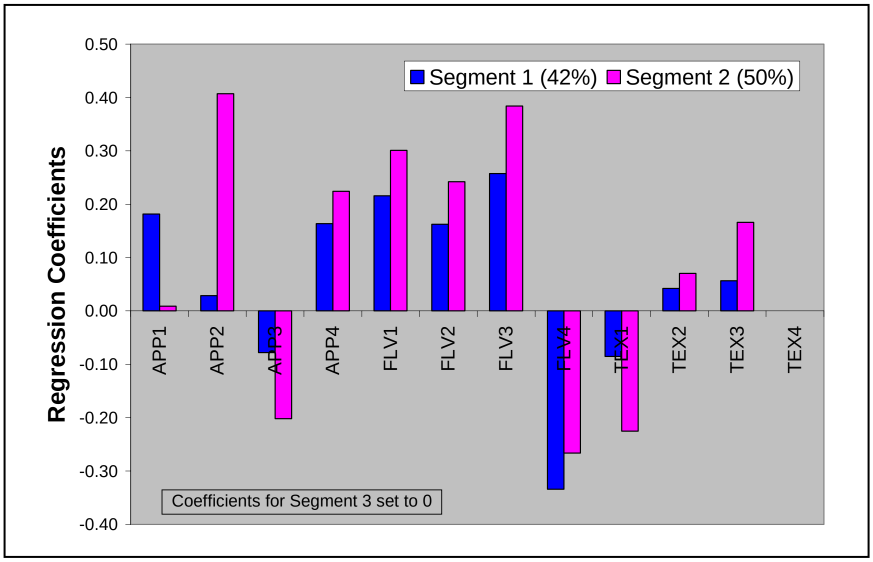


Figure 10. Regression Coefficients for Regression Model 3.

Model 4

The random intercept was again highly correlated with individual respondents' average liking (>0.99). As was the case with Model 2, the addition of a second C-Factor improved the fit (according to the BIC), but a third factor did not. The model R^2 was 0.38, slightly lower than for Model 2, which used the nominal PRODUCT variable as the predictor instead of the 12 sensory attributes.

Two segments were formed on the basis of C-Factor #2 by assigning respondents with scores less than zero to one group and respondents with scores higher than zero to the other. Figure 11 shows the average product liking scores for the two groups and indicates a pattern similar to that for the latent class model using the sensory predictors (Model 3).

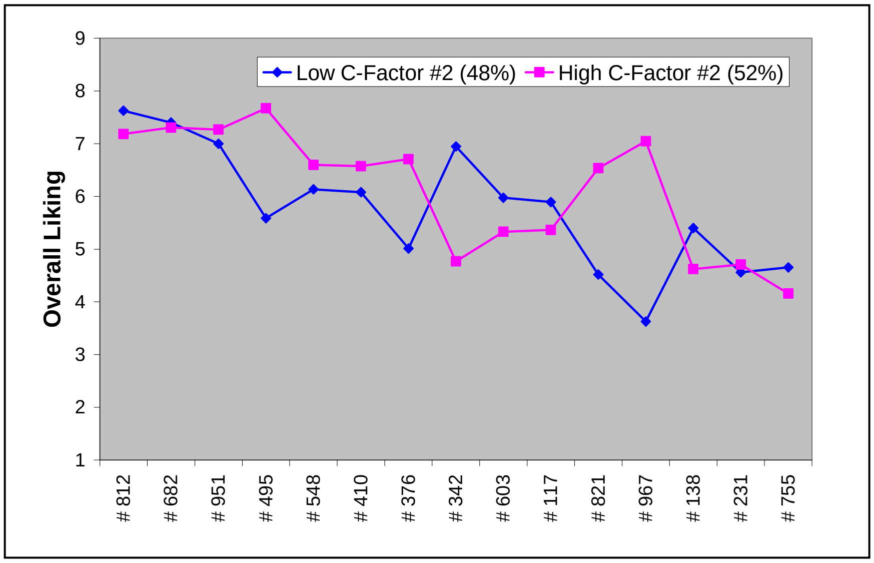


Figure 11. Regression Model 4 Results.

Table 3 provides a statistical assessment of the sensory predictors in terms of main effects and interactions with C-Factor #2. A significant *main effect* indicates that higher levels for the associated attribute significantly increase (or decrease) the rating for all respondents. All sensory predictors except for TEX4 had significant *main effects* on overall liking, but only one attribute (APP2) yielded a significant *interaction effect* ($p<0.05$). For two other predictors, APP1 and APP3, the interaction with C-Factor #2 approached statistical significance ($p\leq 0.10$).

The overall effect of an attribute is given by the main effect plus the C-Factor #2 score multiplied by the interaction effect for a given respondent. For example, for a respondent scoring one standard deviation *above* the mean on C-Factor #2 (C-Factor #2 score = +1), the overall effect of appearance attribute APP2 can be computed as 0.21 +1*0.21 = 0.4. Similarly, for a respondent scoring one standard deviation *below* the mean on C-Factor #2 (C-Factor #2 score = -1), the overall effect of APP2 is 0.21 -1* 0.21 = 0. Thus, respondents scoring higher on C-Factor #2 are more favorably affected by APP2, whereas those scoring low on the factor (say C-Factor #2 = -1) tend to be neutral to APP2.

Note that in addition to the random intercept, only a single additional C-Factor was required to account for respondent heterogeneity. Thus, this model is substantially more parsimonious than a traditional HB (Hierarchical Bayes) type model, which would be equivalent to including 12 additional C-Factors, one for each attribute.

Table 3. Regression Model 4: Main Effects and Interaction Effects.

Product	Main Effect	p-value	Interaction Effect	p-value
APP1	0.08	0.02	-0.08	0.10
APP2	0.21	0.0001	0.21	0.0008
APP3	-0.13	0.0001	-0.07	0.06
APP4	0.19	0.0000	0.02	0.67
FLV1	0.24	0.0000	0.04	0.18
FLV2	0.19	0.0000	0.03	0.44
FLV3	0.29	0.0000	0.07	0.31
FLV4	-0.27	0.0000	0.04	0.34
TEX1	-0.14	0.0180	-0.06	0.45
TEX2	0.05	0.0026	0.01	0.63
TEX3	0.10	0.0047	0.05	0.27
TEX4	0.00	0.95	0.01	0.85

COMPARISON OF REGRESSION MODELS 3 AND 4

The two models are similar in R^2 and are similar in parsimony (Model 3 used 33 parameters compared to 34 in Model 4). The BIC was somewhat better for Model 4 (BIC(3)=9,531 vs. BIC(4)=9,525). The models differ in their use of a discrete (Model 3) vs. continuous (Model 4) measure of respondent heterogeneity. A weakness of Model 4 is that the continuous factor (C-Factor #2) does not yield clearly differentiated segments. Figure 12 shows that the distribution of C-Factor #2 scores is normal, with many scores around zero. This makes the choice of a cut-point for formation of segments arbitrary, as was the case for Model 2. Model 3 provides clear segment differentiation. The strength of that segmentation is further indicated by the posterior membership probabilities (see Figure 13), which show that the average estimated probability of cluster membership is high (>0.8) for the two large segments (Segments 1 and 2), and above 0.5 for Segment 3.

While Model 4 does not provide guidance with respect to segment formation, the segments identified in Model 3 have a close correspondence with C-Factor #2 scores estimated in Model 4. Figure 14 shows that respondents assigned to Segment 1 in Model 3 score high on C-Factor #2, those assigned to Segment 2 score low on C-Factor #2, and those assigned to Segment 3 score around zero on C-Factor #2. Thus, it should not be surprising that Regression Models 3 and 4 provide very similar inferences about which attributes are most important (significant main effects) and which are most involved in respondent heterogeneity (significant interaction effects).

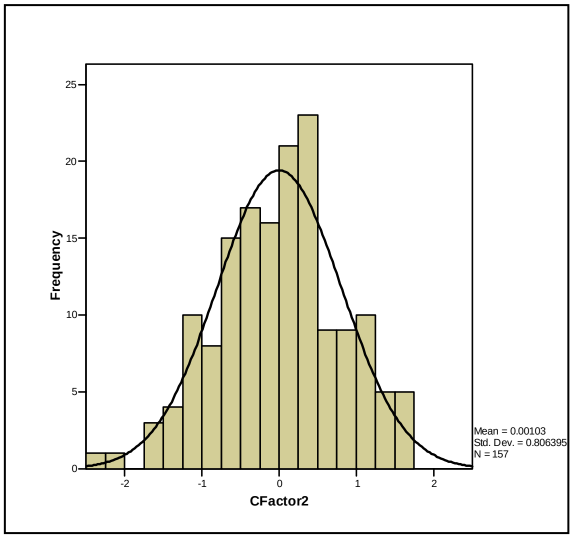


Figure 12. Distribution of C-Factor #2 Scores in Regression Model 4

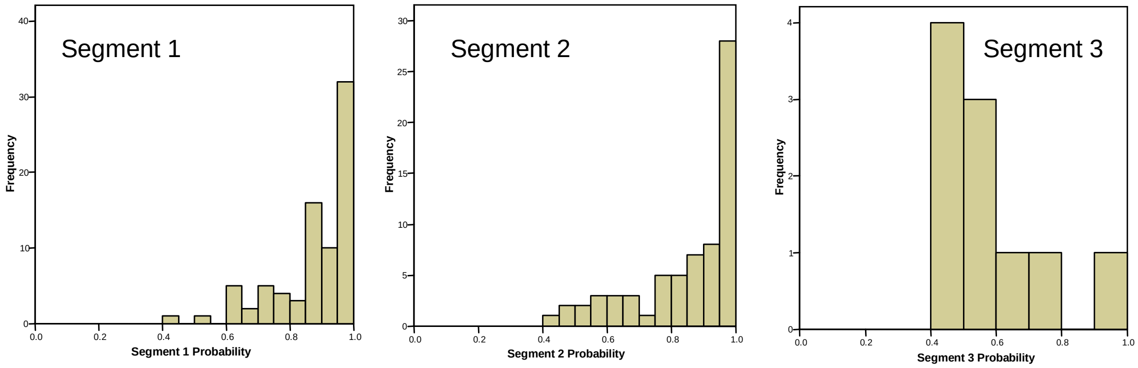


Figure 13. Posterior Membership Probabilities for Segments Determined in Regression Model 3.

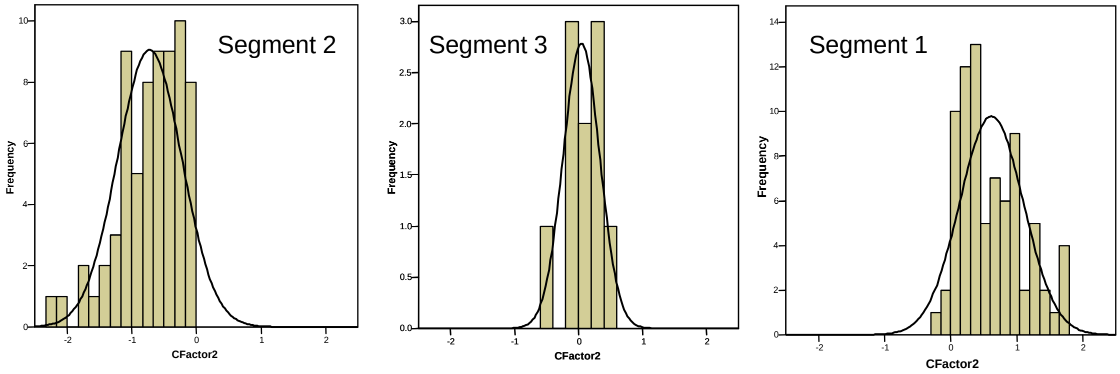


Figure 14. Distribution of C-Factor #2 Scores (Regression Model 4) for Segments Determined in Regression Model 3.

General Discussion

Four alternative approaches were presented for segmenting consumers on the basis of their overall liking ratings without consideration of the products' sensory attributes: the LC Cluster model, the LC Factor model, and the two regression models (Models 1 and 2) that used the nominal product variable as the predictor. All models provided evidence of the existence of segment differences in consumers' liking ratings. While some products appealed to everybody, other products appealed much more to one segment than another.

The LC Cluster model identified two segments that differed in their average level of liking and showed some relative differences in liking as well. The LC Factor model used two factors to isolate segment differences associated with average response level from differences in relative product liking. The segments given by the second factor provided a clearer picture of relative product differences than did the LC Cluster solution.

The regression models with a random intercept yielded segments that provided an even clearer picture of the relative differences than that given by the LC Factor model, since they allowed for a cleaner separation of the response level effect. The correlation of the random intercept was in excess of 0.99 for both the LC Regression model (Model 1) and the non-LC regression model (Model 2). Both correlations exceeded the correlation of 0.87 between the first factor of the factor model and average liking.

Including a random intercept is conceptually similar to mean-centering each respondents' liking ratings. A LC Cluster analysis of the mean-centered liking data would yield similar results to those obtained with LC Regression Model 1. However, there are two reasons to prefer the regression approach in general. With the regression approach, it is possible to maintain the ordinal discrete metric of the liking data. Subtracting an individual's mean from each response distorts the original discrete distribution by transforming it into a continuous scale that has a very complicated distribution. Secondly, in studies where a respondent only evaluates a subset of products, mean-centering is not appropriate since it ignores the incomplete structure of the data. Thus, the regression approach provides an attractive model-based alternative for removing the response level effect.

The use of a continuous vs. discrete random product effect in Regression Model 2 (compared to Regression Model 1) led to a slightly improved model fit, but at a price. The non-LC regression approach does not determine the cut-points to use to identify segments. An arbitrary choice of cut-points is likely to lead to substantial misclassification, in the event that distinct segments do in fact exist.

Replacing the nominal PRODUCT predictor with the twelve quantitative appearance, flavor and texture attributes made it possible to relate liking directly to these attributes. This allowed for the identification of both positive and negative drivers of liking. Segments reacted similarly to the variations in flavor and texture, but differed with regard to how they reacted to the products' appearance. Based on such insights, product developers can proceed to optimize products for each of the identified segments.

Replacing the nominal PRODUCT variable with the sensory predictors did not lead to any substantial loss in model fit. The R^2 for Model 3 was the same as for Model 1, and the fit for Model 4 only slightly below that of Model 2 (0.39 vs. 0.41).

The non-LC regression models (Model 2 and 4) can be compared to Hierarchical Bayes (HB) models. The HB models are equivalent to regression models containing one continuous factor (C-Factor) for each (non-redundant) predictor regression coefficient plus one additional C-Factor for the intercept (15 C-Factors for Model 2 and 13 for Model 4). Since in the analysis of these models the BIC did not support the use of more than two C-Factors, Models 2 and 4 offer much more parsimonious ways to account for continuous heterogeneity than HB. HB would likely overfit these data by a substantial margin, and at least some differences in liking suggested by an HB model would therefore not validate in the general population.

Since no group of quantitative predictors is going to be able to exceed the strength of prediction of the nominal PRODUCT variable with its fourteen degrees of freedom, Models 1 and 2 provide an upper bound on the R^2 for Models 3 and 4, respectively. A comparison of the R^2 of Models 3 and 1 (and Models 4 and 2) provides an assessment of how well the sensory predictors perform relative to the maximally achievable prediction. In this case study, the twelve sensory attributes captured almost all the information contained in the nominal PRODUCT variable that was relevant to the prediction of overall liking. The inclusion of additional predictors (for example, quadratic terms to model a curvilinear relationship between liking and sensory attributes) is therefore not indicated, although in other applications cross-product terms or quadratic terms could be very important in improving model fit or optimizing the attribute levels in new products.

The use of restrictions in LC Regression Model 3 improved the fit over an unrestricted model and allowed for the identification of a third segment, one whose overall liking of the products was not influenced by the sensory attributes. While this group was small, in certain applications such a group could be of substantive interest and warrant follow-up. If nothing else, the members of such a group can be excluded as outliers.

Regression Models 3 and 4 differed in their use of a discrete vs. continuous measure of respondent heterogeneity. The models yielded similar fit statistics and conclusions about the attribute effects and heterogeneity. Since Model 3 yielded clear segments more easily than Model 4 and the fit was almost the same, Model 3 was preferable for our purposes.

Among all the models tested (cluster, factor and the four regression models), Regression Model 3 yielded the most insight into the consumer liking of the products: the model provided clear segment differentiation, it isolated the response level effects from the sensory attribute effects that were of more substantive interest, and it identified the sensory drivers of liking for each segment.

ACKNOWLEDGMENT

The authors wish to thank The Kellogg Company for providing the data for this case study.

REFERENCES

Magidson, J. and Vermunt, J.K. (2002). Latent class modeling as a probabilistic extension of K-means clustering. Quirk's Marketing Research Review, March 2002, 20 & 77-80.

Magidson, J. and Vermunt, J.K. (2002). Latent class models for clustering: A comparison with K-means. Canadian Journal of Marketing Research, 20, 36-43.

Vermunt, J.K. and Magidson, J. Latent class cluster analysis. (2002) In: J. Hagenaars and A. McCutcheon (eds.). Applied latent class models, 89-106. Cambridge University Press.

Skrondal, A. & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.London: Chapman & Hall/CRC.

Appendix

1) Latent Class Cluster Model

If the ratings were treated as *continuous*, the normal error assumption yields a linear model. In this case, the predicted rating for cracker t for latent class x, is expressed as:

$$E(Y_{i.t}) = \alpha_t + \beta_{xt}, \qquad t = 1,2,\ldots,15 \qquad (1)$$

and effect coding is used for parameter identification:

$$\sum_{x=1}^{K} \beta_{xt} = 0$$

so the intercept α_t corresponds to an overall (average) rating effect for cracker t over all cases and for K latent class segments, the β_{xt} effects x = 1,2,…,K represent the segment differences.

The prediction for individual i, is generated by weighting these class level predictions by the corresponding posterior membership probabilities obtained for that individual.

