PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

March 2006

FOREWORD

This volume reflects the proceedings of the twelfth Sawtooth Software Conference, held in Delray Beach, Florida, March 29-31, 2006. This conference was our most successful in terms of attendance in quite some while, drawing 170 participants to the beautiful beachside Delray Beach Marriott hotel.

The focus of this conference continues to be quantitative methods in marketing research. The authors were charged with delivering presentations of value to both the most and least sophisticated members of the audience. Topics included conjoint/choice analysis, market segmentation, MaxDiff, general web interviewing, agent-based simulation, customer satisfaction, brand image research, and "build your own" choice tasks.

The authors also served as discussants to other papers delivered at the conference. These discussants spoke for about five minutes to express contrasting or complementary views. This year, we urged the discussants to put even more emphasis on a thoughtful critique of the papers, and we think that the commentary achieved a higher level than in past conferences. A few of the discussants have prepared written versions of their comments for this volume.

The papers and discussant comments are in the words of the authors, with generally very little copy editing performed. We express our gratitude to these authors for continuing to make this conference one of the most useful and practical quantitative methods conferences in the industry.

Sawtooth Software

July, 2006

CONTENTS

Summary of Findings

The twelfth Sawtooth Software Conference was held in Delray Beach, Florida, March 29-31, 2006. The summaries below capture some of the main points of the presentations. Since we cannot possibly convey the full worth of the papers in such few words, the authors have submitted complete written papers within this 2006 Sawtooth Software Conference Proceedings.

Putting the Ghost Back in the Machine (Andrew Jeavons, Nebu USA): In this presentation, Andrew emphasized the difference in the level of human interaction with market research surveys as we've progressed from paper-based instruments, to phone, to CAPI, and now to web-based surveys. Human interviewers lead to different biases, but also certain benefits (such as ability to keep respondents from terminating prematurely). Andrew also reported results from a study that segmented respondents into introverts and extroverts. Respondents saw one of two versions of the web-based questionnaire: a colorful version or a plain version. He compared how introverts vs. extroverts responded to the plain or colorful surveys. He suggested that web surveys in the future might adapt to the personality and characteristics of the respondent to yield the highest completion rates and best quality data. He also considered the fact that the change in question modality between the web and CATI could cause a "cognitive shift" which could affect the processing of questions due to the absence of an interviewer and this may be the basis for so-called modality effects. It may be that the web causes a cognitive shift that for some types of surveys may be unwanted.

Scalable Preference Markets (Ely Dahan, UCLA, Arina Soukhoroukova and Martin Spann, University of Passau): Ely presented a novel way to measure preferences for attribute levels or product concepts in the form of a stock-trading game. Respondents are first trained how to use an internet-based stock trading system. However, rather than trade stocks, the respondents traded product concepts or attribute levels. The price of a “stock” is established based on how each trader thinks the general market (the other participants in the game) will end up valuing each concept or level. Each respondent buys and sells stocks with the goal of winning: having the most total wealth at the end of the game (cash plus value of stocks). Through actual tests of this process, Ely and his colleagues found that respondents enjoyed the stock-trading game much more than taking conjoint surveys. They also found that the resulting preferences from the stock-trading game and conjoint surveys were quite similar. One weakness of the approach is that it provides no ability to estimate individual-level models.

Assessing the Integrity of Convergent Cluster Analysis (CCA) Results (John A. Fiedler, Oreon Inc.): Is cluster analysis art or science? That was John’s initial question to the audience. Cluster analysis procedures always find clusters, whether actual delineation exists or not. The question becomes how to quantify the delineation between clusters, and whether such measures can justify the existence of segments. John cited early work by Sawtooth Software founder Rich Johnson on a measure called Cluster Integrity. Intuitively, it is the density of the mid region (between segments) compared to the densities of the regions near the cluster centers. John compared this measure of cluster integrity to the standard measure of reproducibility offered in CCA software (the ability of the algorithm to assign respondents into the same cluster given different starting points). He found little correlation between the measures for synthetic and real data sets. John felt that Cluster Integrity is a useful metric and could be employed when finding cluster solutions. He voiced his desire to see a new version of CCA software that featured both algorithmic and user-interface improvements. And, he concluded that cluster analysis really should be science.