Since the ratings are *not* continuous but discrete (Y=m; m=1,2,…,9), we instead assume the following adjacent category logit model which treats the ratings as *ordinal*. For K = 2 classes, we have:

$$logit(Y_{im.t}) = \alpha_{tm} + \beta_{xt}, \qquad t = 1,2,\ldots,15; \quad m = 2, 3,\ldots, 9; \quad x=1,2$$

where

$logit(Y_{im.t})$ is the adjacent category logit associated with rating Y = m (vs. m-1) for cracker *t*,

and effect coding is used for parameter identification:

$$\sum_{m=1}^{M=9} \alpha_{tm} = \sum_{x=1}^{K} \beta_{xt} = 0$$

so, similar to the continuous case, the intercepts capture average response levels.

Thus, for segment x :

α_{tm} m = 1,2,…,9 denote the intercepts associated with product *t*
and β_{xt}, x = 1,2 represents the effect of product *t*

2) LC Factor model

For ordinal ratings, the model is a restricted LC cluster model, where the segment differences are characterized by 2 independent discrete factors (D-Factors). The first D-Factor consists of 4 ordered levels, the second is dichotomous:

$logit(Y_{im.t}) = \alpha_{tm} + \beta_{xt}$ t = 1,2,…,15; m = 2, 3,…, 9;

where:

$\beta_{xt} = \beta_{t1}X_1 + \beta_{t2}X_2$

X_1 = 0, 1/3, 2/3, 1 for levels x_1 = 1,2,3 and 4 respectively of D-Factor X_1

X_2 = 0, 1 for levels x_2 = 1, 2 for dichotomous D-Factor X_2

where D-Factors X_1 and X_2 are stochastically independent of each other

3) Ordinal Regression Models

For each of the following regression models, the ratings are treated as ordinal. If instead they were treated as continuous the models would be linear regression models in which case the subscript m would not appear on the intercept and logit($Y_{im.t}$) would be replaced by E($Y_{i.t}$) as in the case of the LC cluster model. In addition, each regression model contains one or two continuous factors (C-Factors). For further details on the use of C-Factors see [4].

Model 1: LC Ordinal Regression with Random Intercept and Discrete Random PRODUCT Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{xt}$$
$$\alpha_{im} = \alpha_m + \lambda F_i$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$
$$V(\alpha_{im}) = \lambda^2$$

where m=2,3,…,9 and V denotes the variance.

logit($Y_{j.k}$) is the adjacent category logit associated with rating Y = m (vs. m-1) for product *t*

F_i is the C-Factor score for the *i*th respondent

β_{xt} is the effect of the t^{th} product for class x
and effect coding is used for parameter identification:

$$\sum_{m=1}^{M=9} \alpha_{im} = \sum_{t=1}^{T=15} \beta_{xt} = 0 \text{ for each segment x = 1,2,…,K.}$$

Model 2: LC Ordinal Regression with Random Intercept and Continuous Random PRODUCT Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{it}$$
$$\alpha_{im} = \alpha_m + \lambda_{10} F_{i1} + \lambda_{20} F_{i2}$$
$$\beta_{it} = \beta_{t0} + \lambda_{2t} F_{i2}$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$
$$V(\alpha_{im}) = \lambda_{10}{}^2 + \lambda_{20}{}^2$$
$$E(\beta_{it}) = \beta_{t0}$$
$$V(\beta_{it}) = \lambda_{2t}{}^2$$

where: *logit($Y_{im.t}$)* is the adjacent category logit for product t
C-Factor score F_{i1} is associated with the intercept
C-Factor score F_{i2} is associated with the T product effects

and $(F_{i1}, F_{i2}) \sim BVN(0, I)$

Model 3: LC Ordinal Regression with Random Intercept and Discrete Random Product Attribute Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{x1} Z_1 + \beta_{x2} Z_2 + ... + \beta_{xT} Z_Q$$
$$\alpha_{im} = \alpha_m + \lambda F_i$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$
$$V(\alpha_{im}) = \lambda^2$$

where:

logit($Y_{im.t}$) is the adjacent category logit for product *t* with attributes $Z_1, Z_2, \ldots, Z_Q$
β_{xq} is the effect of the *q*th attribute for class *x*

and C-Factor *score* F_{i1} is associated with the intercept

Model 4: Ordinal Regression with Random Intercept and Continuous Random Product Attribute Effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + ... + \beta_{iQ}Z_Q$$

$$\alpha_{im} = \alpha_m + \lambda_{10}F_{i1} + \lambda_{20}F_{i2}$$

$$\beta_{iq} = \beta_{q0} + \lambda_{2q}F_{i2}$$

$$E(\alpha_{im}) = \alpha_m$$

$$V(\alpha_{im}) = \lambda_{10}{}^2 + \lambda_{20}{}^2$$

$$E(\beta_{iq}) = \beta_{q0}$$

$$V(\beta_{iq}) = \lambda_{2q}{}^2$$

where: *logit*$(Y_{im.t})$ is the adjacent category logit for product *t* with attributes $Z_1, Z_2, \ldots, Z_Q$

C-Factor F_{i1} is associated with the intercept

C-Factor F_{i2} is associated with the Q product attribute effects

and $(F_{i1}, F_{i2}) \sim BVN(0, I)$

ASSESSING THE IMPACT OF EMOTIONAL CONNECTIONS

PAUL CURRAN, GREG HEIST,
WAI-KWAN LI, CAMILLE NICITA, BILL THOMAS
GONGOS AND ASSOCIATES, INC.

BACKGROUND:

In today's dynamic competitive environment, more and more brand managers are moving their positioning strategy from traditional value propositions (quality, value, etc.) in favor of more emotionally-charged value propositions. Mega brands such as McDonalds, Cheerios, Pillsbury, and Pepsi are seeking ways to touch consumers' emotions as well as communicate traditional cost-benefit selling propositions.

In our work at Gongos and Associates, we have helped many of our key clients use qualitative research to better understand the emotional ties that consumers have with their brands. In our research, we have found that consumers are motivated to purchase products based on attributes that reinforce a specific "emotional connection." We define "emotional connections" as consumer-identified emotions and values that connect the consumer with a product or brand. These connections represent the personally relevant role the product/brand plays in the life of a consumer.

While qualitative methodologies are certainly valid, it would be very helpful for a brand manager if the impact of these emotional connections could be quantified. This paper attempts to accomplish two objectives: (1) report a choice-based conjoint application that gauges the relevance of various emotional connections for a product category and (2) assess the impact of emotional connections on purchase consideration.

In a typical quantitative survey, respondents would be asked to reflect their opinions on an X-point scale. However, as argued in Gerald Zaltman's thought-provoking book, How Customers Think, up to 95 percent of consumer thinking happens in our subconscious. This begs the question: Does commercial market research rely solely on logical reasoning and rational thought processes to gauge consumer behavior or design products? If yes, to what extent are we missing out on a powerful component of consumer behavior—one's emotional connection to a brand?

This argument is probably especially true for product categories and brands for which consumers express a high degree of emotional intensity. In cases such as these, a traditional quantitative survey, which requires respondents to read the survey questions and provide answers consciously, would not be the best tool in quantifying emotional connections and brand associations. To truly assess consumers' emotional connections to a brand we need to devise a way to collect respondents' opinions while holding conscious thinking to a minimum.

For several years, Gongos and Associates has been using a technique for understanding—qualitatively—consumers' emotional connections to brands. In this approach, we explore and identify relevant emotional connections and interrelationships within a category and for specific brands.

The qualitative approach is generally done through a series of one-on-one, in-depth interviews. A typical IDI lasts approximately two hours. Prior to coming to the interview, respondents are asked to put together a collage of images, pictures and/or words that illustrate their feelings toward an ideal brand/product. The images, pictures and/or words help respondents express the emotional ties more easily as it is often difficult to articulate such feelings. The collage also identifies inter-relationships among the images and aids in the development of qualitative output such as consensus maps.

Using the research tool described above, clients are able to help define the brand by validating current beliefs and creating a future positioning strategy and product development "roadmap." This research also allows us to understand both the corporate and customer perspectives, which can identify realignment opportunities to infuse emotional connections into all marketing activities.

The key in understanding consumers' emotional connections to brands is to make a link between a product/service/brand and the consumer's life. It is not unreasonable to suggest that understanding these emotional links and translating them into tactical actions (e.g. product characteristics, the retail environment, packaging, communications, etc.) significantly increases product success rates, customer satisfaction, and brand loyalty. Furthermore, understanding emotional connections can likely make the difference between a product that looks good on the drawing board but fails on the store shelf. In the end, client organizations are better able to develop products and services that win the heart and mind (and pocketbook) of the consumer.

Still, even with the powerful insight these qualitative sessions provide, it is difficult for brand managers to move forward with marketing outlays without the support of some type of quantitative measures. Quantitatively, corporate-level decision makers often seek to:

- Assess the impact of emotional connections on traditional market research metrics (purchase likelihood, customer satisfaction, brand loyalty, etc.).
- Prioritize the relative importance of various emotional connections.
- Determine which emotional connections are associated with which brands—Which emotional connections does their brand "own" and which are "owned" by the competition?
- Pinpoint an important yet unclaimed emotional connection—identify any "white space" in the market that our client's brand can grab?

Knowing clients' desire for this type of quantitative support, a team of researchers at Gongos and Associates set out to develop a technique that would provide the desired output but stay true to the conceptual framework of Zaltman's work and, by association, our well-established qualitative approach. With that said, we knew that any technique we were to develop would have to follow the following guidelines:

- Use a measurement tool that provides high discrimination but minimal conscious effort on the part of the respondent.

- Use visual stimuli to reflect emotional values, replicating the qualitative process whereby respondents use images and words to represent emotional values.
- Hinder overly-conscious thinking to ensure respondents react to stimuli in an emotionally-charged manner.

The major goal of our quantitative approach is to allow respondents to reflect their responses with minimal conscious effort. In other words, we want them to let us know what they feel—which should only involve minimal conscious thinking. To accomplish this, we rely on three specific techniques:

- Use pictures collected from qualitative research as stimuli to reflect the emotional connections. By eliminating as much verbiage as possible, the conscious thinking should then be minimized.
- Use elements of "negative priming" (Titz, Behrendt, and Hasselhorn, 2002; Janiszewski 1990) to inhibit conscious thinking.
- Use a choice-based conjoint with extremely simple paired-comparison choice tasks so as to minimize the amount of cognitive effort needed to respond.

THE SURVEY INSTRUMENT:

In 2001, Saturn wanted to better understand how its customers connect with their brand and product offerings. The research used our qualitative emotional connections technique described previously. A series of IDIs were completed with Saturn owners as well as owners of competing makes. From this research, we were able to identify the following emotional connections relevant to their brand as well as the product category:

- Success/Accomplished
- Independent/ Self-Reliant
- Comfortable/Relaxed/Happy
- Peace of Mind
- Smart/Practical
- Care for Others/Family
- Fun to Own
- Saturn Family/Community

Given the amount of qualitative research we had already completed, we felt Saturn would be an ideal client for which we could develop our quantitative survey instrument. Armed with the results of our qualitative research and the guidelines we had established, we set out to design a quantitative survey instrument that would meet our research objectives.

A key component to the approach involved the selection of the competitive set. In this case, we selected three competitive, yet distinctive, vehicle brands: Honda, Volkswagen, and, of course, Saturn. Selection of the competitive set is critical since an underlying assumption is that each brand has a certain level of emotional intensity.

After some initial screening questions, our instrument begins by presenting respondents a series of vignettes. These vignettes were, more or less, actually built by the IDI participants during the qualitative phase. Each vignette included a photograph along with a "story" as written/described by the IDI participant. For each emotional value, the respondent identifies the vignette to which they most strongly relate. One respondent-identified vignette is used to represent each emotional value being tested. The images associated with the vignettes become proxies for the specified emotional connection later in the survey. Once the respondent-specific stimuli are established, the respondent is taken through the negative priming exercise and then onto the two choice exercises.

Perhaps the most provocative component of our approach is the negative priming exercise. "Priming" is a facilitated cognitive process that produces a faster reaction time or higher recall rate, due to the cues provided by the "primer." Suppose, for example, you are asked to pick out breakfast items from the following list of words: butter, bread, paper, orange juice, phone, and computer. To prime you, before entering the lab you are kept in a waiting room that is infused with the scent of freshly baked pastry. While you are waiting your mind is focusing on all these bakery items, such as bread and butter. In theory, your reaction time to the experimental task would be much faster than that of a respondent kept waiting in a scent-free environment. The discrepancy between the unprimed and the primed respondent would be referred to as a "priming effect."

Negative priming refers to a slowed response to visual stimuli that is initially ignored and later presented as the target. This decrement in performance is due to the interference of the irrelevant stimulus on the processing of the relevant stimulus. While some argue that the interference slows the processing of relevant stimulus, this is actually what we seek to accomplish. That is to say, we show irrelevant pictures to interfere with respondents' ability to cognitively process the relevant pictures (those the respondent associates with the emotional connections).

Negative priming can then be considered a suppressed cognitive process, and, while its exact impact has had mixed assessments in the academic literature, there are some established mechanisms for generating the desired distraction/interference effect. Consistent with the work of Titz, Behrendt, and Hasselhorn (2002), we developed a negative priming exercise administered prior to each choice exercise. Respondents were challenged to remember various innocuous details from a set of 12 pictures. Respondents were told there would be quiz at the end of the survey in which they would be asked to recall as many of the details as possible. In theory, respondents would be distracted by the irrelevant pictures, which should effectively suppress them from thinking too much about the stimuli presented in the choice tasks. In addition, in their attempt to remember all the details from these irrelevant pictures, respondents would become mentally overloaded—leaving limited mental capacity directed towards the stimuli presented in the choice tasks task. In this case, negative priming is used to suppress unwanted conscious processing of the stimuli representing the emotional connections.

Key to our research was the use of paired comparisons in the choice tasks. Paired comparisons are often used in psychometric and econometric research studies. In 2003, Bryan Orme presented a paper in which he compares monadic ratings and paired comparisons. In this paper, Orme establishes how paired comparisons allow respondents to better discriminate among many similar items. Likewise, Burton (2003) describes various categories of paired comparisons and makes an argument for the use of a type of factorial design know as incomplete cyclic designs (ICDs). In testing his theories on ICDs, Burton finds that semantic similarity judgments can be collected with high reliability when using efficient designs and robust sample sizes. These bodies of work lend support to our technique in that we use highly efficient experimental designs and choice tasks that provide considerable discrimination.

Our first set of paired comparisons explores the "importance' dimension. Here we present respondents with a choice of two emotional connections and ask: How do you want to feel about a vehicle? After completing the first choice exercise, respondents are re-primed.

To measure the second dimension—brand association—each respondent is randomly assigned a brand from the competitive set. Respondents are then asked: Which image do you associate more with brand X? (Where X is either Saturn, Honda, or Volkswagen.)

OUR TEST

As noted previously, our primary objective was to develop a quantitative survey instrument to assess the impact of emotional connections previously uncovered in a series of qualitative interviews. It is important to note that the process we lay out should be considered a follow-up to thorough qualitative research and not a "stand-alone" approach.

The survey was administered over the Internet. We acknowledge that the Internet may not be the ideal methodology to administer this type of study compared to a more controlled central location test. However, given this is our first attempt to quantify emotional connections in such a non-traditional manner, we felt the Internet would be the most viable methodology.

The survey instrument was programmed using Gongos*Online* web tools, a set of proprietary online research tools developed at Gongos and Associates.

Respondents were recruited from Survey Sampling International's SurveySpot online consumer panel. SSI maintains an automotive sub-panel from which highly targeted automotive samples can be drawn. A total of 1,662 Honda, Saturn, and Volkswagen owners participated in this study. Respondents were screened in such a way as to match the demographic composition of the qualitative phase. To this end, our final sample had the following characteristics:

- 831 owners of a new, not previously owned Honda
- 529 owners of a new, not previously owned Saturn
- 302 owners of a new, not previously owned Volkswagen
- All respondents were the primary driver and decision maker of their qualifying vehicle.

- Qualified respondents were screened to ensure they were not employed in/affiliated with the automotive or marketing services industries.

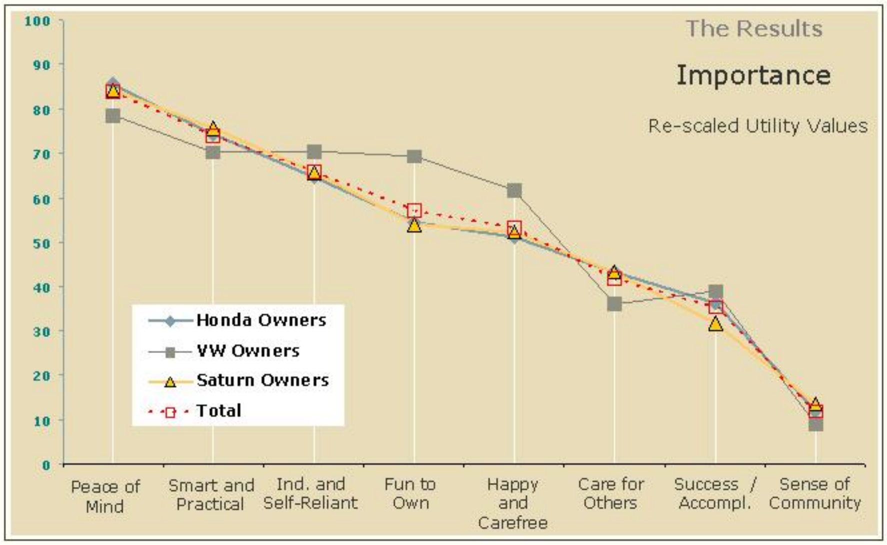
On average, the survey took respondents 16 minutes to complete.

Sawtooth Software's CBC System for choice-based conjoint was used to create the experimental design. Using CBC, the study specifications called for a two-attribute study where attribute 1 was the eight emotional connections (eight levels) and attribute 2 was a dummy attribute not shown to respondents (two levels). We use the CBC paper-and-pencil module to output the design and then complete the programming using the GongosOnline web tools. To ensure a highly efficient design, a total of seven versions of the paper-pencil study were exported from the CBC system.

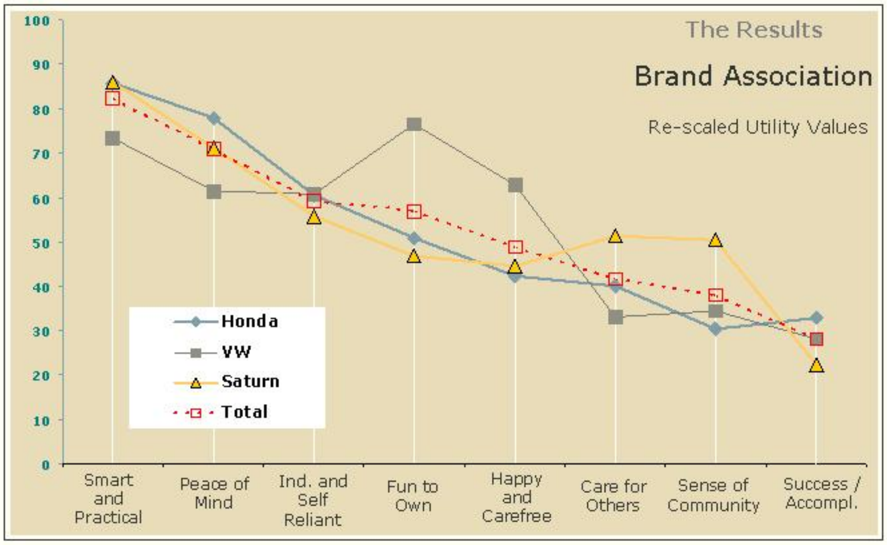
KEY FINDINGS

Using Sawtooth Software's Hierarchical Bayes system, individual-level utilities were estimated for each of the two choice exercises, providing us a set of part-worths for each of the two dimensions we explored—importance and brand association. Using these utilities, we completed four separate analyses to assess the impact of various emotional connections. Given the nature of the analyses selected for this research, we rescaled both sets of utilities such that each respondent's lowest rated emotional connection received a value of 1 and the highest value received a value of 100.

From the importance utilities, we are able to assess the relevance of eight emotional connections for automobiles among our client's target market. The emotional connection "peace of mind" is shown to have the most impact while "sense of community" has the least. These findings confirm what was hypothesized in the qualitative phase.

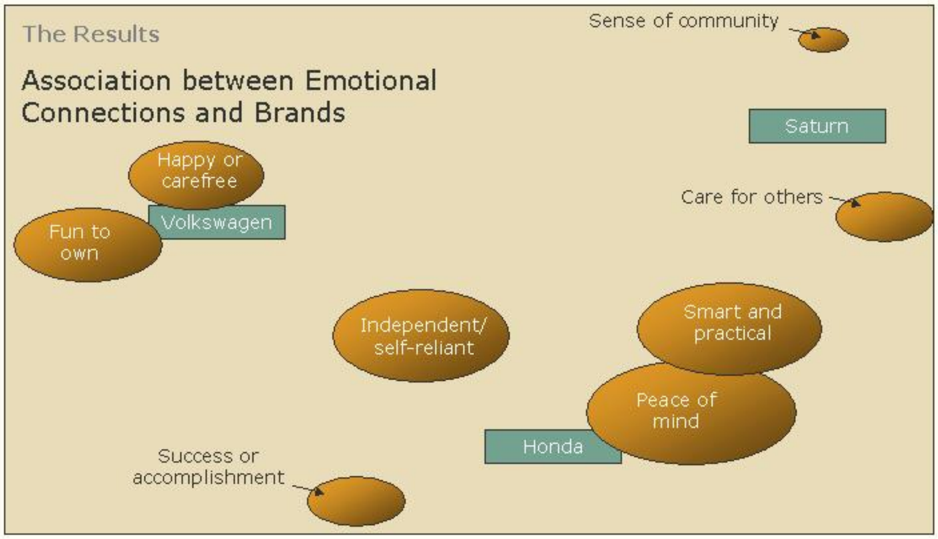


Using the brand association utilities, we were able to assess the differentiating relevance of the eight emotional connections for each brand. We found Volkswagen to be highly associated with "fun to own" while Honda is associated more with "peace of mind." This analysis also paints the competitive landscape by revealing the "gaps" across brands on each emotional connection. We observed a big "gap" for "fun to own," but a much narrower "gap" for "independent and self-reliant."



Our third analysis, also from the brand association utilities, is a correspondence analysis. Using correspondence analysis we are able to build a perceptual map to help pinpoint which brand "owns" which emotional connection. While Volkswagen owns "fun to own," Honda owns "peace of mind." In addition, the results also reveal the unoccupied market niche, should a brand wish to reposition to gain better customer appeal and/or avoid competition. We are also able to incorporate the importance dimension into the map by varying the size of the elements on the perceptual map (where the size of the bubble reflects the importance of the emotional connection).

The perceptual map shows which brands are associated with which emotional connections; the size of the bubble reflects the relative importance of the value



Our final analysis uses a metric we call "emotional congruence." Emotional congruence refers to the extent to which the brand satisfies one's emotional needs. Emotional congruence is a derived measure based on the gap between a desired emotional connection (importance utilities) and the perceived emotional associations with the brand (association utilities). A match between the respondents' preferred emotional connection and the perceived brand-specific emotional connection would result in a high emotional congruence score.

By correlating emotional congruence to purchase consideration of a brand, we can measure the impact to which emotional connections drive brand consideration. In our case, we observed significant moderate correlations between emotional congruence and purchase consideration (0.36 – 0.44), suggesting that emotional connections do, in fact, drive purchase consideration.

We were quite thrilled that each of the analyses provided our client a sound quantitative measure upon which they could build their brand strategy. Importantly too, we rely on the various analyses to triangulate the findings as we sought robust external validity for both the methodology and the results.

CONCLUSION

Traditional conjoint/choice research largely focuses on pragmatic, concrete product attributes (e.g., price or gas-mileage), although it is well documented that consumers' choices and decisions are often affected by "soft" elements such as emotional connections. Emotional connections and other similar soft elements are difficult to quantify, and thus researchers tend to rely on qualitative methodologies to gather insights on such dimensions.

Through our work, we have demonstrated a unique application of choice-based conjoint and Hierarchical Bayes that allows one to gauge the relevance of various emotional connections for a product category. Further, by adapting this approach as a follow-up to a robust qualitative study, we show how the impact of emotional connections on purchase consideration can be established.

To that end, we feel we have been successful in:

- The development of a unique methodology for measuring soft elements, such as emotional connections, that impact purchase consideration.
- Outlining a plan of analysis to assess the relevance of various emotional connections and the extent to which emotional connections shape the competitive landscape.

REFERENCES

Janiszewski, C. (1993) "Preattentive Mere Exposure Effects," Journal of Consumer Research, 20, p. 376-392.

Burton, Michael L. (2003) "Too Many Questions: The Uses of Incomplete Designs for Paired Comparisons," Field Methods, Sage Publications

Orme, Bryan (2003) "Scaling Multiple Items: Monadic Ratings versus Paired Comparisons," Sawtooth Software Proceedings, Sequim: Sawtooth Software.

Zaltman, Gerald and Robin Hige Coulter (1995) "Seeing the voice of the customer: Metaphor-based advertising research" *Journal of Advertising Research*; New York; July/August 1995.

Zaltman, Gerald (1997) "Rethinking market research: Putting people back in." Journal of Marketing Research (JMR), November 1997.

MCMC METHODS

ITEM RESPONSE THEORY (IRT) MODELS: BASICS AND MARKETING APPLICATIONS

LYND BACON, JEAN DURALL
LYND BACON & ASSOCIATES LTD.
PETER LENK
UNIVERSITY OF MICHIGAN

Item Response Theory (IRT), also called *latent trait models*, is a class of psychometric measurement models that were developed for standardized testing in education. In their simplest and most conventional form, they estimate continuous latent variables[1] from observed binary or ordinal responses to test or survey items. Several alternative methods are available for inferring subject characteristics from multiple responses. However, the strength of IRT resides in simultaneously estimating characteristics of both the respondents and the items. The former is the inferential goal of marketing studies and usually receives the most attention. Methods for investigating the properties of survey items are less rich and are often disconnected from the inferential model. IRT models unify the two in a single framework, which offers distinctive advantages to marketing researchers. In this paper we will review the simplest IRT models, illustrate their application to marketing research, and delineate their relation to similar, widely-used marketing research models.

STICKING WITH YOUR BRAND

Consider the following hypothetical situation. You have a research client that wants you to develop a scale based on a set of binary response items that she has collected data on. She is convinced that some combination of these responses can measure an underlying characteristic of her customers that has important marketing implications. She calls this unobservable characteristic “brand adhesion” because she thinks of it as the stickiness of the relationship between a customer and a brand. How should you proceed to construct the scale that your client wants?

There are of course at least a couple of time-honored and relatively well known approaches you could use. You could contemplate just adding the responses across the items to get a total score for each respondent, but without doing anything else you’d be simply assuming that the items in combination form an internally consistent and reliable scale. In order to address the reliability issue, you could do conventional item reliability analysis using Cronbach’s Alpha coefficient and item-total correlations. One weakness of this approach is that your correlational analysis is based on binary data. Another is that you are assuming that the reliability of your measure is constant across the range of possible score values. This needn’t be the case, and in fact it can be very useful to know whether this assumption is tenable. We’ll consider this issue of constant reliability further in what follows, below.

[1] A history of IRT models is available at www.uic.edu/classes/ot/ot540/history.html

Another approach you could take is to use principal components analysis, or common factor analysis. These techniques are applied to the correlation or covariance matrix of the data. However, because the data are binary, there are restrictions on the correlations or covariances, which are functions of the marginal distributions of the binary responses. Then there is still the issue of assuming that the reliability of the scale you'd compute is constant across the range of possible score values.[2]

But isn't this always the case, that reliability is constant? Absolutely not! It depends on how you model your scale. An in fact, a particular strength of IRT models is that you can choose *how* reliability is distributed over your scale when using IRT models for scale development.

The simplest IRT model you could use in addressing your client's request is a *Rasch,* or one parameter (1P), IRT model. The core of the one parameter IRT model relates the probability, p, of observing a "correct" response from respondent "i" to an item "j" to an unobserved score on the unobserved trait.

Suppose now that you have J binary survey or test items with data coded on them such that a "1" means a positive response (e.g. "correct," "yes," or "true") and a "0" means a negative response ("incorrect," "no," or "false"). A one parameter IRT model for these data can be expressed as the set of relationships between the observed responses $\boldsymbol{X_{ij}}$ and model parameters as:

$$p(x_{ij} = 1 | \theta_i, \alpha_j) = F(\theta_i - \alpha_j) \tag{1}$$

The function **F** in Equation (1) is traditionally either a logistic or cumulative normal density function for each of the J items.[3] $\boldsymbol{\theta_i}$ is the respondent score on the underlying trait, and $\boldsymbol{\alpha_j}$ is the item-specific parameter that is usually called the item "difficulty" parameter. The reason for this name is that for any given, fixed respondent score $\boldsymbol{\theta_i} = \boldsymbol{\theta}$ *****, the larger an item's $\boldsymbol{\alpha_j}$ becomes, the smaller the probability is that respondents will answer it correctly/affirmatively. This parameter acts as a cut-off or threshold parameter in the sense that the probability of a positive response is greater than 0.5 when $\boldsymbol{\theta_i > \alpha_j}$, and it is less than 0.5 when $\boldsymbol{\theta_i < \alpha_j}$.

Note that the subject parameters $\boldsymbol{\theta_i}$ and the item parameters $\boldsymbol{\alpha_j}$ are measured in terms of the same units of measurement, the units of the latent trait. This is one of the defining characteristics of IRT models.

[2] Space prohibits detailed discussion of the use of item reliability analysis and factor analysis for scale construction. If you are interested you should see Bollen (1989) for an in-depth, traditional coverage of these and many other latent variable modeling issues. For a recent comparison of factor models and IRT models applied to ordinal data, see Moustaki, Jőreskog & Mavridis (2004).

[3] This is in fact the functional form of the most common kind of IRT model, the cumulative model. There have been many other forms defined. For a review of different cumulative IRT model forms, see Heinen (1996).

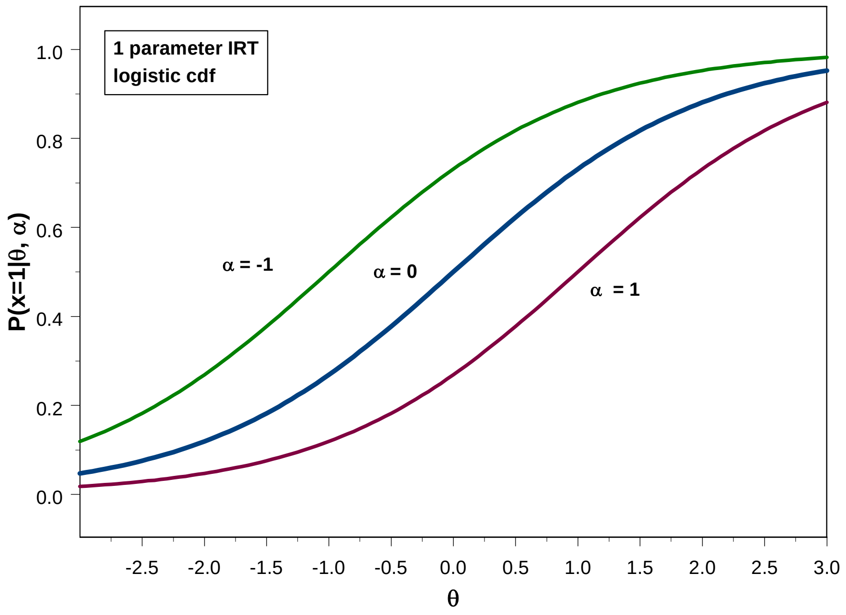


Figure 1. Item characteristic curves (ICCs) for three, one parameter (1P) IRT model items. Abscissa ("θ") is the value on the latent trait. Ordinate is the probability of a "correct" response given the value of θ, and also α, the item "difficulty," or threshold, parameter. ICCs for three different values of α are shown.

IRT models are sometimes visualized in terms of *item response functions,* or *item characteristic curves* (ICCs). Figure 1 shows ICCs for three items that differ in their level of difficulty. In this graph, the value of the latent trait is on the abscissa (labeled $\boldsymbol{\theta}$), and the probability of a positive (correct or true) response to an item as a function of $\boldsymbol{\theta}$ is on the ordinate.[4] The $\boldsymbol{\alpha_j}$ values for the items shown are -1, 0, and +1.

Another defining characteristic of IRT models is the notion of *local* scale reliability. Figure 2 shows the *item information functions* (IIF) for the three items whose ICCs are in Figure 1. These functions are the Fisher's information functions and indicate how well each item discriminates between high and low values of $\boldsymbol{\theta_I}$ as a function of location on the latent trait. Larger values of IIF indicate greater degrees of discrimination. The maximum discrimination a particular item can afford is at the location of the item's difficulty parameter, $\boldsymbol{\alpha_j}$. Information defined in this manner is analogous to measurement reliability, but in this case it is reliability as a function of item location on the underlying trait. The further away you get on the latent trait from the value of an item's difficulty parameter, the less well that item discriminates between values of the latent trait for different subjects.

[4] The abscissa is the "x" axis, and the ordinate is the "y" axis.

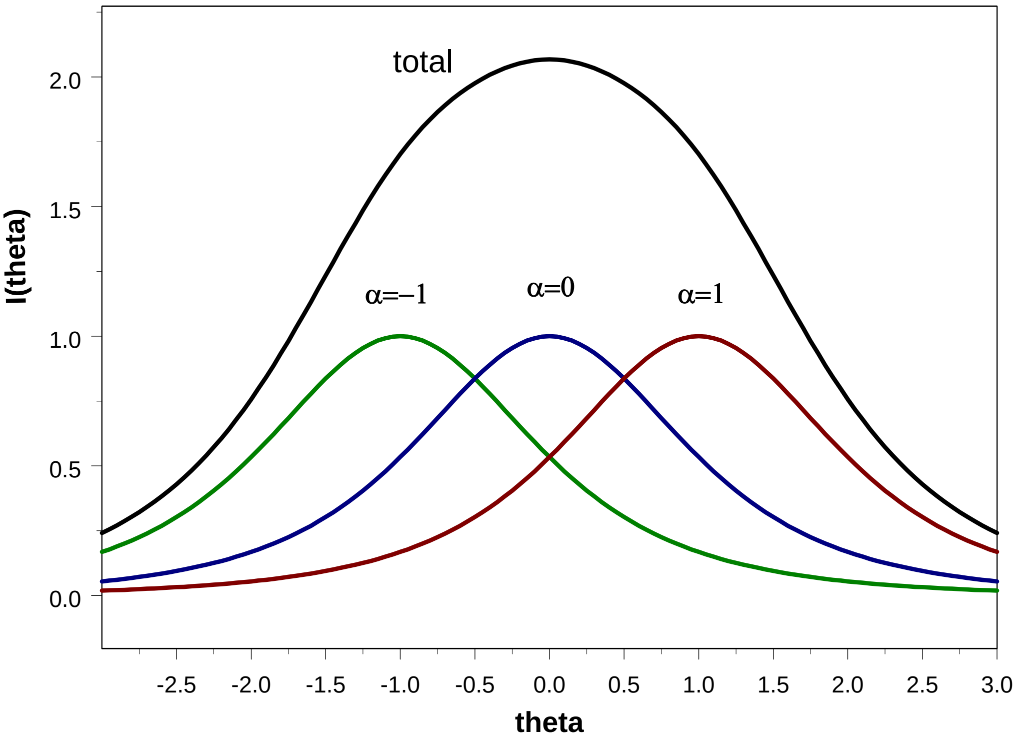


Figure 2. Item and total information functions for 1P model items shown in Figure 1. Abscissa is value on the latent trait. Ordinate is Fisher's information.