Identification of Segments Determined through Non-Scalar Methodologies (John Pemberton and Jody Powlette, Insight Express): Many researchers are turning to non-scalar methods (such as MaxDiff and ranking tasks) to measure the importance of items and for use in developing segments. However, creating subsequent typing algorithms to assign new respondents into existing segments with these non-scalar techniques is much more challenging than with traditional ratings-based scale techniques (which can use methods such as discriminant analysis). John fielded a comparison study and found that 5-pt Likert scales, ranking tasks, and MaxDiff tasks produced very similar relative weights for the items. John described two methods for finding an efficient subset of pairwise comparisons that could identify segment membership: a discriminant-based approach and a Bayesian updating approach. The Bayesian approach performed slightly better in tests to predict segment membership within holdout samples.

Reverse Segmentation: An Alternative Approach (Urszula Jones, Curtis L. Frazier, Christopher Murphy, Millward Brown, and John Wurst, SDR/University of Georgia): One of the classic problems with cluster segmentation research is that segments developed based on attitudes or behavior variables rarely show many important differences when cross-tabbed against demographic or other targetable characteristics. Of course, clients prefer segments that differ on attitudes and that are also identifiable in the marketplace. Ula presented a method called Reverse Segmentation that helps solve these issues. Rather than cluster on each respondent, respondents are collapsed into objects based on the demographics or other targetable variables that seem to show significant differences among the attitudinal variables (determined through a series of ANOVA runs). For example, one respondent object may be composed of respondents with college education, low income, male gender, and having children. For each object, means are computed on the attitudinal variables. These objects are then clustered (based on the means of attitudinal variables). Ula presented results for a real dataset showing that the new approach produced segmentation schemes with significant differences among both basis and demographic variables.

Testing for the Optimal Number of Attributes in MaxDiff Questions (Keith Chrzan, Maritz Research and Michael Patterson, Probit Research): MaxDiff is becoming a mainstream procedure for estimating the importance or preference of items. One of the key decisions in designing a study is how many items to show within each set. Keith presented results from three different methodological studies that featured different numbers of items per set. Respondents were randomly divided into groups receiving as few as three to as many as eight items per set, with the total number of sets held constant. Increasing the number of items had a strong effect on time to complete each set. Three items per set seemed to have lower predictability. But, four to eight items per set performed about at parity. Keith highlighted that the extra time to ask seven or eight items per task may be put to better use asking more sets containing four to five items. He concluded that four to five items per set is about the right number to balance respondent difficulty and achieve high precision of estimates. One concerning point from this research is that the parameters often showed statistically significant differences depending on the number of items per set.

Product Line Optimization through Maximum Difference Scaling (Karen Buros, Data Development Worldwide): Karen illustrated the use of multiple variations on TURF searches for developing optimal product lines. The variations included: reach customers who like at least one item in the line; maximize the number of different items liked in the line; maximize the number who find their favorite item in the line; and maximize the "share of requirements" (similar to "share of preference") satisfied by the line. Karen argued that MaxDiff is well-suited to TURF type optimizations, because it discriminates well among items and can lead to strong predictions at the individual level. When the search size is too large, exhaustive search may be infeasible. In those situations, Karen resorted to genetic algorithms (through a modified version of the Sawtooth Software ASM module). She also showed how client-friendly spreadsheet simulators could be delivered allowing managers to play what-if analysis around optimally-designed product sets, where the results were reported as percent of respondents reached, etc.

Agent-Based Simulation for Improved Decision-Making (David G. Bakken, Harris Interactive): Agent-based simulators represent a cutting edge technology that some leading researchers are investigating for market forecasting. David described agent-based simulations as representing complex systems (such as competitive markets) that result from decisions made by a collection of autonomous actors. The actions of these actors are controlled by specific decision rules and influenced by stochastic processes. David showed examples including word-of-mouth, the interaction between consumers and automakers, and the effect of advertising on awareness. The examples were programmed in Microsoft® Excel, but David also described NetLogo, an agent-based simulation toolkit. David concluded that the goal of agent-based simulations should not be a point-estimate prediction, but to achieve a distribution of outcomes that reflect the impact of varying conditions.