An IRT model's total information is the sum of its individual item information functions. You'll see an example of a total information function in Figure 2. In this graph, the value on the latent variable is the abscissa, and information is a function of the latent trait. Again, noting that the value (as indicated by the height) of this information function indicates local reliability, you'll observe that by combining these three particular items into a single scale, you've concentrated your measure's reliability in the middle of the range of possible values on the latent trait.[5] As a result, your three-item scale is most reliable at distinguishing differences in θ_I 's that are located close to θ_i = 0, the mode of the total IIF, and less reliable as you move further away from zero. If all three difficulty parameters were negative, then the IIF's mode would be less than zero, and if all three difficulty parameters were positive, then the IIF's mode would be positive.

There is no reason why you must have a scale with a total information function as in Figure 2, however. If you happen to have a set of items available that span a range of difficulty parameter values, you can select a subset of them that gives you the information function that you want. For example, you can increase or decrease the reliability of your measure above or below θ_i= 0. If your available items can't provide

[5] Possible values on the trait are real numbers between minus infinity and infinity.

you with the kind of information function you want, you can develop and test more. This flexibility in scale construction is a hallmark of the IRT measurement approach.

IRT models, like other kinds of models, do require some assumptions. An important one is the assumption of *local independence*[6]. The idea is that observed responses are only related through the unobserved trait, and so once the trait is taken into account, the observed responses are independent. You can investigate how tenable this assumption is for any particular model and data set.

Another assumption is the notion of *item independence*[7]. This is the assumption that you would estimate the same value of $\boldsymbol{\theta_I}$ for respondent "i" by using different subsets of a pool of items specifically developed to measure the latent trait of interest. It is the tenability of this assumption that enables to use alternate forms of standardized tests.

It is important to distinguish between local independence and item independence. Local independence allows the researcher to combine items in such a way as to improve his or her ability to estimate latent traits in various ranges of the scale. At a technical level, it is the mathematical property that allows the total information function to be the sum of item-specific information functions. It also allows you to use standard, maximum likelihood or hierarchical Bayes software to estimate the model without additional contortions. In contrast, item independence is a more fundamental concept in scale design. Item independence means each item in a set of items measures the same latent trait and only that latent trait in the simple Rasch model. If local independence does not hold, then a clever statistician could rewrite the estimation code to accommodate correlated data. If item independence does not hold, then the survey items are poorly constructed, and the validity of the survey is questionable if one is using single-latent trait IRT models: one should then consider multiple trait IRT models as presented below.

Several different methods have been used for estimating the parameters of IRT models. These include conditional and restricted maximum likelihood procedures. More recently, IRT modelers have been using Markov Chain Monte Carlo (MCMC) methods to fit full probability, hierarchical Bayes models. Our preference is to use this more contemporary approach to estimating IRT models. In the simple examples that follow, this is the estimation method we use. Readers interested in traditional estimation procedures may want to look into the references and resources available on www.edres.org, and www.rasch.org.

Now, turning back to your client's "Brand Adhesion" scaling problem, suppose further that she has already collected data from 560 customers on each of the following four "true/false" questions:

1. *"I regularly buy Brand X" (X is your client's brand)*
2. *"I buy X even when other brands are more expensive."*

[6] "Local independence" is also known as "conditional independence" in the Bayesian community or "independent observations" in the maximum likelihood community. This property allows one to add individual IIF to obtain the total IIF.

[7] "Item independence" is a different concept than "independence" in the probabilistic sense, as in the previous footnote. "Item independence" refers to the construction of the items and what you intend for them to measure. For example, you may have a bank of items to measure "Verbal Comprehension" and a bank of item to measure "Analytical Reasoning." In this case, a test designer could not substitute items from the first bank for the second.

3. *“If the store I regularly shop in were to be out of Brand X, I’d go to another store before buying a different brand.”*
4. *“I’d borrow money to be able to buy Brand X.”*

One thing you’ll note about these questions is that they imply a sort of “escalation” of behaviors and commitment in regard to Brand X. Responding “true” to question 4 clearly requires more of a commitment to Brand X. IRT models are particularly appropriate when you have items logically related in this way.

$$\textbf{logit}(\textbf{p}(\textbf{x}_{\textbf{ij}} = 1)) = \theta_i - \alpha_j$$

$$\textbf{x}_{\textbf{ij}} \sim \textbf{bernoulli}(\textbf{p}(\textbf{x}_{\textbf{ij}} = 1))$$

i=1, 560 (N)

j=1,4 binary items

$$\theta_{\mathrm{i}} \sim \boldsymbol{N}(0,1)$$

$$\alpha_j \sim \boldsymbol{N}(0,0.001) \quad \leftarrow \textbf{ Normal(mean,precision)}$$

<u>Figure 3.</u> Specification for Brand Adhesion 1P IRT model. In the specifications of the prior distributions for theta and alpha, the second parameter is the precision, which is 1/variance. The sample size is N=560.

Figure 3 gives the details of a one parameter IRT model for your client’s data. You have i = 560 subjects, and j = 4 binary response items. *F* in equation (1) is the logistic function. The latent trait θ_i can be viewed in two ways: as a random effect or as a subject-specific parameter that varies across the population. We will specify its distribution of heterogeneity as a normal distribution with mean zero and precision[8] one. The reason for this specification is that the latent trait has undefined units. Setting the mean of the normal distribution to zero fixes the origin for the latent trait, and setting the variance to one fixes the spread. Without these or other assumptions, the Rasch model would not be identified. As part of fully specifying this model in the Bayesian tradition, we have specified fairly noninformative, normal prior distributions for the item difficulty, or *threshold*[9], parameters α_j.

Supposing now that we picked some reasonable initial values for all of the parameters of this model, we ran our MCMC procedure for 10,000 iterations, and then we “thinned” our chains of sampled values by taking every 3rd value from an additional 3,000 draws. These thinned chains are our approximations to the posterior densities of our item and respondent parameters.

[8] Precision is one over the variance. The Bayesian literature often uses precision instead of variances because in some simple models, the posterior precision is the sum of the prior precision and the precision of the maximum likelihood estimator.

[9] “Threshold” is a more appropriate name for the difficulty parameter in most marketing applications, and so we will use it in all that follows.

Threshold estimates	Mean	2.5%	97.5%
Borrow money	1.061	0.879	1.223
Go to different store	0.810	0.637	0.956
More expensive	0.348	0.205	0.516
Buy regularly	-2.220	-2.487	-19.85

Table 1. Estimates of threshold (difficulty, alpha) parameters for Brand Adhesion 1P IRT model example. 10,000 burn-in iterations, followed by every 3rd of 3,000 additional draws. Means are posterior means. 2.5% and 97.5% are percentile scores from the thinned chain values for each. -2LL at posterior means = 2538.3. DIC=2543.7, pD=5.46.

Table 1 gives the means of the thinned chains for each of the four item's threshold parameters, and also the 2.5 percentile and 97.5 percentile values from the corresponding chain.[10] At the bottom of Table 1 are a few statistics that summarize the model's overall fit to the data: -2 times the log likelihood, or the *deviance,* evaluated at the posterior means is used like the deviance in non-Bayesian models where the smaller the better, ceteris paribus.

"DIC," the Deviance Information Criterion, is analogous to the Akaike Information Criterion (AIC), which also penalizes models that are more complicated. DIC is computed as the difference between two times the average deviance (D_{ave}) over the sampled chain minus the deviance calculated using the posterior means (D_{post}). pD is an estimate of the effective number of parameters for the model. You can think of it as summarizing the dimensionality of the model parameter space. pD is calculated as $D_{ave} - D_{post}$.[11]

You should note something about the pattern of the mean thresholds in Table 1. They order the items on the underlying latent trait in a logical way. Items that should signify a low level of "brand adhesion," buy the brand regularly," for example, have low thresholds, meaning that respondents overall are more likely to endorse these items. Items that should correspond to a high degree of brand adhesion, like "borrow money to buy X," have higher threshold values, meaning that relatively fewer respondents endorsed them. Although the ordering of the items in terms of their (mean) thresholds makes sense, there was nothing in our model that forced this ordering.

[10] These results and all subsequent results were obtained using WinBUGS, a program for using MCMC methods that is in the public domain. See References for additional information on WinBUGS.

[11] If you are interested in learning more about summary model fit measures and posterior predictive checking, Gelman et al.'s (2003) Chapter 6 provides good coverage. Note that we've given very short shrift to critical issues in model fitting here due to space limitations.

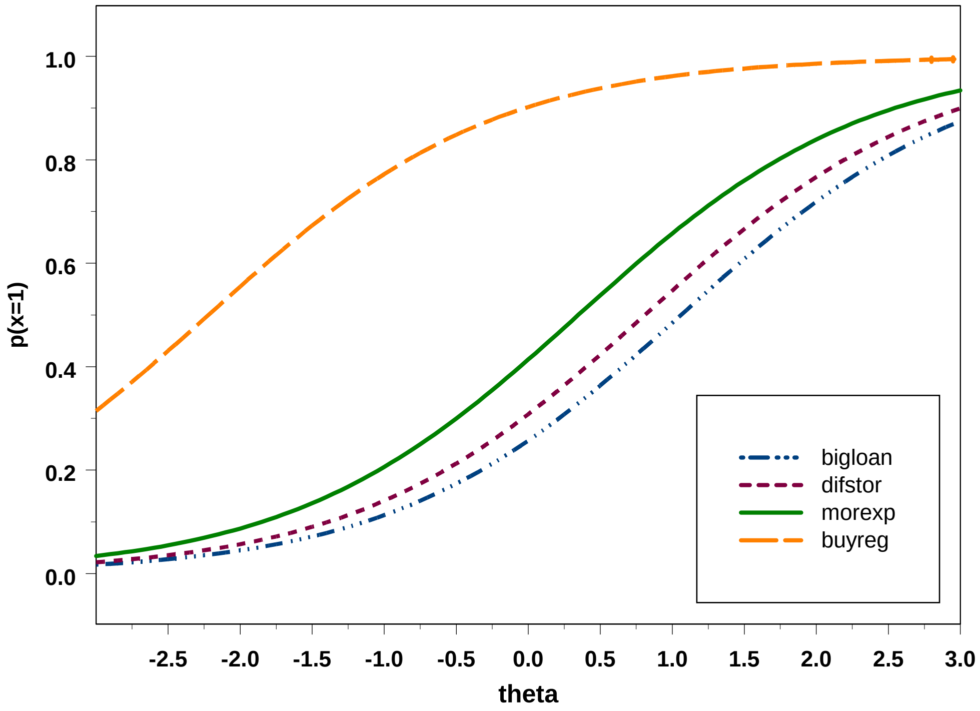


Figure 4. Estimated ICCs for Adhesion 1P model. Abscissa is value on the latent trait. Ordinate is the predicted probability of a positive response (i.e. an endorsement of an item statement) as a function of posterior means of the estimates of θ an α.

Figure 4 shows the ICCs for our one parameter model. Strictly speaking, these are ICCs based on the posterior means of the MCMC chains for the items' threshold parameters. You can see in Figure 4 that the ICCs of the items are ordered left to right in the way that you'd expect given the posterior means of their threshold values. This graph makes it obvious that the items are not equally spaced out along the latent trait, and that for these respondents the rate with which they endorsed the "buy regularly" item was quite high relative to the other three items, which differed from each other a lot less in this respect.

Upping your adhesion ante through increased model complexity

Upon seeing your one parameter model results, your client asked whether her four questions were all equally good indicators of adhesion, or whether perhaps some were more informative than others.

Like any good modeler, your inclination is to apply a more complicated model, hoping for more billable hours and creating client lock-in for your services. Fortunately for you, an IRT model is readily available to you that will meet your needs:

$$p(x_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = F(\beta_j \theta_i - \alpha_j) \quad (2)$$

Equation (2) shows what is called the “two parameter” (2P) model because it has two parameters for each binary item. If you compare Equation (2) to Equation (1) you’ll see that we added an additional parameter for each item’s model, β_j, which is constrained to be positive. Traditionally, this parameter is called an item’s *discrimination* parameter, since the larger it is, the more accurately the item distinguishes between respondents with θ_i greater than the value of the item’s threshold parameter α_j, from those whose θ_i are less than α_j. As β_j gets larger, the item’s ICC function looks more and more like a “step” function. Note that since β_j is item-specific, different items can have different values of β_j, and they can vary in terms of how well they discriminate between high and low θ_i.

Returning to your client, we now fit the 2P model, summarized in Figure 5, to her data, using the procedure we describe earlier. Table 2 and Figure 6 summarize the estimated parameters. In Table 2 the β_j are constrained to be nonnegative by the specification of truncated normal priors, “normal(0,0001)I(0,)” in WinBugs notation, in the model statement of Figure 5.

$$\textbf{logit(p(x}_{\textbf{ij}} = 1\textbf{))} = \beta_j(\theta_i - \alpha_j)$$

$$\textbf{x}_{\textbf{ij}} \sim \textbf{bernoulli(p(x}_{\textbf{ij}} = 1\textbf{))}$$

$$\textbf{i=1, 560 (N)}$$

$$\textbf{j=1,4 binary items}$$

$$\theta_{\mathrm{i}} \sim N(0,1)$$

$$\alpha_j \sim N(0,0.0001)$$

$$\beta_j \sim \textbf{\textit{normal}}(0,0.0001)\ \textbf{I(0,)}$$

Figure 5. Specification for Brand Adhesion 2P IRT model. N=560 (same data as for the 1P model). The discrimination parameters (betas) have truncated normal priors so that they are nonnegative.

	Threshold (α)			**Slope (β)**		
Variable:	Post. Mean	2.5%	97.5%	Post. Mean	2.5%	97.5%
Borrow money	1.527	1.163	1.912	6.626	1.646	17.000
Go to different store	1.019	0.718	1.007	1.255	0.268	3.275
More expensive	0.546	0.250	0.844	0.040	-0.008	0.130
Buy regularly	-3.092	-3.038	-2.406	6.908	2.847	15.680

Table 2. Estimates of threshold (difficulty, alpha) parameters for Brand Adhesion 2P IRT model example. 10,000 burn-in iterations, followed by every 3rd of 3,000 additional draws. Means are posterior means. 2.5% and 97.5% are percentile scores from the thinned chain values for each. -2LL at posterior means = 2276.8. DIC=2471.3. pD=194.6.

Table 2 gives posterior means for each item's threshold parameter α_j and discrimination parameter $\boldsymbol{\beta}_\mathbf{j}$. You'll see that the $\boldsymbol{\beta}_\mathbf{j}$ differ quite a bit across the items, and that the $\boldsymbol{\beta}_\mathbf{j}$ for the "I buy it when it's more expensive" item doesn't seem to be different from zero. This means that this item isn't telling you anything reliable about respondents' $\boldsymbol{\theta_i}$. The $\boldsymbol{\beta_j}$ are a lot like loadings in factor analysis in the sense that they tell how strongly related a particular item is to the latent trait.

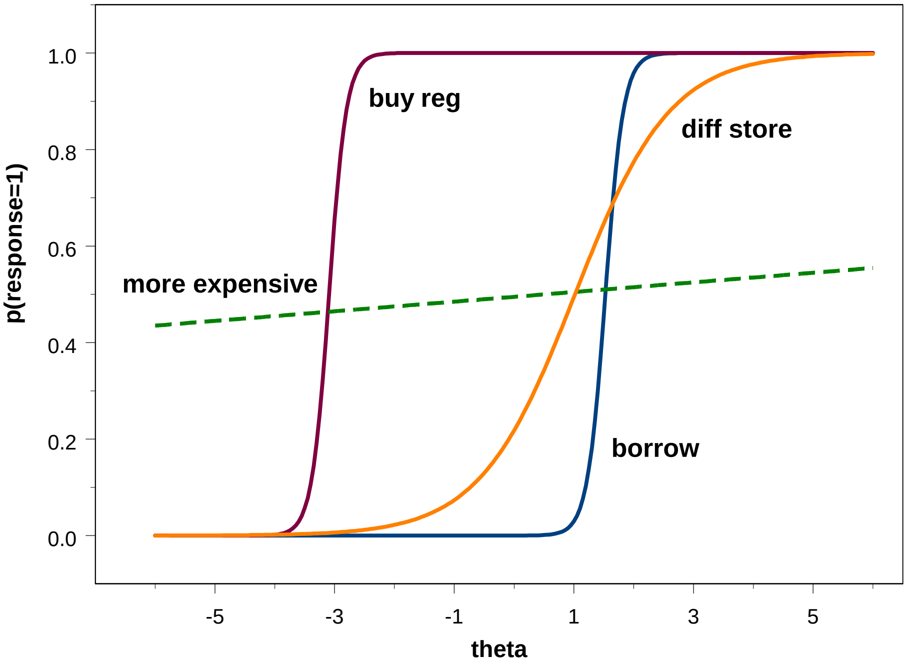


Figure 6. Estimated ICCs for Brand Adhesion 2P model. Details are as in Figure 4.

Figure 6 shows the ICCs for the four items and highlights the considerable differences among them. You can see that the "more expensive" item is essentially useless in terms of the model you've developed for your client, and that two of the items, "buy regularly" and "borrow money," are particularly strong discriminators among respondents with latent trait in the three regions: less than -3.5, between -3.5 and 1, and greater than one. That is, if your latent trait is zero, you will almost certainly respond "Yes" to "buy regularly" and "No" to "borrow money." On the other hand, there is about a 40% chance of you responding "Yes" to "different store."

RUM it up with IRT

The two-parameter IRT/HB model is strongly related to hierarchical Bayes, choice based conjoint models (CBC/HB™ Sawtooth Software trademark.) that are derived from McFadden's (1974) random utility theory (RUM). The development of the 2P IRT model assumes that there is a latent variable for subject i's ability to answer item j correctly:

(3) $Y^*_{ij} = \beta_j\ \theta_i + \varepsilon^*_{ij}$ for subject i and item j.

where θ_i is a measure of the subject specific latent trait; β_j is the item-specific discrimination parameter that is positive, and ε^*_{ij} is a random error term. As mentioned above, the model is normalized so that the average of θ_i across subjects is zero and its variance across subjects is one. For item j, this latent trait is moderated by the item's discrimination parameter.

IRT is based on a cut-point model. Subject i "correctly" responds to item j (X_{ij} =

1) if $Y^*_{ij} > \alpha_j$, and "incorrectly" otherwise. Assuming that the Y^*_{ij} has a logistic distribution given by Equation (4):

$$P(Y^*_{ij} \le y) = \frac{\exp\left(y - \beta_j\theta_i\right)}{1 + \exp\left(y - \beta_j\theta_i\right)} \tag{4}$$

then the probability of a correct response is the two parameter IRT model:

$$P(X_{ij} = 1) = P(Y^*_{ij} > \alpha_j) = \frac{\exp\left(\beta_j\theta_i - \alpha_j\right)}{1 + \exp\left(\beta_j\theta_i - \alpha_j\right)} \tag{5}$$

As test items increase in difficulty, the cut-point or difficulty parameter α_j will become larger. The difficulty parameter controls the location of the logistic curve, and the discrimination parameter β_j controls its shape. Items with larger discrimination parameters provide sharper information about the value of a subject's latent trait. As we previously mentioned, the item characteristic functions and item information functions depend on these item-specific parameters.

IRT can be related directly to RUM by introducing two latent variables:

$Y_{ij} = \beta_j\theta_i + \varepsilon_{ij}$ for subject i and item j
$Z_{ij} = \alpha_j + \delta_{ij}$ for subject i and item j.

The first is a measure of subject's i latent ability for item j, and the second is the difficulty of item j. Instead of logistic distribution, McFadden's (1974) standard RUM assumes that ε_{ij} and δ_{ij} are independent with extreme–value distributions. Then the distribution of the two, latent variables are:

$$P(Y_{ij} \le y) = \exp\left[-\exp\left(-\left\{y - \beta_j\theta_i\right\}\right)\right] \text{and } P\left(Z_{ij} \le z\right) = \exp\left[-\exp\left(-\left\{z - \alpha_j\right\}\right)\right],$$

with densities (first derivatives of the above cumulative probability functions):

$$\begin{aligned} f_{Y_{ij}}(y) &= \exp\left[-\exp\left(-\left\{y - \beta_j\theta_i\right\}\right)\right]\exp\left[-\left(y - \beta_j\theta_i\right)\right] \\ f_{Z_{ij}}(z) &= \exp\left[-\exp\left(-\left\{z - \alpha_j\right\}\right)\right]\exp\left[-\left(z - \alpha_j\right)\right] \end{aligned} .$$

Subject's i response to item j is "correct" if $Y_{ij} > Z_{ij}$, then one again obtains the two parameter IRT model:

$$P\left(X_{ij}=1\right)=P(Y_{ij}>Z_{ij})=\int_{-\infty}^{\infty}\int_{-\infty}^{y} f_{Z_{ij}}(z)\,dz\;\; f_{Y_{ij}}(y)\,dy$$

$$=\int_{-\infty}^{\infty}\exp\left[-\exp\left(-\left\{y-\alpha_j\right\}\right)\right]\exp\left[-\exp\left(-\left\{y-\beta_j\theta_i\right\}\right)\right]\exp\left[-\left(y-\beta_j\theta_i\right)\right]dy$$

$$=\exp\left(\beta_j\theta_i\right)\int_{-\infty}^{\infty}\exp\left[-\exp\left(-y\right)\left\{\exp\left(\alpha_j\right)+\exp\left(\beta_j\theta_i\right)\right\}\right]\exp\left(-y\right)dy$$

$$=\frac{\exp\left(\beta_j\theta_i\right)}{\exp\left(\alpha_j\right)+\exp\left(\beta_j\theta_i\right)}.$$

In the two-option framework, the condition for a "correct" response in RUM can be rewritten as:

$$Y_{ij}>Z_{ij}\Leftrightarrow\beta_j\theta_i+\varepsilon_{ij}-\delta_{ij}>\alpha_j.$$

One can show that the difference in the error terms, assuming independent, extreme-value distributions, has a logistic distribution. Hence, defining $\boldsymbol{Y^*_{ij}=Y_{ij}-\delta_{ij}}$, one obtains the standard IRT model.

Although the development of IRT and RUM have different behavioral assumptions[12], they are mathematically identical models. That is, based on the data alone, one could not distinguish between the two behavioral stories. The IRT/HB model class can be viewed as a special case of CBC/HB™, as derived from RUM, and the estimation of the latent trait and test item parameters is very familiar to users of Sawtooth Software's CBC/HB™ package. However, IRT goes beyond standard choice based conjoint models by providing a rich framework for test item construction. Information about test items is summarized by the item characteristic curves.

The path just taken

To recap where we have been so far, we've described 1P and 2P parameter models for binary response data. In these models, there are parameters that describe individuals' locations on an observed, continuous trait, and parameters that relate the binary items to the same trait. A key characteristic of these models is their ability to estimate local reliability, and as a result you can customize the reliability of a scale you construct in terms of where on the trait of interest it is most reliable at discriminating differences between respondents.

We have given simple examples of 1P and 2P models. In these applications, we specified full probability models, and used MCMC methods, by means of WinBugs, to estimate the models' parameters. Finally, we have summarized how IRT and RUM models are related.

[12] IRT assumes the students has a "Correct" response when his or her latent ability variable crosses a difficulty threshold, while RUM specifies that the subject selects "Yes" or "No" to maximize two, competing utilities. The behavioral assumptions behind RUM seem to be more reasonable for selecting one of several product concepts as most preferred, while the behavioral story behind IRT seems more compelling for standardized testing.

There are four more things to note about these models for binary data and our modeling approach. First, among the other IRT model specifications in the literature there are 3P and 4P models that accommodate "guessing" and "random mistakes." We think that the 3P models are one way to take into account some kinds of response bias (e.g. yea-saying). Another method of adjusting for yea-saying is to introduce a subject-specific random effect for scale usage in 2P models. This will not work in 1P models, unless one is estimating multiple traits with different banks of items, because the latent trait and the scale-usage parameters are confounded

Second, when you fit IRT models using Bayes procedures, you have a very natural way to deal with missing responses. You simply estimate them like any other model parameter.

Third, the single trait model can be generalized to a multiple trait model. That is, instead of one latent trait driving the response, there may be multiple traits. This extension goes from a single factor model to a multiple factor model.

Fourth, IRT models are easily extended to include covariates for either the person or the item parameters. And, they can be very easily extended to ordinal response data. We turn to this, next.

Life is (even more) complicated

You can use the IRT approach for ordinal data like what rating scales usually should be taken to produce. We can write the following response function for this kind of ordinal data:

(6)
$$P(x_{ij} = k \mid \theta_i, \beta_j, \alpha_{\delta.j}) = \frac{\exp \sum_{l=1}^{k} (\beta_j \theta_i - \alpha_{\delta lj})}{\sum_{m=1}^{k} \exp\left(\sum_{l=1}^{m} \beta_j \theta_i - \alpha_{\delta lj} \right)}$$

Equation (6) expresses the probability of respondent i answering item j with response category k as a function of the respondent's $\boldsymbol{\theta_i}$, the item's discrimination parameter $\boldsymbol{\beta_j}$, and a set of threshold parameters $\boldsymbol{\alpha_{\delta j}}$ for the category cut-offs between the response categories. There are k-1 of them for a response scale with k categories. These are item-specific parameters, just like the $\boldsymbol{\beta_{j.}}$ An IRT model of this sort is sometimes called an "ordered polytomous response" IRT model.

Note that there is nothing that requires that all items have the same number of response categories. As you might imagine, if there are k response categories for an item, the item has k-1 ICCs. It is usually assumed that the discrimination parameters are equal across the ICCs of a given item, and hence there is only one $\boldsymbol{\beta_j}$ per item.

Getting personal with your laundry

Your client is back again, and because of your stellar work on brand adhesion measurement she has another request for you. She has an internal client interested in how attached customers are to her brand of laundry detergent, and so she has used 5 point Likert scales to collect users’ responses to the following:

1) *“If this detergent were a person, I’d want it to be my friend.”*
2) *“I recommend this detergent to all my friends and relatives.”*
3) *“This detergent is more effective than others.”*
4) *“I’ve been happy with the performance of this detergent.”*

The response scale labels ranged from “1, strongly disagree” to “5, strongly agree.” She has 420 complete responses.

So, being one step ahead of her because you already know how to fit an appropriate IRT model to her data, you fit the model specified in Figure 7, and obtain the results in Table 3 and Figure 8.

$$\mathrm{logit}(\mathrm{q}(\mathrm{x}_{\mathrm{ij}} = k)) = \beta_j \theta_i - \alpha_{jk}$$

$$p(x_{ij} = k) \text{ obtained from } \mathrm{q}(x_{ij} = k)$$

$$x_{ij} \sim \mathrm{categorical}(p(x_{ij} = k), \sum_{k=1}^{K} p(x_{ij} = k) = 1)$$

$$\theta_i \sim \mathrm{normal}(0,1)$$

$$\beta_j \sim \text{truncated normal}(0,0.001)$$

$$\alpha_{\mathrm{jk}} \sim \mathit{normal}(0.0.001)$$

<u>Figure 7.</u> **Specification summary for ordinal response IRT model for Likert rating scale data on detergent attachment.** J=4 rating scale items, k=5 ordinal response categories on each Likert scale. Each Likert item has a single discrimination parameter, beta. The betas are modeled as truncated normal so that they are nonnegative.

Scale:	β Posterior mean	2.5%	97.5%
Been happy	6.406	5.460	7.449
Better performance	7.066	5.863	8.289
Would recommend	7.726	6.682	8.911
Be my friend	7.214	7.187	8.359

<u>Table 3.</u> Posterior mean discrimation parameters for each of the four rating scale items about detergent attachment. Percentiles are as in Tables 1 & 2.

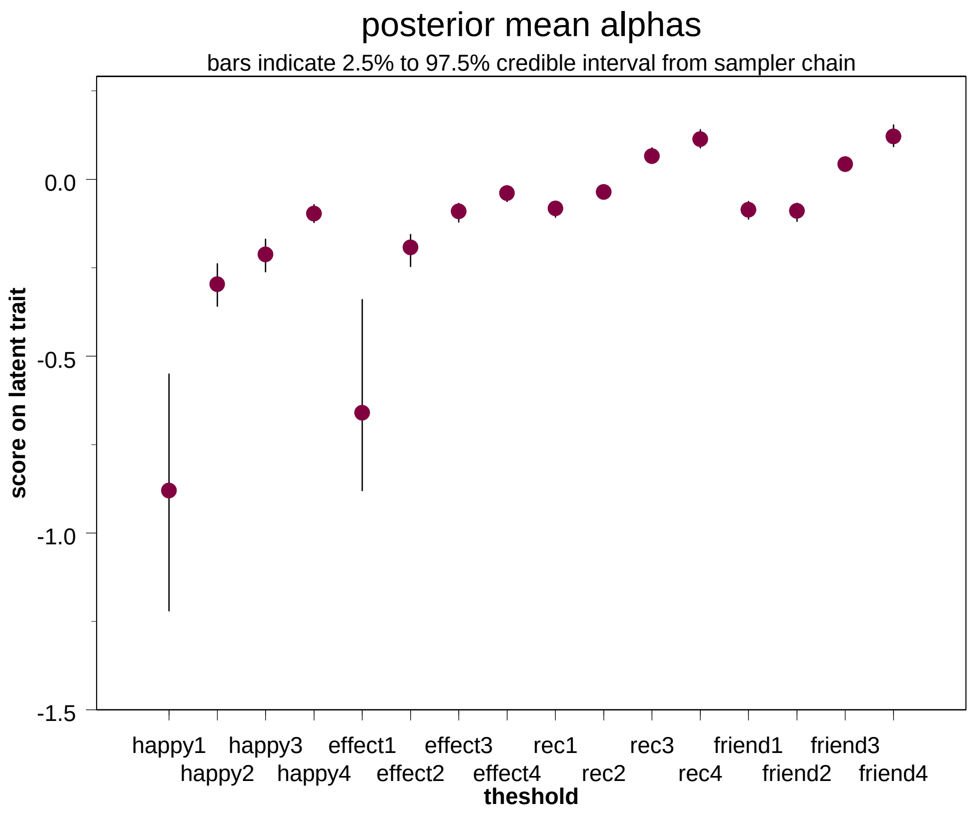


Figure 8. Posterior mean thresholds, or category cut-offs, for each of the Likert rating scale items used to measure detergent attachment. Thresholds are on the abscissa, with scales ordered from left to right in terms of increasing value of average within scale alpha. Posterior mean alphas (thresholds) are plotted on the ordinate, with 2.5 percentile and 97.5 percentile values from each thin chain shown as a bar.

Table 3 gives the posterior means of the discrimination parameters $\boldsymbol{\beta_j}$ for each of the four items. As you can see, each item proved to be quite effective at discriminating between respondents whose $\boldsymbol{\theta_i}$ fell above and below the item's thresholds.

Figure 8 summarizes the distributions of the threshold estimates for the four items. Plotted on the abscissa are the $\alpha_{\delta j}$ for all four items. The items are ordered from lowest to highest within-item average $\alpha_{\delta j}$, from left to right. On the ordinate are the posterior mean $\alpha\delta j$ plotted as filled symbols, along with the 2.5%-97.5% interval shown as a vertical bar. You'll note from Figure 8 that the items differ in the mean values of their thresholds, and that some thresholds are a lot less precise than are others. Those that separate the lowest response categories for the "happy with performance" and the "more effective" scales are particularly imprecise. A likely cause is that responses in the lowest category for these two items were very scarce, and the wide spread in their posterior distributions mostly reflects their prior distributions.

Going exogenous

Another way that IRT models can be extended is by adding covariates to the parameters. You can, for example, specify an IRT structural equation model in which latent trait models are the measurement models, and the structural equations describe linear or nonlinear relationships between the latent traits.

In our last example here we’re going to describe the simplest IRT model with covariates. In it, we express θ_i as the dependent variable in a regression equation in which the predictors are covariates.

It’s a good thing that we have this covariates model handy, because your client now has an even more interesting challenge for you. She has some rating scale data from 380 in-home product testers of a new concept called “SuSu Nirvana,” a nutri-stimulant beverage that will be sold over the counter.

Her ratings are on 5 point Likert scales, and were elicited by the following questions:

1) *“A SuSu in the morning gives me a jump on the day.”*
2) *“I feel more optimistic and confident when I’ve had some Susu.”*
3) *“SuSu helps me peak creatively.”*

Your client wants to know whether these items can be combined to provide an overall “SuSu impact” score (sort of a measure of connection with the intended positioning), and whether the score is related to two other variables, heavy caffeine use (a binary measure), and a score on a standardized sensation-seeking scale.

Figure 9 shows a specification for a model with one latent trait, and the two covariates. This model has the same IRT specification as in our previous example. The three regression coefficients (the ϕ’s) are given a multivariate normal prior.

The regression model:

$\theta_i \sim normal(\mu,\tau)$

$\mu = \phi_0 + \phi_1 z_1 + \phi_2 z_2$

z1 - standardized sensation-seeking score

z2 - 1=caffiene user, 0=non-user

$\phi_0,\phi_1,\phi_3 \sim MVN(\bar{\phi},R)$

$R \sim Wishart(diag(0.2),3)$

Latent trait model:

same as before, but with 3 rating scales

Figure 9. **Specification summary for IRT model with covariates.** The latent trait part is similar to Figure 7 of the previous example, but with three rather than four ordinal rating scale items. In the regression part of this model, the regression coefficients, the ϕ’s, are modeled as having a multivariate normal prior. The observed covariates, Z1 and Z2, are treated as continuous and binary variables, respectively.

Scale	β Posterior mean	2.5%	97.5%
Susu in the a.m.	1.872	1.078	3.167
More optimistic	2.609	1.702	3.848
Creative peaking	2.440	2.355	3.758

Table 4. Posterior mean discrimination parameters for each of the three rating scale items used to measure SuSu Nirvana affinity. Percentiles are as in Table 3. 11,000 sampler draws as burn-in, then every 20th of an additional 3,000. DIC=185.9, pD=88.56

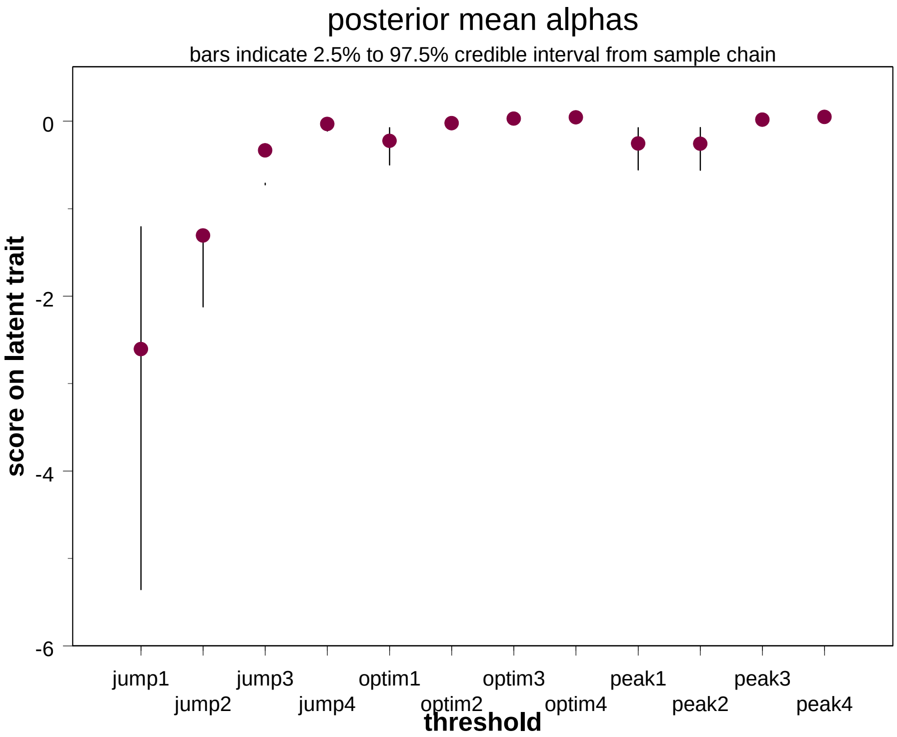


Figure 10. Posterior mean alphas for the thresholds of the Susu Nirvana Model. Details are as in Figure 8.