How Many Choice Tasks Should We Ask? (Marco Hoogerbrugge and Kees van der Wagt, SKIM Analytical): Marco began by citing earlier work on this same subject by Johnson and Orme, which focused on this question with respect to aggregate models. The key difference in this presentation was the emphasis on individual-level models (as estimated through HB). Marco presented a new way to think about the required number of tasks per respondent. He proposed that for each respondent, we can observe whether the addition of new tasks improves the ability to predict holdout tasks. Once the holdout predictability improves little, then he argued that the results have stabilized (converged) and additional tasks are of little value. Marco and Kees used cluster analysis to summarize the marginal improvements of additional tasks in predicting holdout tasks for relatively homogeneous groups of respondents. They found that for a number of data sets, holdout predictability stabilized for all groups after about the tenth choice task.

Sample Planning for CBC Models: Our Experience (Jane Tang, Warren Vandale, and Jay Weiner, Ipsos Insight): Jane and Jay shared their experience of being in a company that executes many conjoint-related studies every year. A time-consuming part is discussing sample size during the early planning and bidding of projects. A client may have a general idea about an attribute list and may want early direction regarding rough sample size. To respond to these repeated requests in their organization, they built a spreadsheet tool that uses simple inputs to suggest reasonable sample sizes. They stressed that this tool does not replace the formal tests for design efficiency utilized by marketing scientists; but rather that it is useful for quick and rough initial direction. Jane and Jay extended the formula originally proposed by Johnson for aggregate CBC sample planning. Their extension accounts for percent usage of None, increased model complexity due to many attributes, and the effect of projected homogeneity of the sample. They demonstrated use of the tool with some practical examples.

Brand in Context: Brand Definition in Volatile Markets (Andrew Elder, Illuminas): What happens when an established brand in one marketplace seeks to leverage its brand equity to branch out into a new product category not typically associated with that brand? Such is the case with the "triple-play," wherein telecom, cable, and internet providers seek to leverage their brand equity to provide services in related, but somewhat different spaces. Andy reported the results of a web-based survey that asked respondents about their perceptions and preferences on issues related to triple play, and examined the impact of brand modeling across multiple categories. When modeling adoption of the merged triple-play offering, traditional brand attributes carried similar weight as in a traditional "single-category" model, demonstrating that core branding components are not strictly bound to their established category. Yet Andy also found that certain attributes related to the category itself had a significant impact on triple-play adoption, suggesting clear contextual effects that affected the ability for certain brands to leverage their reputation into an emerging category. Andy concluded that brand relevance is shaped by perceptions and usage of the category with which the brand is associated, and that category characteristics should be considered an integral part of brand modeling and brand communication.

Brand Positioning Conjoint: A Revised Approach (Curtis L. Frazier, Urszula Jones, and Katie Burdett, Millward Brown): In a previous Sawtooth Software conference, Frazier and co-authors showed how to decompose the brand part worth into image elements through a two-stage approach. In the first stage, individual-level part worths for brand were estimated using a standard CBC approach followed by HB estimation. In the second stage, the part worths for brand were regressed on the ratings for brands on a variety of image attributes. In this updated presentation, the authors showed how the same type model could be built under single stage estimation. This is done by incorporating the brand ratings on the multiple items within the same design matrix as the CBC tasks. They conclude that the one stage approach is more robust. The information provided by this model helps clients understand the components of brand strength, and better understand the impact on product choice by strengthening the brand in terms of the image elements.

Rethinking (and Remodeling) Customer Satisfaction (Lawrence Katz, IFOP (Paris)): Lawrence suggested that there often is a misunderstanding concerning the goals of satisfaction research. The models often used reflect a structural analysis of the drivers of satisfaction which while interesting to clients at first are too stable over time to be useful as a tool for tracking. One common question is whether to place the overall satisfaction question at the beginning of a battery of image components or at the end. Lawrence suggested that this was more than just a simple methodological issue and that the two placements measure two distinctly different psychological constructs that he called "surface" and "deep" satisfaction. Whereas the latter is more appropriate for structural modeling using the usual additive models, it is surface satisfaction that best captures what respondents spontaneously think (and say) and the short-term consequences that might result. Surface satisfaction can also be modeled, but by using methods based on open-ended response data and coding schemes that allow for asymmetric attribute effects on satisfaction. Lawrence concluded by suggesting that a program for measuring satisfaction should focus on regular studies of surface satisfaction and relatively long periods between modeling of deep satisfaction.