Regression Coefficient:	Mean posterior ϕ	2.5%	97.5%
Intercept	-0.859	-1.060	-0.716
Heavy caffeine user (Z2)	0.581	0.436	0.741
Sensation seeking (Z1)	0.616	0.438	0.825

Table 5. Regression coefficient estimates for SuSu Nirvana model.

Tables 4 and 5 and Figure 10 provide some model results. Table 4 and Figure 10 summarize the item parameters for the model. You'll note from them that all three rating scale items are discriminating between respondents with high values on the latent trait, and that the "optimistic" and "peaking" items are providing redundant information in the sense that their thresholds are very similar. Table 5 summarizes the regression part of the model. You can see that both covariates are positively related to the latent trait. We've assumed a causal direction here in the case of both. It could be the case that SuSu causes sensation-seeking, of course, if it did, it should not be over-the-counter, and one could extract additional consumer surplus.

So, does your client know how smart you are, now, or what?

Extensions, applications in marketing research, and final comments

A variety of other extensions to basic IRT models have been described. These include nonparametric IRT models, multivariate IRT models in which at least some of the observed variables "load" on more than one latent trait (e.g. Bacon, Rivers, Jackman & Hunter 2004), unfolding IRT models, IRT models that use items that are heterogeneous in their types (Bacon & Haddock 2004), and finite mixture IRTs.

We think important applications of IRT models in applied marketing research include the following:

- Tracking studies
- Adaptive surveying by applying a pool of developed items and reflective of specific measurement objectives
- Causal modeling
- Market segmentation
- Multi-lingual and multi-modal research

Here are some resources you might find useful if you decide to pursue IRT modeling. They include public domain sources of software:

- Social Science Measurement Project at sourceforge.net
  http://sourceforge.net/projects/ssm/

- www.rasch.org
- www.edres.org
- www.psychometricsociety.org
- BUGS www.mrc-bsu.cam.ac.uk/bugs
- R http://lib.stat.cmu.edu/R/CRAN/
- http://www.unt.edu/rss/class/rich/5840/
- BUGS XLA www.pipshome.freeserve.co.uk/stats/
- Free IRT Project http://freeirt.free.fr/

On some of these links you'll find commercial software for IRT modeling. These programs use traditional (i.e. non-Bayesian) modeling approaches. They include BILOG and MULTILOG, which are available from Scientific Software Inc., ProGamma, and Assessment Systems Corporation. Statistical Innovation's LatentGold also has IRT capability.

General References

Bacon, L., D. Rivers, S. Jackman and J. Hunter (2004), "Market Structure Analysis and IRT Models," in *Joint Statistical Meetings*, Toronto ONT: American Statistical Association, August.

Bacon, L and N. Haddock (LBA Ltd.; Atomic Intelligence) (204), "Taking Customers at Their Word: Natural Language Processing and the Analysis of Text Data," in *Advanced Research Techniques Forum*, Whistler B.C.: American Marketing Association, June.

Bock, D. and M. Aitken (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika*, 46, 443-459.

Bollen, K. (1989), Structural Equations with Latent Variables, New York NY: Wiley.

Congdon, P. (2003), *Applied Bayesian Modeling*, Wiley Series in Probability and Statistics, New York: Wiley.

Gelman, A., J. Carlin, H. Stern and B. Rubin (2004), *Bayesian Data Analysis*, Texts in Statistical Science, 2nd ed., Boca Raton FL: Chapman & Hall/CRC.

Heinen, T. (1996), *Latent Class and Discrete Latent Trait Models: Similarities and Differences*, Advanced Quantitative Techniques in the Social Sciences, vol. 6, Thousand Oaks CA: Sage.

Lord, F.M. and M.R. Novick (1968), *Statistical Theories of Mental Test Scores*, Reading MA: Addison-Wesley.

McFadden, D. (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” *Frontiers in Economics*, editor P. Zarembda, New York: Academic Press, 105-142

Moustaki, I., K. Joreskög and D. Mavridis (2004), "Factor Models for Ordinal Variables with Covariate Effects on the Manifest and Latent Variables: A Comparison of LISREL and IRT Approaches," *Journal of Structural Equation Modeling*, 11(4), 487-413.

AVOIDING IIA MELTDOWN: CHOICE MODELING WITH MANY ALTERNATIVES

GREG ALLENBY
OHIO STATE UNIVERSITY
JEFF BRAZELL
THE MODELLERS
TIM GILBRIDE
UNIVERSITY OF NOTRE DAME
THOMAS OTTER
OHIO STATE UNIVERSITY

ABSTRACT

Choice predictions from conjoint experiments can be inaccurate when there are many alternatives available in the marketplace. One reason for this is the large increase in the error space that typically accompanies each choice alternative. By restricting the error space, IIA (independence of irrelevant alternatives) meltdown associated with proportional draw can be avoided.

INTRODUCTION

The introduction of hierarchical Bayes (HB) methods in marketing has stimulated the use of choice models, particularly in the analysis of conjoint data. The hierarchical nature of HB models, which combines respondents' responses using random-effects, has been shown to improve the predictive accuracy of conjoint part-worth estimates. Typically, choice models are specified in logit form (see below), which yields a closed-form expression for the choice probability. The availability of a closed-form choice probability facilitates the development of marketplace simulators and other methods of exploring the effects of changes in product formulation on expected market shares.

While the logit model has many advantages over other models of choice, a drawback is the property known as the independence of irrelevant alternatives (IIA). The IIA property is related to the form of the choice probability, where the ratio of the probability of any two choice alternatives, i and j, does not depend on the value of other, irrelevant alternatives. The IIA is both a blessing and a curse. It is a blessing in that it allows study of a subset of brands available in the marketplace, yielding utility estimates that are not affected by the missing brands. It is a curse in the sense that the IIA property can be overly restrictive in reflecting substitution patterns among brands when some of the brands are more similar than others. Classic examples of IIA problems include the "red bus, blue bus" example in transportation, and discussions of the competitive structure between Pepsi, Coke and 7up, where changes in the price of Pepsi are expected to have an impact on the ratio of the choice probability of Coke to 7up. The presence of IIA implies that an increase in the probability of an "irrelevant" alternative (e.g., due to a change in price) is obtained by decreasing the probability of choice alternative i and j by an amount that is proportional to their choice probability. This proportional draw property insures that the ratio of choice probabilities for alternatives i and j remains constant.

The IIA property becomes extremely restrictive when there are many choice alternatives under study. The presence of many choice alternatives, coupled with the existence of the IIA (proportional draw) property, can lead to the unreasonable predictions when a new item is introduced into the choice set, or when the attributes of a choice alternative are altered. The presence of many choice alternatives increases the likelihood of dissimilar offerings for which the IIA property should not hold.

The purpose of this paper is to propose an approach to dealing with IIA meltdown within the context of a logit model. The IIA property can be traced directly to model error term assumptions. In the next section, we discuss alternative approaches to altering assumptions about the error term that can overcome IIA problems, and propose our approach. In section 3 we apply our method to a conjoint choice study of automobile purchases where the marketplace literally consists of thousands of choice alternatives. Section 4 discusses extensions for application to the analysis demand for packaged goods using scanner panel data. Concluding remarks are offered in section 5.

CHOICE MODELS AND ERROR TERMS

The choice probability of selecting alternative i can be expressed generically as:

$$\Pr(i) = \Pr(V_i + \varepsilon_i > V_k + \varepsilon_k \text{ for all k}) \qquad (1)$$

where i and k denote different choice alternatives, V is the measurable portion of utility that is revealed to the researcher, and the error term, ε, reflects the portion of utility known to the decision maker but unknown to the researcher. Choice models associated with equation (1) are sometimes referred to as random utility models because of the error term, ε. Regardless of the distributional assumptions made about ε, it is important to realize that the choice probability is "built up" by finding the set of error terms, $\{\varepsilon\}$, that correspond to the probability statement, i.e., $V_i + \varepsilon_i > V_k + \varepsilon_k$ for all k. This set of error terms for which equation (1) is correct is actually a subset of all the possible values of $\{\varepsilon\}$, and the choice probability is obtained by computing the probability associated with this set. A remarkable fact is that this computation leads to the logit choice model when the error terms are assumed to be distributed according to an extreme value distribution.

$$\Pr(i) = \exp[V_i] / \Sigma_k \exp[V_k] \qquad (2)$$

The IIA property of the logit model can be seen by taking the ratio of the choice probabilities of alternative i to alternative j:

$$\Pr(i)/\Pr(j) = \exp[V_i]/\exp[V_j] \qquad (3)$$

which does not depend on any of the other choice alternatives. The IIA property is unreasonable in large choice sets because it becomes unreasonable to expect that the ratio of choice probabilities would depend on none of the other choice alternatives. As the number of choice alternatives increase, there should be higher likelihood that at least one of the choice alternatives would have a differential effect that violates the IIA property.

The IIA property is not a problem in typical conjoint studies because choice sets usually have a small number of choice alternatives and researchers usually avoid including dissimilar choice alternatives. The goal of most conjoint analyses is to understand the drivers of preference among substitute offerings, and as a result the offerings are typically thought to be members of the same product class.

However, researchers are sometimes interested in using choice models to understand the impact of changes to brand formulation on market-wide demand. Many conjoint applications are aimed at simulating share movements for such a choice set. Whereas IIA may be behaviorally defensible in situations where similar alternatives are compared, it is less so in situations where consumers are likely to realize that some of the offerings are more similar than others.

Nested Logit Model

One approach to dealing with the IIA property when it becomes too restrictive is to assume that the error terms, $\{\varepsilon\}$, in equation (1) are correlated. The extreme value distribution allows for dependent error terms that take on a correlational structure known as an "intra-class" structure. The intra-class structure can be thought of as a block correlational structure, where the correlation between any two errors within a block is equal, and correlations among blocks are also similarly constrained. The resulting choice probability takes on a nested form:

$$\Pr(i) = \Pr(i|A)\ \Pr(A) \qquad (4)$$

where Pr(i|A) is the choice probability from within a choice set corresponding to the block, and Pr(A) is the choice probability among blocks. Moreover, each factor on the right side of equation (4) is logit in form. Thus, the nested logit model has a "local IIA" property. Details of the local IIA property associated with the nested logit model can be found in Allenby (1989).

While the nested logit model is effective at allowing for differential patterns of substitution within and among blocks specified by the researcher, it suffers from the same "IIA meltdown" as the standard logit model in large choice sets. When the number of choice alternatives numbers in the hundreds and above, the "local IIA" property can easily be assumed to apply to 10 or 20 choice alternatives within each block, which may be unreasonable.

Proposed Approach

Our solution to the IIA meltdown problem can be thought of as an extreme case of the correlated error assumption of the nested logit model. Rather than have correlated error terms associated with each choice alternative, we propose that choice alternatives within the same block have the *same realization* of the error term. This corresponds to a correlation of 1.0, the limiting case of the nested logit model.

For illustrative purposes, consider a choice set comprised of three alternatives in which two, 1a and 1b, share the same brand name, but differ in some minor way. These two choice alternatives share the same error realization, while the third alternative has a different error realization. Then, the choice probabilities associated with the three alternatives are:

$$\begin{aligned}
&\Pr(1a) = \Pr(V_{1a} + \varepsilon_1 > \{V_{1b} + \varepsilon_1, V_2 + \varepsilon_2\}) \\
&\Pr(1b) = \Pr(V_{1b} + \varepsilon_1 > \{V_{1a} + \varepsilon_1, V_2 + \varepsilon_2\}) \\
&\Pr(2) = \Pr(V_2 + \varepsilon_2 > \{V_{1a} + \varepsilon_1, V_{1b} + \varepsilon_1\}) \qquad (5)
\end{aligned}$$

Now, since the same error term, ε_1, is associated with two of the choice alternatives, the expression for the choice probability will be different than in equation (1). Assuming the ε_1 and ε_2 are distributed extreme value, the expressions for the choice probabilities becomes:

$$\begin{aligned}
&\Pr(1a) = \exp[V_{1a}] / D \qquad \text{if } V_{1a} > V_{1b} \text{ else } 0 \\
&\Pr(1b) = \exp[V_{1b}] / D \qquad \text{if } V_{1b} > V_{1a} \text{ else } 0 \\
&\Pr(2) = \exp[V2] / D \\
&\qquad D = \max\{\exp[V_{1a}],\exp[V_{1b}]\} + \exp[V2] \qquad (6)
\end{aligned}$$

That is, if $V_{1a} < V_{1b}$, the choice probability for alternative 1a is zero. The choice probability for 1a is non-zero only if its value, V_{1a}, is greater than the other choice alternatives that share the same error realization. Thus, the choice probabilities can be thought of as arising from a two-step process:

1. The "best on the block" is determined for each set of choice alternatives that have the same error realization, and
2. Logit choice probabilities are formed based on the maximum from each block.

The first step in computing the choice probability involves a deterministic search among the alternatives within the block, and the second step resembles the standard logit model. We emphasize, however, that this description is of the solution procedure for computing the choice probabilities, not the assumed behavioral story of the decision maker. The decision maker is simply assumed to be a utility maximizer, and that alternatives within blocks share the same error realization.

The strength of our approach stems from strong assumptions about the origins of the error term. We assume the error term arises because of an important attribute that is shared among offerings within the same block. The most likely candidate for such an attribute is a brand name, or a quality level, whose value is difficult to quantify. For example, when buying durable goods the manufacturer trademark may be responsible for the error in researcher knowledge of the consumer decision process. When buying packaged goods, the error may not be associated with the quantity present in different package sizes, but rather the quality of brand. Below we discuss application of our proposed model in two such applications, and illustrate how parsimonious specification of error term realizations can help deal with IIA meltdown.

AUTOMOBILE STUDY

The automobile market in the US consists of thousands of offerings. To model the entire market, autos are broken down in categories by brand (e.g. Toyota, Ford) and nameplate (e.g. Camry, Corolla, Focus, F-150). Most manufacturers have even narrower classifications defined as "trim level concepts" (e.g. Camry SE, Camry LE) where nameplates differ only on option packages, engine sizes or other features.

The standard approach to simulating market share would be to assume each automotive offering as a distinct alternative among the thousands available. Because the market is so large and autos have so many features, it is impossible to capture effects for every distinction in the modeling process. The large number of alternatives available within brand or nameplate that are closely related on features *not* in the choice model suggests it may not be valid to assume each concept has a unique error term.

In general terms, the approach taken here to avoid IIA meltdown involved a two-stage estimation of market share. First, we estimate utilities for each auto independently. Then, within nameplate categories, we select only the alternative with the highest utility. Once each nameplate has its "winning" alternative, we allow these winners to compete in stage two. Consider a simple example with three nameplates. Using the standard choice modelling approach, we would have nine competing alternatives with nine error terms. In the proposed model, we have only three in the second stage.

Standard Approach

Concept	Utility	Share Estimate
Chevy Malibu	0.75	1.7%
Chevy Malibu Maxx	7.03	16.3%
Subtotal		**18.1%**
Ford Focus	0.85	2.0%
Ford Focus ZX4	8.39	19.4%
Ford Focus ZX4 ST	6.85	15.9%
Ford Focus SVT	2.66	6.2%
Subtotal		**43.4%**
Toyota Camry	6.90	16.0%
Toyota Camry LE	3.90	9.0%
Toyota Camry SE	5.82	13.5%
Subtotal		**38.5%**

Two-Step Approach

Concept	Utility	Share Estimate
Chevy Malibu	0.00	0.0%
Chevy Malibu Maxx	7.03	31.5%
Subtotal		**31.5%**
Ford Focus	0.00	0.0%
Ford Focus ZX4	8.39	37.6%
Ford Focus ZX4 ST	0.00	0.0%
Ford Focus SVT	0.00	0.0%
Subtotal		**37.6%**
Toyota Camry	6.90	30.9%
Toyota Camry LE	0.00	0.0%
Toyota Camry SE	0.00	0.0%
Subtotal		**30.9%**

At the simulation stage, this methodology is applied at the individual level, so the shares smooth out significantly across the entire sample.

The two-stage approach also helps us avoid overestimation of nameplate market share when that nameplate has a high number of different trim levels on the market. Using the above example, Ford could simulate what would happen if they introduced four new trim level concepts of their Focus line. If they used the standard approach, the total share of Ford Focus would increase drastically for no other reason than that the nameplate now has more alternatives; this overestimation might occur even if the newly introduced line garnered lower preference than the old lines. However, because only the winning alternative from each nameplate competes in the two-stage model, the total share of Ford Focus in the model adjusted for IIA Meltdown would only increase if the new alternatives had higher utilities in parts of the marketplace.

PACKAGED GOOD STUDY

The analysis of demand for packaged goods must also deal with the issue of large choice sets. Consider, for example, the number of brand-pack combinations available in most product categories. An extreme example is the beer category, where offerings of a specific brand are available in multiple package sizes (e.g., 6-pack, 12-pack, 24-pack, etc.). In a recent study by Allenby, Shively, Yang and Garratt (2004), the three major brands of light beer (Bud, Miller and Coors) have 84 brand-pack combinations among them.