Dual Response "None" Approaches: Theory and Practice (Chris Diener, King Brown Partners, Inc., Bryan Orme, Sawtooth Software, Inc., and Dan Yardley, King Brown Partners, Inc.): Dual-response None involves asking about the None alternative in a separate question. Respondents first choose among available products and then separately indicate if they would actually buy the product they chose. Another way of phrasing the second-stage question is to ask whether respondents would buy *any* of the products available in the task. These two approaches reflect ways to ask questions and to model the results. Dual-response None provides a safety net when the None usage is relatively high, because information about the other attributes and levels is not lost when a None response is recorded. The None parameter is much higher with the dual approach. Dual-response None may be modeled with standard MNL software, customized approaches (with the likelihood function as the joint probability across the two questions), and also using Sawtooth Software's CBC/HB v4. All lead to similar results. Chris also found that modeling the dual-response None as a choice between the chosen alternative and None vs. a choice between all alternatives and None changed the size of the None parameter. Chris suggested asking respondents to consider the None with respect to the chosen alternative, but to model the results as if all alternatives were being compared to None.

"Must Have" Aspects vs. Tradeoff Aspects in Models of Customer Decisions* (John R. Hauser, MIT, Ely Dahan, UCLA, Michael Yee, MIT, and James Orlin, MIT): Standard conjoint analysis has been analyzed using a compensatory model, where deficiencies in one attribute can be made up for by strengths in other attributes. However, evidence suggests that respondents use non-compensatory strategies to choose product concepts. John described different heuristic rules that respondents might apply to screen products on certain characteristics. For example, a buyer might say: "I will consider flip phones, with mini-keyboards, from Blackberry." John reviewed a practical method to infer the best lexicographic description of respondents' (partial) rank data and showed results demonstrating that a non-compensatory model produces hit rates on par or better than going best-practices compensatory models. The non-compensatory model additionally yields insights for management regarding what aspects respondents screen on. The described method can be used with either traditional card-sort conjoint or choice-based conjoint.

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

External Effect Adjustments in Conjoint Analysis (Bryan Orme and Rich Johnson, Sawtooth Software, Inc.): The market simulator is the most practical and useful deliverable of a conjoint analysis study. However, due to assumptions in conjoint analysis, the results usually don't match actual market shares. Many researchers adjust conjoint models to better predict actual market shares. Bryan showed a method for adjusting for unequal distribution that avoids IIA assumptions and incorporates appropriate differential substitution effects. He also addressed the issue of scale factor, and how it related to random noise in buyer behavior. Bryan argued that adjustments for distribution and scale factor are theoretically defensible, given appropriate data. After that, some researchers additionally adjust the models to predict shares. Bryan showed that the standard Sawtooth Software external effect adjustment doesn't perform as well as adjustments made to individual-level part worth utilities. Respondent weighting was also tested, but shown to change the behavior of the simulator in some extreme ways. Bryan emphasized that the use of external effects is a dangerous practice, and should be avoided whenever possible. But, if a project requires adjustments for forecasting purposes, some adjustments work better than others.

Confound It! That Pesky Little Scale Constant Messes up Our Convenient Assumptions (Jordan Louviere, University of Technology, Sydney and Thomas Eagle, Eagle Analytics, Inc.): Jordan reviewed the issue of scale factor: the size of MNL parameters is inversely related to error. As error in responses increases, the size of the estimated MNL parameters decrease, and vice-versa. For that reason, Jordan explained, MNL model parameters cannot be identified unless the scale factor is set to a constant. Without comparable scale factors, it is not appropriate to directly compare part worths from choice studies across respondents. Predictions from simulators can also differ significantly due to scale factor. Scale factor varies between consumers, between questionnaire instruments, and due to environmental differences. These issues affect HB and latent class models as well. Jordan described the use of covariance heterogeneity models that capture scale effects as well as mean effects. Using these models, he showed that respondents reflected a distribution of scale factors as well as a distribution of parameter estimates. Jordan argued that failure to pay attention to this (in random coefficient models) can result in biased and misleading inferences and predictions.

Estimating Attribute Level Utilities from "Design Your Own Product" Data—Chapter 3 (Jennifer Rice and David G. Bakken, Harris Interactive): Jennifer and David presented a third paper in their series of investigations into Design Your Own Product (DYOP) questions. They discussed how this question type may be more realistic for certain purchase contexts. This final chapter focused on trying to estimate stable parameters at the individual level, and comparing the parameters to those from a standard CBC experiment. To estimate parameters at the individual level with DYOP, they included self-explicated questions on each of the levels in the study. They also employed HB analysis to combine information from the self-explicated questions with the DYOP choices. To estimate price parameters for each item, they included the relative price of each item (with respect to the total configured product's price) in the design matrix. They achieved reasonably good predictions of holdouts with their model. The price sensitivity parameters differed significantly between CBC and DYOP, and DYOP price sensitivities were often much higher than those from CBC.

Simulating Market Preference with "Build Your Own" Data (Rich Johnson and Bryan Orme, Sawtooth Software, Inc. and Jon Pinnell, MarketVision Research): BYO (Build Your Own) tasks have received some interest over the last 10 years in the literature, and also at this conference. In BYO tasks, respondents design their optimal product, based on the features specified at given prices. Rich described an experiment that aimed to measure price sensitivity by feature by varying the prices across respondents. Respondents were also given a CBC questionnaire, to compare the results to BYO. BYO data can be analyzed using counts or through MNL by assuming that the respondent made one choice from the universe of all possible product design combinations. But, such an MNL model is often impossible to estimate with standard MNL software, as there can be billions of alternatives. Rich showed that counting data give essentially the same answer as the complex MNL. He also showed that simulators can be built using the logs of count probabilities as pseudo utilities. The part worths differed significantly in some cases from CBC, and there were definite context effects in BYO data. The between-respondents price variations did not lead to stable price sensitivity estimates for BYO data for Rich's study, and he noted that much larger sample sizes would be needed. Rich suggested that use of CBC or BYO should depend on the choice process one wants to model, and one is not simply a substitute for the other.

PUTTING THE GHOST BACK IN THE MACHINE

ANDREW JEAVONS
NEBU USA

In 1967, the philosopher and polymath Arthur Koestler wrote a polemic diatribe against what he saw as the dehumanization of man by contemporary psychology. The title of this book was "The Ghost in the Machine." Its targets were the theories of B.F. Skinner, the well known behaviorist psychologist.

Skinner's theories had gained ascendancy since the mid 1930's. Skinner sought to explain all human behavior in terms of stimulus-response (SR) relationships. He believed that operant conditioning, the pairing of learned responses to stimuli, was the basis of all human behavior. In this focus Skinner sidelined conscious cognition as an irrelevant construct. Indeed it could be said that he saw conscious and unconscious thought as an epiphenomenon of the SR relationships that really governed human behavior.

In 1957 Skinner published the book "Verbal Behavior." The aim of this book was to explain human speech and its acquisition in SR terms. Not unsurprisingly, the book met with some fierce criticisms from a wide range of philosophers and psychologists, Koestler being one of them. Noam Chomsky, a noted philosopher of language, wrote a withering review of Skinner's book (Chomsky 1959). This review is now seen by many as the birth of "cognitive psychology," in reaction to Skinner's sterile view of human behavior.

Koester, among others, thought that Skinner's view of the mechanisms governing how people behaved was simplistic, naïve and failed to explain the totality of human action. The rush to appear "scientific," he thought, caused us to miss what is fundamental to human behaviour. Why was there war? Why did people love and hate? For Koestler, behaviourism couldn't explain these highly important behaviors and so to him was bankrupt.

At the time it was written it could be argued that the Koestler book was bordering on the quixotic. The answers to the question he was asking have been debated for millennia; there's no reason to think that this shouldn't continue. However Skinner's theories (although Skinner always argued that he was atheoretical) did cause psychology to ignore some avenues of research (such as memory and reasoning), and in this respect Koestler and more importantly Chomsky played an important part in re-orienting psychology.

Koestler's title—"The Ghost in the Machine" refers to the element of humanity in human behavior. He was trying to re-introduce a view of people and how they behave that included elements of rationality and emotion, rather than an emphasis on pure SR mechanisms which leave no room (deliberately) for any intervention of thought or other "woolly" concept in behavior. His attack was on the concept of people as "black boxes," where no attention is paid to what may mediate behavior.