One approach to modeling demand for packaged goods is to treat each brand-pack combination (i.e., stock keeping units, or sku) as a unique choice offering. But, as with automobiles, the assumption that each of these offerings is associated with a unique error term ignores the fact that the same offering is present in any of the package sizes. To reduce the error space and avoid IIA meltdown, it is useful to consider the economic foundation of random utility models. A random utility interpretation of equation (1) is that $V_i + \varepsilon_i$ is that V_i is the logarithm of the marginal utility of offering i. More generally, discrete choice arises when this utility is assumed to be linear:

$$U(x_i) = \psi_i x_i \qquad (6)$$

where $U(x_i)$ is the utility associated with "x" units of offering "i", and ψ_i is the marginal utility of the offering (equal to V_i in equation 1). The assumption that marginal utility is constant (i.e., equal to ψ_i for any amount of consumption), leads to linear indifference curves and utility maximizing solutions where only one of the offerings is selected.

If we assume that marginal utility, ψ_i, is random with a multiplicative error, we have,

$$U(x_i) = \psi_i e^{\varepsilon i} x_i \tag{7}$$

Similar to the automobile example above, equation (7) associates an error term with each brand "i" regardless of the package size "x." Taking the logarithm of equation (7) leads to a linear equation:

$$\ln U(x_i) = \ln \psi_i + \ln x_i + \varepsilon_i \tag{8}$$

As discussed by Allenby, et al. (2004), the choice probability can be expressed in logit form (assume the error is distributed extreme value):

$$\Pr(x_i) = \frac{\exp[\psi_i + \ln(x_i) + \alpha \ln(T - p(x_i))]}{\sum_{k=1}^{K} \exp[\psi_k + \ln(x_k^*) + \alpha \ln(T - p(x_k^*))]} \tag{9}$$

where x* is associated with the package size that maximizes the "best on the block" search, $p(x_i)$ is the price of purchasing x_i units of the brand, T is a budgetary allotment, and the term "αln(T-p(x))" reflects the presence of an outside good whose contribution to a consumer's utility is affected by how much money is spent in the product category.

Application of equation (9) to packaged goods data results in an extremely parsimonious model, with marginal utility ψ associated with each brand and not each brand-pack offering. For the 84 brand-pack combinations in the light beer category, Allenby, et al., estimate a demand model with only six parameters. The model is shown to be predictively accurate, and does not have unreasonable implications

CONCLUDING REMARKS

Our approach to dealing with models associated with large choice sets is radical. We make strong assumptions about the error term to reduce their number and avoid the restrictive influence of IIA. The reduction is accomplished by considering the origin of the error term, and assuming that the realizations of the error are the same for offerings within similar "blocks" of offerings. This reduction in the error space results in improved predictions and more reasonable forecasts than standard models.

Considering the origins and properties of the error term in models of consumer behavior is a fruitful avenue for future research. Sawtooth Software's "Randomized First Choice" option is another example where predictive improvements are obtained by modifying the error term. Data in marketing is characterized by a large number of respondents, but a relatively small number of observations per respondent. It is therefore beneficial to be judicious in including error terms in models so that the individual respondent-level data is not overwhelmed by noise.

REFERENCES

Allenby, Greg M. (1989) "A Unified Approach to Identifying, Estimating and Testing Demand Structures with Aggregate Scanner Data," *Marketing Science*, 8, 265-280.

Allenby, Greg M., Thomas S. Shively, Sha Yang and Mark J. Garratt (2004) "A Choice Model for Packaged Goods: Dealing with Discrete Quantities and Quantity Discounts," *Marketing Science*, 23, 1, 95-108.

ADVANCED TOPICS IN CHOICE MODELING

A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs)

Rich Johnson
Sawtooth Software, Inc
Joel Huber
Duke University
Bryan Orme
Sawtooth Software, Inc.

Abstract:

At the previous Sawtooth Software conference, Johnson, Huber & Bacon introduced an algorithm for adaptive choice design which produced share predictions directionally better than those of standard CBC. This paper reports results of a further study which uses a slightly more powerful algorithm, but less self-explicated information for preliminary estimation of partworths. This paper also studies both Full Profile and Partial Profile choice designs.

We have not yet been able to produce an adaptive design algorithm that is more effective than CBC. Results suggest that the design algorithm we tested would benefit from better preliminary estimates of respondents' partworths.

Also, Full Profile designs are found to be more effective than Partial Profile designs at predicting Hit Rates for Full Profile holdout choice tasks.

Introduction:

Choice-Based Conjoint analysis (CBC) has achieved a dramatic increase in use during the past decade. According to a recent Sawtooth Software newsletter, "Among Sawtooth Software customers, its use has now eclipsed the use of ACA" (Sawtooth Software 2004).

There are several reasons for the growth of CBC. One important reason is that choices are more like actual marketplace behavior than are the rankings or ratings used by other conjoint methods. Another reason is that the recent availability of Hierarchical Bayes (HB) methods permits estimation of partworths for individuals, where previously it had seldom been possible to collect enough choices from each individual to support individual-level analysis.

Choice designs are inherently less efficient than other conjoint methods. Each choice task requires that the respondent become familiar with several alternatives, but the resulting data indicate only which alternative is chosen, rather than anything about intensity of preference. Despite the important contribution of HB, there remains considerable incentive to make choice designs more efficient.

At the previous Sawtooth Software conference, Johnson, Huber, & Bacon (JHB) introduced an adaptive algorithm for design of choice studies (ACBC) which customized the design for each respondent based on self-explicated information about his/her partworths. They compared results for groups that differed in method of design, some receiving customized designs and others receiving standard CBC designs. They also examined two other experimental groups for whom choice designs were further manipulated to produce increased utility balance. Their basic conclusions were:

- ACBC and CBC achieved similar Hit Rates, though CBC was slightly higher.
- ACBC produced better share predictions than standard CBC. (Though the authors were not able to test whether the improvement was statistically significant.)
- When designs were manipulated to increase utility balance, and thus to increase indifference among alternatives, Hit Rates and share predictions were both degraded unless the missing information was re-introduced in some way. This was done by using self-explicated desirability judgments as constraints on estimated partworths. Hit Rates and share predictions for all design methods improved with the reintroduction of this information.

These results seemed promising, and suggested a further look at potential gains from adaptive choice design. We note that in an interesting recent article, Toubia, Hauser and Simester (2004) have also described success in designing Adaptive CBC questionnaires by a different method, though their results involved synthetic data rather than actual respondents.

This paper reports on a second study which differs in several ways from that of JHB.

- Since modifying designs to increase utility balance was found to degrade performance unless the missing information was re-introduced during estimation, we decided not to use utility balance as a criterion to alter designs.
- Partial Profile choice designs present a way to accommodate larger numbers of attributes while retaining manageable choice tasks. We decided to test ACBC vs. CBC with both Partial Profile and Full Profile choice tasks.
- JHB found that desirability ratings for levels within attributes were useful for improving Hit Rates and share predictions but that ratings of attribute importance were not useful. Therefore we decided to retain desirability ratings but to omit attribute importance ratings.
- We made improvements to the adaptive design algorithm, which we expected would produce slightly better performance. These are described in the Appendix.

In the balance of this paper we describe the adaptive algorithm and the study design, present results of the study, and then discuss implications of those results and what we believe to be appropriate next steps.

THE ACBC ALGORITHM:

We provide a non-technical description of the ACBC design algorithm used in this study, which attempts to maximize "D-Efficiency." A more complete description is provided in the Appendix. The logic follows Huber and Zwerina (1996).

The Design Matrix for a choice task is a matrix of ones and zeros. There is a row for each alternative, and a column for each attribute level. Every row contains zeros except for the attribute levels that appear in that alternative, where the entry is a one.

Suppose we have two attributes, one with two levels and one with three levels, and that there are two alternatives in the choice task. Then the design matrix for a choice task might look like this:

	--Attribute1--		----Attribute 2----		
	Lev1	Lev2	Lev1	Lev2	Lev3
Alternative 1	1	0	0	1	0
Alternative 2	0	1	0	0	1

The first alternative has level 1 of attribute 1 and level 2 of attribute 2. The second alternative has level 2 of attribute 1 and level 3 of attribute 2.

Suppose we have preliminary estimates of this respondent's partworths. Then the respondent's utility for each alternative would be obtained by adding up the partworths corresponding to ones in that row. For example, suppose this respondent's partworths are as shown in the following table. Then utilities for the alternatives are also as shown:

	--Attribute1--		----Attribute 2----				
	Lev1	Lev2	Lev1	Lev2	Lev3		
Est Partworth	0.5	-0.5	-0.3	0.0	0.3		
						Utility	Probability
Alternative 1	1	0	0	1	0	0.5	.67
Alternative 2	0	1	0	0	1	-0.2	.33
Prob-Wtd Mean	.67	.33	.00	.67	.33		

According to the logit model, the probabilities of the respondent choosing each alternative are proportional to the exponential transformations of their utilities. Exponentiating each utility produces values of 1.65 for the first alternative and .82 for the second. Percentaging these to add to 1.0 gives choice probabilities of .67 for the first alternative and .33 for the second, so this respondent should be twice as likely to choose the first alternative as the second.

We are interested in assessing the amount of information about the respondent's partworths that can be contributed by the answer to this choice task. If one probability is very much larger than the other, we are confident the respondent will choose that alternative, so observing that choice won't increase our knowledge. Therefore, other things being equal, it is good that no alternative be strongly dominant.

It is also good for the alternatives to be as different from each other as possible. For example, since neither alternative includes the first level of attribute 2, the respondent's choice can't tell us anything about that level.

To measure the amount of information provided about each attribute level we compute something analogous to the variance of the design parameters in each column. We first compute the probability-weighted mean for each column, by multiplying each element of the design matrix by the choice probability for that alternative, and then adding the column sums. The probability-weighted means are in the first row of the next table. Then each column's probability-weighted mean is subtracted from each element, to get the table of differences from that mean.

	--Attribute1--		----Attribute 2----		
	Lev1	Lev2	Lev1	Lev2	Lev3
Prob-Wtd Mean	.67	.33	.00	.67	.33
Differences from Mean					
Alternative 1	.33	-.33	0	.33	-.33
Alternative 2	-.67	.67	0	-.67	.67
Sum of Squares	.56	.56	.00	.56	.56

Finally, the sum of squared differences is accumulated for each column. This results in something analogous to a variance for that attribute level, providing information about how different the alternatives are on that attribute level. Notice that the sum of squares is zero for the first level of attribute 2, because the alternatives do not differ with respect to that level. However, the choice of one alternative over the other would permit us to infer something about the other four levels, each of which have sums of squares greater than zero. In general, we would like these sums of squares to be as large as possible.

We would also like the sum of cross-products of each pair of columns to be as close to zero as possible. In other words, we would like the ones in the various columns of the design matrix to appear independently of one another. Of course, this cannot be done when there are only two alternatives in a single task. But suppose there were many different choice tasks. Then it might be possible to arrange a design in which, across many tasks, the columns of the design matrices are uncorrelated.

There is one more thing to consider—we have not yet accounted for the fact that an alternative which is seldom chosen is less informative than one chosen more often. In a sense, the information value of an alternative is proportional to its likelihood of being chosen. We account for this by weighting each alternative by its choice probability when accumulating sums of squares and cross products.

Now, suppose we compute the probability-weighted mean for each task, column-center the design matrix for that task, and accumulate the sum of squares and cross products over tasks, weighting the results for each alternative by its choice probability. That would result in a 5 x 5 matrix in which the diagonal contains weighted sums of squares like those above, but accumulated over many choice tasks. The off-diagonal elements would be weighted sums of cross products for pairs of columns. Such a table is a called the "Information Matrix" for that design.

We would like diagonal elements of this Information Matrix to be as large as possible, and its off-diagonal elements as close to zero as possible. When that occurs, the design is optimally efficient, in the sense that ensuing estimates of partworths have the smallest standard errors. A mathematical function that measures the extent to which that occurs is the determinant of the Information Matrix, and designs that maximize the determinant are said to be "D-Efficient." Our ACBC design algorithm chooses the design matrix for each new task in a way that attempts to maximize D-Efficiency for the design consisting of previous tasks plus the new task. Details are provided in the appendix. Here we shall just say that the algorithm attempts to find a design matrix for the next task which is in the subspace defined by the characteristic vectors corresponding to the smallest roots of the current Information Matrix.

The actual efficiency of a choice design depends upon an individual respondent's part worths, because they are needed to estimate choice probabilities, which are in turn needed to column-center each task and to weight the squares and cross products for each alternative. Therefore, a design can be more efficient for a particular respondent if we have a good preliminary estimate of his or her partworths.

In this study, as in the JHB study, we used self-explicated judgments to estimate those partworths. It would be possible to update estimates of partworths as the interview progresses, but in the interest of simplicity we did not do that, and the initial estimates were used throughout the interview. Also, the JHB study asked for both attribute importance judgments and judgments of desirability of levels within attributes, but this study did not ask for attribute importance judgments. Therefore in this study all attributes were assumed to be of equal importance for the preliminary estimates of partworths.

The Design of the Current Study

The subject of this study was laptop computers. There were 9 attributes with a total of 31 levels, as shown in Table 1. Respondents were randomly assigned to four treatment cells:

1. Regular CBC, Full Profile
2. ACBC, Full Profile
3. Regular CBC, Partial Profile
4. ACBC, Partial Profile

Apart from these treatment differences, the questionnaire flow was identical in each cell.

> First each respondent answered questions about general characteristics he/she sought in computers and the type of usage expected for the computer to be purchased.
>
> Next each respondent indicated the relative desirability of levels within each of four attributes: Brand, Operating System, Pointing Device, and Exterior Material.

> (For the other five attributes we assumed respondents would prefer greater capability.)
>
> Next each respondent answered two Full Profile holdout tasks. These, as well as all others, contained three choice alternatives.
>
> These were followed by 14 calibration tasks. For cells 1 and 2 these were Full Profile tasks, and for cells 3 and 4 they were Partial Profile tasks presenting four attributes at a time.
>
> Finally there were four additional Full Profile holdout tasks, two of which were identical repeats of the first two holdouts.

The calibration tasks were unique for each respondent, using either the "complete enumeration" method of CBC for Full Profile, the "shortcut" method for Partial Profile, or the adaptive design algorithm for ACBC. The holdout tasks were not unique for each respondent. There were four different versions of holdout tasks and each respondent was randomly assigned to one version or another. This meant that although each respondent only saw six holdout tasks, we had aggregate data for a total of 24 holdout tasks with which to test share predictions. Having six independent holdout groups results in less precision for each group but increases assurance that the results are not a function of the particular holdouts chosen.

We employed web-based survey administration using AOL's Opinion Place panel, managed by a joint AOL-SPSS team. We are indebted to Andrea Durning and Becky Cunningham of SPSS for arranging the fieldwork and absorbing the data collection costs for this methodological study. The sample was drawn using a "river" methodology, wherein respondents enter Opinion Place from a variety of popular portals on the Web. Respondents were screened for age (above 18), PC usage, purchase interest for laptops, and comfort level with comparing basic features of laptop computers. There were a total of 1009 respondents, with at least 247 in each treatment cell.

EQUIVALENCE OF SAMPLES IN TREATMENT CELLS:

Since different treatments were administered to different subsets of respondents, it was important to ascertain that the respondents in all cells were equivalent. We checked several measures of this:

The average number of minutes per interview was 7.7 minutes, with a maximum of 7.8 and a minimum of 7.6.

Reliability percentages for the repeated holdout tasks averaged 70%, with maximum for treatment groups of 71% and minimum of 69%. These values provide no evidence of systematic difference among treatment cells.

Respondents were asked to classify themselves into one of the following three categories:

> I'm principally concerned that technology products are easy to use, that they work consistently, and that I can receive helpful technical support if I need it

I like to have the latest and greatest technologies, and am willing to pay a premium for them.

It is very important for me to get a good value for the money, even if that means not having the absolute latest technology.

Overall, 14% of respondents chose “ease of use”, 32% chose “latest technologies” and 54% chose “value.” There were no significant differences between cells, with the Chi Square value smaller than would be expected due to chance 70% of the time.

Respondents were also asked about the expected use of the computer they might buy:

Primarily for business-related activities

Primarily for home use (not for business-related activities)

For both business and home use

Other

Few respondents chose “Other” so it was combined with “Both.” Overall, 6% of respondents chose “Business”, 42% chose “Home,” and 52% chose “Both or Other.” This Chi Square value ***was*** significant at the .01 level. Cell 1 (CBC Full Profile) respondents were more likely to choose “Both or Other”, and Cell2 (ACBC Full Profile) respondents were more likely to choose “Home.” However, a covariance analysis was done to assess the effect of this imbalance on Hit Rates, and it was found to have no effect.

Available evidence suggested that the four treatment cells contained equivalent samples of respondents, so we may proceed to examine the effects of the experimental treatments.

EXPERIMENTAL RESULTS:

Individual partworths were estimated using Sawtooth Software’s CBC/HB software, with separate analyses conducted for each cell. We also estimated each respondent’s partworths two ways: without constraints, and also (using a special version of the software) with constraints that forced partworths within attributes to have specified rank orders. These rank orders were determined by self-explicated desirability ratings, or by levels of capability for those attributes where order was thought to be obvious.

The HB runs used 2000 “burn in” iterations, and the final estimates for each individual were obtained by averaging 5,000 random draws from each individual’s posterior distribution, with a draw on each tenth iteration.

Hit Rates for constrained and unconstrained solutions are shown in the tables below, rounded to the nearest percentage. Significance tests have been conducted using more precise numbers.

Hit Rates – Unconstrained

	Full Profile	Partial Profile	Average
CBC	71	63	67
ACBC	68	63	66
Average	69	63	66

Hit Rates – Constrained

	Full Profile	Partial Profile	Average
CBC	70	64	67
ACBC	70	64	67
Average	70	64	67

The most striking result is that Hit Rates for Partial Profile respondents are much lower than for Full Profile respondents. This is true for both CBC and ACBC, and both constrained and unconstrained methods of estimation. Of course, the holdout tasks employed Full Profiles, so to some extent this result may reflect a "methods bias." However, almost all conjoint validation exercises involve predictions of Full Profile choices, so this result cannot be dismissed as inconsequential. Most comparisons to date between Full Profile and Partial Profile CBC have found strong congruence between the parameters, with a few isolated significant differences between the parameters after adjusting for scale (Chrzan 1999, Patterson and Chrzan 2003). These same authors have also argued that Partial Profile CBC tasks are overall more efficient than Full Profile tasks. The differences in predictive accuracy between Full Profile and Partial Profile for our study seem at odds with those conclusions.