Even Skinner had cause to acknowledge that all behavior is not quite simply a question of stimulus and response. A Skinnerian researcher discovered what was termed "superstitious behavior." In a normal operant conditioning paradigm a pigeon (or some other animal) would be placed in a box (called a Skinner box). A stimulus, such as a light or noise, would then be used to cause the animal to produce the desired response (pushing a lever, for instance). One researcher decided to film what the animal actually did in the Skinner box when it was performing the desired response. What they saw surprised them. The animal was performing the desired response, but it was also performing activities that had no relation to the desired response. Standing on one leg, turning around, a variety of behaviors were observed. These were termed superstitious behavior in that for some reason the animals had learned them, but they had no relationship to the stimulus and the reward they were given. The black box wasn't so simple, while the desired behavior had been learnt, others had been "attached" to it.

WEB INTERVIEWS

In the early to mid 90's the advent of web interviewing was the answer to a prayer for much of the market research industry. The decline in response rates for CATI and mall intercept was creating a crisis for data collection. When it became clear that the web could be used to conduct surveys, web interviewing quickly became a mainstream data collection method.

Using the web for interviewing reverses a trend in personal contact. Below is a table showing the communication mechanisms that mediate the various forms of interviewing:

Type of Interviewing	**Communication Modes Used**
Paper interview (face to face)	Verbal, paralinguistic, interpersonal
CAPI	Verbal, paralinguistic, interpersonal
CATI	Verbal, paralinguistic
Web (WAPI)	Written, graphic

In Europe, CAPI interviewing increased during the 90's, as it was regarded as producing higher quality data than CATI. In the commercial sector of the USA, CAPI never enjoyed much success, although paper interviewing and mall intercept had been popular ways of collecting data. Declining cooperation rates in the malls and economic factors lead to the growth of CATI, but this then began to suffer from a decline in cooperation. It should be noted that in the government/social research domain CAPI is used widely due to data quality concerns and still enjoys a high profile in the USA and around the world.

Using the web is a volte-face in terms of communication with respondents. CATI restricted the mode of interaction with respondents, but there was still a live person mediating the interview. WAPI replaces the interviewer with a block of technology; there is no verbal, paralinguistic or interpersonal communication, just a web browser.

We lost the ghost in our machine—the interviewer, and replaced a concept of the respondent as part of an interaction with the "data source" theory of respondents. Respondents in web interviews push buttons and we collect responses. We assume they are valid and that we can cut off the human interaction aspect of interviewing with impunity. There are many economic reasons to get rid of interviewers, not least the huge management cost and headache of interviewers. It's easier to deal with technology (although that is not without its caprices) than interviewers. Interviewers may be seen to corrupt or vary the way the interview is conducted, and this quest for being scientific (the use of the terminology of the "instrument" for a questionnaire points to this) leads to the idea that eliminating interviewers can be a good thing. It will lead to more accurate or truthful data, and we are being more scientific with reliable technology than fallible, human interviewers.

Interviewers are useful in some contexts; the concept of refusal converters belies the view of interviewers as pure sources of noise in data collection. Interviewers exert an effect on an interview and it is not all to the detriment of the data.

THE WEB RESPONDENT LANDSCAPE

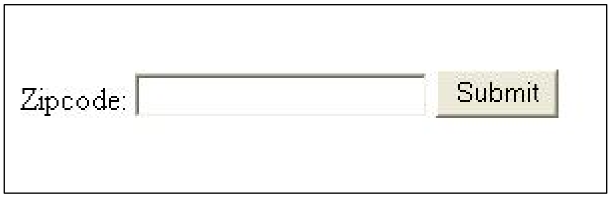
Whatever modality we use we need completes. In CATI and CAPI, interviewers—if they were good—produced completed surveys. The implication of variable performance of interviewers implies that some were better than others, and the variance these "good" interviewers produced was beneficial. With the web we have also seen the growth of panels, and it's clear that how companies treat their panelists has a commercial impact. One researcher coined the phrase "spanking the panel." By that they meant exposing panelists to a survey that may hasten their exit from the panel, but having no choice because of client demands. Where no panel is used, incidental interviewing akin to mall intercept still demands that respondents are engaged and motivated to complete the survey accurately.