We examine possible reasons for the difference in performance between Full and Partial Profile in the Discussion section.

Looking at the row averages of the two tables, there is little difference between Hit Rates for CBC and ACBC, but a slightly better result for CBC, both similar to previous findings of JHB. Also, not surprisingly, constrained estimation is slightly more successful, on average, than unconstrained estimation.

It is also noteworthy that Hit Rates for the Full Profile cells average about 70%, which is also the average reliability for the repeated holdout tasks. Although JHB reported higher reliability and Hit Rate percentages, they also found Full Profile Hit Rates to be about equal to average reliabilities.

Differences among the cells of these two tables were tested with an analysis of covariance, using as independent variables several measures on which we examined the equivalence of groups in the previous section, as well as which of the six holdout sets each respondent saw. All Full Profile cells are significantly higher than any Partial Profile cells, but there are no other significant differences among treatment cells.

Share Predictions were made using Sawtooth Software's "Randomized First Choice" method of simulation. This entails making thousands of "random draws" from each respondent's distribution of partworths, achieved by perturbing his point estimates by specified amounts of normal random variation. For each draw the respondent is allocated to the product with apparent maximum utility. By averaging over ten thousand draws, an estimate was made of each individual's share of preference for each product. The amount of random variation used for each treatment cell was that which produced the best share predictions for that cell.

The quality of predictions is measured by Mean Absolute Error (MAE), and we again show results separately for unconstrained and constrained solutions

Mean Absolute Error predicting Choice Shares

Share Predictions (MAEs) – Unconstrained

	Full Profile	Partial Profile	Average
CBC	4.6	7.1	5.8
ACBC	6.5	7.2	6.8
Average	5.6	7.2	6.4

Share Predictions (MAEs) – Constrained

	Full Profile	Partial Profile	Average
CBC	5.2	7.2	6.2
ACBC	6.1	6.3	6.2
Average	5.6	6.8	6.2

These MAE values are about two and a half times as large as those reported by JHB. One reason for this is that we divided our respondents among six groups of holdout concepts, whereas JHB showed the same holdout tasks to all 1000 respondents. Thus we have more holdout tasks than JHB, but fewer answers to each of them. The larger number of holdout tasks permits us to more easily test the significance of differences among treatments, but at the cost of less precision in our share predictions.

Since MAEs are computed from aggregate values rather than at the individual level, the differences among these cells cannot be tested using an individual-level analysis of covariance, as we did with Hit Rates. However, since results for each treatment are based on predictions of the same 24 holdout tasks, it is possible to do a two-way analysis of variance of MAEs for treatments crossed with holdout concepts. The interaction of treatments x holdouts is used as the error term for the between-treatments contrast. Testing differences between all 8 cells obtained by combining constrained and unconstrained estimation methods, the differences among cell means did not achieve significance at the .05 level. (This is in contrast to a similar test on aggregate Hit Rates, where the difference among cells was significant at far beyond the .01 level.)

Nonetheless, let us look at the differences as they appear. CBC seems to have a slight advantage over ACBC when not constrained, but there is no difference when constraints are used. Full Profile seems to have an advantage over Partial Profile for both constrained and unconstrained solutions. Finally, constraints seem to make little systematic difference.

These findings are at odds with the earlier findings of JHB, where ACBC produced better share predictions than CBC, especially when constraints were applied. This result is disappointing, and leads us to wonder what has caused it. In the next section we examine that question, as well possible reasons for the disappointing performance of Partial Profile.

DISCUSSION:

Here we consider possible reasons for ACBC's failure to produce better share predictions than CBC. As we have already commented, this study differed in two ways from the previous one, both of which were expected either to have no impact or to have a small positive impact.

The first possibility concerns a small change in the design algorithm which is described in the Appendix. Although anything is possible, it does seem extremely unlikely that that change could have caused the problem.

The second possibility is that in this study we did not collect self-explicated attribute importance judgments, whereas JHB did. JHB found that information not to be helpful in constraining partworth estimates, although within-attribute desirability ratings were found to be helpful. However, the preliminary partworths JHB used for questionnaire design did employ those importance ratings, and ours did not. Our preliminary partworths assumed that all attributes were of equal importance. Omitting those judgments may have been a mistake.

Recall that the adaptive design algorithm attempts to maximize D-Efficiency. We have examined the average D-Efficiency of the designs used in all four treatment cells, assuming partworths equal to those we used in the adaptive design. These efficiency values have been scaled with respect to the value for the CBC Full Profile cell.

Average D-Efficiency, Using Preliminary Partworths

	Full Profile	Partial Profile	Average
CBC	1.00	0.86	0.93
ACBC	1.99	1.65	1.82
Average	1.50	1.26	

The adaptive design algorithm appears to have done what it should. The ACBC values are approximately twice as large as the corresponding CBC values, meaning that the design algorithm has doubled the efficiency of the designs, assuming the preliminary partworths are valid.

Also, as one would expect, the Partial Profile designs have less statistical efficiency than the Full Profile designs. Researchers who advocate Partial Profiles expect this, but believe that a corresponding reduction in respondent confusion will more than compensate for the loss in statistical efficiency.

Our next step was to repeat the efficiency calculation, but using respondents' final (unconstrained) partworths in the computation of D-Efficiency. Those results are shown in the following table.

Average D-Efficiency, Using Final Partworths

	Full Profile	Partial Profile	Average
CBC	1.00	0.84	0.92
ACBC	1.09	1.14	1.12
Average	1.05	0.99	

When using final partworths the ACBC cells are only slightly more D-Efficient than the CBC cells. This suggests that the preliminary partworths may have been rather poor estimates of the final partworths. To assess this possibility we computed the correlation for each individual between preliminary and final (unconstrained) partworths, and then averaged those values within each treatment cell. The results are shown in the following table:

Average Correlations Between Preliminary and Final Partworths

	Full Profile	Partial Profile
CBC	.445	.643
ACBC	.446	.572
Average	.445	.608

Two things are interesting about these numbers. First, the correlations between preliminary and final partworths are quite low. This suggests that the preliminary partworths were indeed poor estimates of the final values, and that it may not have done much good to construct adaptive designs based on those estimates. Second, the correlations are higher for Partial Profile than for Full Profile, indicating that the designs may have been more "on target" for the Partial Profile respondents than for the Full Profile respondents. This is borne out by the fact for adaptive designs, D-Efficiency is higher for the Partial Profile cell than for the Full Profile cell when using final partworths.

Tables 2 and 3 contrast average partworths and attribute importances for the Full Profile and Partial Profile respondents. These are unconstrained estimates of partworths, and are re-scaled using the default method in Sawtooth Software called "Zero-Centered Diffs" (partworths were scaled for each individual so that the sum across partworths for each attribute is zero and the average range within attributes is 100). In both cases Brand

and Price were the most important attributes, but for the Full Profile respondents Brand was more important and Price was much more important. The two sets of partworths are significantly different ($p<0.05$) for 31 of the 38 levels, including all levels for price.

Among the Partial Profile respondents, attribute importances were generally flatter than among Full Profile respondents. This characteristic has also been noticed in other conjoint studies, where it is often observed that attribute importances computed from ACA partworths (also a partial profile technique) are flatter than attribute importances from Full Profile conjoint studies. The partial profile exercise may also have resulted in understating the importance of price, similar to what is often seen in ACA, due to "double counting" of non-price attributes. However, without further information about real-world purchase decisions, we cannot know which set of part worths is more correct.

This characteristic may help explain two phenomena.

1. Adaptive designs were more efficient for Partial Profile respondents than for Full Profile respondents. This may have been because the assumption of equal attribute importances used in the design was closer to the truth for Partial Profile respondents.
2. Partial Profile respondents had lower Hit Rates than Full Profile respondents when responding to Full Profile holdout tasks. This may have been because respondents place more importance on Brand and Price when responding to Full Profiles (where Brand and Price are always present) than when responding to Partial Profiles (where Brand and Price appeared only about half of the time). This would suggest that partworths developed from Partial Profiles would be at a disadvantage when used to predict responses to Full Profiles.

CONCLUSIONS AND RECOMMENDATIONS:

Based on the results of this study, we have reached several general conclusions.

Despite our considerable efforts to improve on CBC’s standard approach to design, we were not able to do so in this study. This finding is the basis for our subtitle, “The Surprising Robustness of Standard CBC Designs.”

The adaptive algorithm does appear to be effective at increasing D-Efficiency with respect to the partworths it is given to work with. Thus there is hope for adaptive design, despite its failure to produce better share predictions in this study.

The preliminary partworths used for adaptive design in this study were not effective. If self-explicated partworths are used for study design, it may be necessary to include attribute importance judgments as well as judgments of desirability of levels within attributes. This may be particularly true when the preference order of attribute levels varies across respondents, as in this study.

We did not observe a dramatic improvement in either Hit Rates or share predictions due to use of constraints in estimation. JHB did find constraints to be valuable, particularly when they had manipulated designs to increase utility balance. We did not strive for utility balance in this study, and that may be the reason constraints were not helpful.

We think ACBC is a potentially useful concept, and are interested in exploring different ways to develop preliminary estimates of partworths to be used for study design. Our current inclination is avoid self-explicated information altogether, relying instead on preliminary information from a sample of similar respondents. And we are inclined to update our estimates of each respondent's partworths following each choice task. Two schemes have occurred to us, which we would like to test in a further study.

One method would first conduct a latent class analysis on data from a pilot test of the questionnaire with similar respondents. The initial estimate of each respondent's partworths would be the mean of the partworths for that sample. However, after the respondent answers each choice task, we would compute his probability of belonging to each class, and form a new estimate of his partworths by using those probabilities as weights. From time to time the data for all respondents could be re-analyzed to update the latent class parameters.

The other method would be in the spirit of HB. We would again start with a pilot sample, which would be analyzed to estimate the means and covariances for the distribution of partworths. After the respondent answers each choice task, partworths would be estimated which have maximum posterior probability, considering both the likelihood of his/her responses to date and also the probability distribution of partworths. By the time a few questions had been answered, we should achieve a reasonable estimate of the respondent's partworths. From time to time the data for all respondents could be re-analyzed to update the population parameters.

Both of these schemes would have the advantage of not requiring self-explicated judgments of respondents' partworths, and both would profit from borrowing information from other respondents.

The idea of Adaptive Choice-Based Conjoint analysis remains intriguing. Success appears to be more illusive than we expected when we began this study, but we remain hopeful. Perhaps we will have a positive report at the next Sawtooth Software Conference. In the interim, users of CBC can be happy with the robustness of its design approach.

APPENDIX

THE ADAPTIVE CBC DESIGN PRINCIPLE

Assertion: The most productive task to add to a multinomial logit design consists of one whose design vectors lie in the space spanned by the characteristic vectors of the Information Matrix corresponding to its smallest characteristic roots.

Any positive definite symmetric matrix **S** can be expressed in terms of its characteristic roots and vectors

$\mathbf{S} = \mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}'$

Where $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}$,
$\mathbf{\Lambda}$ is diagonal with diagonal elements $\lambda_i >= 0$ and in descending order.
The roots of **S** are diagonal elements of $\mathbf{\Lambda}$, and the vectors of **S** are columns of **Q**.

The determinant of a product is the product of the determinants. The determinant of an orthogonal matrix such as **Q** is unity. Therefore $|\,\mathbf{S}\,| =$ the product of the roots λ_i.

The partial derivative of $|\,\mathbf{S}\,|$ with respect to each root of **S** is the product of the other roots. Therefore, $|\,\mathbf{S}\,|$ can be increased most rapidly by increasing its smallest roots.

The multinomial logit Information Matrix is obtained by probability-centering each column of the design matrix for each task, and then accumulating sums of squares and cross products of the resulting matrices, where the squares and cross products for each alternative are weighted by its probability of choice.

Let **Z** be a matrix of column-centered design vectors for previous tasks, with each row multiplied by the square root of that alternative's choice probability. Then the Information Matrix $\mathbf{S} = \mathbf{Z}'\mathbf{Z}$. We are looking for a set of design vectors $\mathbf{X_n}$ for a new task, so that when each column of $\mathbf{X_n}$ is probability-centered and each row is multiplied by the square root of its choice probability to get $\mathbf{Z_n}$, the new Information Matrix

$\mathbf{S_n} = \mathbf{S} + \mathbf{Z_n}'\,\mathbf{Z_n}$ will have the greatest determinant. By the argument above, $\mathbf{Z_n}$ should be chosen to lie in the subspace spanned by the characteristic vectors of **S** corresponding to its smallest characteristic roots.

The procedure used in the adaptive design algorithm:

We work with $(\,\mathbf{S} + \mathbf{I}\,)$ rather than **S** itself, since adding a constant to the diagonal of **S** does not affect its characteristic vectors, but does cause all roots to be positive so it can be inverted.

The characteristic vectors corresponding to the largest characteristic roots are easy to find. Since the characteristic vectors of a matrix are the same as those of its inverse, we invert $(\,\mathbf{S} + \mathbf{I})$ and find the largest characteristic roots and corresponding vectors of the inverse. Those are then modified to make rows of the design matrix for the new task.

In the earlier JHB study a simple procedure was used to convert characteristic vectors to design vectors. For a choice task with **k** alternatives, **k – 1** characteristic vectors were used. The $\mathbf{k_{th}}$ vector was taken as the negative of the sum of the first **k – 1** characteristic vectors. All **k** vectors were partitioned into sections corresponding to different attributes.

To create the design vector for the first alternative, the largest element in each section of the first characteristic vector was set to unity and all other elements in that section were set to zero. A similar procedure was used for each successive alternative, except for a procedure that ensured minimal overlap. No element was set to unity if one in the same position had been set to unity previously in that task and any other unused elements were available.

This procedure was modified in the current study. Rather than assuming a one-to-one relationship between characteristic vectors and design vectors, an additional step was used to find design vectors that lay in the desired subspace but did not correspond one-to-one to characteristic vectors. This was accomplished with a rotational procedure similar to the Varimax procedure used in factor analysis. However, rather than maximizing the sum of fourth powers as in Varimax, this procedure sought an orthogonal rotation of the characteristic vectors which maximized the sum of third powers. This tended to produce vectors with few large positive elements and many small negative elements. Following this rotational procedure, the rotated vectors were converted to ones and zeros as before.

Initially all characteristic roots of $(\mathbf{S} \ + \mathbf{I}\ )^{-1}$ are equal to unity, so the characteristic vectors are arbitrary and the first task is chosen randomly. There is considerable randomness in selection of the first few choice tasks, but eventually each new task contributes more critically to the determinant of the Information Matrix.

Table 1
Attributes and Levels

Attribute 1: Brand:
1. IBM
2. Dell
3. Compaq
4. Toshiba
5. Acer

Attribute 2: Display:
1. 14-inch
2. 15-inch
3. 17-inch

Attribute 3: Processor Speed:
1. 2.0 GHz
2. 2.4 GHz
3. 3.0 GHz

Attribute 4: Battery Life:
1. 3 hours
2. 4 hours
3. 5 hours

Attribute 5: Weight:
1. 3 pounds
2. 5 pounds
3. 7 pounds

Attribute 6: Operating System:
1. Windows XP Professional
2. Windows XP Professional + Microsoft Office
3. Windows XP Home

Attribute 7: Pointing Device:
1. Center post
2. Touch pad
3. Track ball

Attribute 8: Exterior Material
1. Black composite
2. Silver aluminum
3. Gunmetal titanium

Attribute 9: Price:
1. $ 800
2. $1,200
3. $1,600
4. $2,000
5. $2,400

Table 2
Average Part Worths

	Partial Profile	**Full Profile**
IBM	16.59	15.33
Dell	61.78	69.09
Compaq	2.18	8.33
Toshiba	-12.80	-18.41
Acer	-67.75	-74.33
14-inch	-38.63	-17.04
15-inch	5.17	1.18
17-inch	33.46	15.87
2.0 GHz	-48.13	-33.48
2.4 GHz	6.74	-0.09
3.0 GHz	41.38	33.57
3 hours	-40.39	-20.67
4 hours	4.70	2.91
5 hours	35.70	17.76
3 pounds	25.67	15.20
5 pounds	13.61	2.70
7 pounds	-39.28	-17.90
Windows XP Professional	-13.26	-5.95
Windows XP Professional + Microsoft Office	27.83	18.83
Windows XP Home	-14.57	-12.88
Center post	-29.71	-13.09
Touch pad	31.42	17.89
Track ball	-1.71	-4.80
Black composite	-1.10	-4.74
Silver aluminum	3.20	1.08
Gunmetal titanium	-2.09	3.66
$800	42.97	70.83
$1,200	46.10	56.16
$1,600	18.38	37.01
$2,000	-26.38	-35.46
$2,400	-81.07	-128.53

Table 3
Average Importances

	Partial Profile	**Full Profile**
Brand	17.35	20.36
Display	9.38	6.20
Processor Speed	11.15	9.63
Battery Life	9.15	6.66
Weight	9.05	6.43
Operating System	10.28	7.79
Pointing Device	10.01	8.72
Exterior	5.20	4.86
Price	18.44	29.35

REFERENCES

Chrzan, K. (1999) "Full Versus Partial Profile Choice Experiments: Aggregate and Disaggregate Comparisons," *Sawtooth Software Conference Proceedings*, Sequim, WA, 235-248.

Huber, Joel and Klaus Zwerina (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, (August), 303-317.

Patterson, M. and K. Chrzan (2003) "Partial Profile Discrete Choice: What's the Optimal Number of Attributes," *Sawtooth Software Conference Proceedings*, Sequim, WA, 173-185.

Toubia, O, Hauser John R. and Duncan I Simester (2004) "Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis," *Journal of Marketing Research*, 41, (February), 116-131.

Sawtooth Software, *Sawtooth Solutions*, Newsletter, (Winter, 2004).