One other phenomenon has developed on the web, that of professional respondents. Web sites such as www.surveypolice.com have developed with the apparent aim to help professional web respondents deal with online survey companies. It may be that professional respondents are not all bad; it depends on the accuracy of the data. There is a body of research that indicates that more experienced respondents give different answers; on the other hand it's been known for years that experienced interviewers tend to edit (or truncate) open ended comments more than newer interviewers. Perhaps we have just substituted one filter for another. Of the greatest concerns for web panels is the problem of engagement and veracity with web respondents. Respondents above all have to be giving accurate information—professional or not. They have to be interested in providing information—not just getting a reward at the end of the survey. These two factors are the biggest challenges facing web interviewing now.

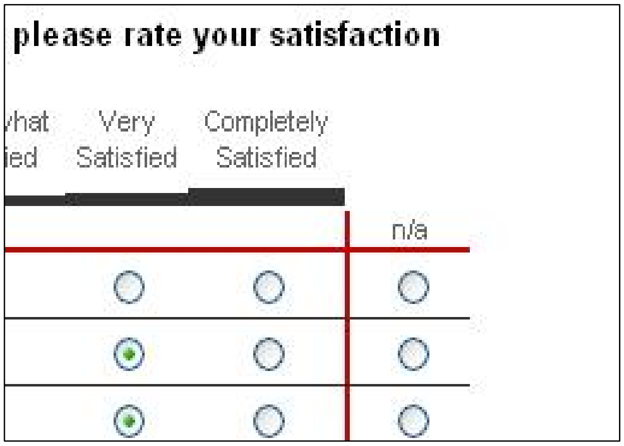
WEB INTERVIEWS AND RESPONDENT VARIATIONS

Given these issues, we have to ask ourselves how good we are at designing web surveys that engage the respondent and promote accurate responding. Sherman *et al.* (2001) looked at web home page creators. Creators of home pages overestimated how effective the pages were in promoting a positive self image of the web page creator. It could be that the same can be said of web surveys. We may overestimate how engaging they are. We may be overconfident in our abilities to design effective web surveys. We may also overestimate the respondent's ability to communicate and navigate the interview.

Here is a "question" from a recent web survey:



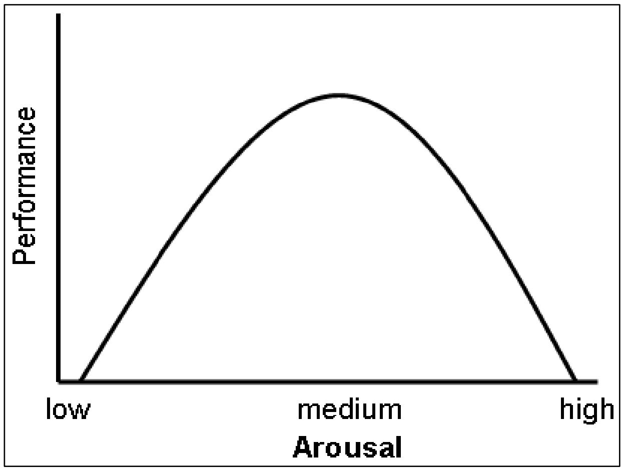
It's hard to see how this is engaging. We wouldn't use such brevity of speech when doing a CATI or CAPI survey.



please rate your satisfaction

/hat ied	Very Satisfied	Completely Satisfied	n/a
	○	○	○
	●	○	○
	●	○	○

The question shown to the left is an improvement. But the meaning of n/a is not clear—no animals? Never Access? No action? The assumption that it means not applicable—which it probably does—is probably only made by researchers, not respondents.



Engagement in a web interview is mediated by many factors. An old, but well known, principle in psychology is the Yerkes-Dodson law of arousal (1908). This states, as shown at left, that there is an optimal level of arousal for all tasks. Too little and we under perform, too much and we under perform. We have to get the right level of stimulation to make sure respondents are in an optimal state to respond accurately.

There are different optimal levels of stimulation for different people. Using the concepts of introversion and extroversion, Eysenck (1967) after Nebylitzin (1976) cited in Bedny and Seglin (1999), some people seek stimulation (extroverts) while others avoid it (introverts). This is linked to their internal activation state. Introverts have no problems absorbing stimuli, thus they need little stimulation. Extroverts absorb stimuli with more difficulty, thus they require more. Therefore using this characterization of respondents we may be able to argue that certain types of questionnaires are optimal for different types of respondents. Of course this is a very, very crude classification, and is not without controversy, but overall it has stood the test of time in various forms, so it is at least worthy of consideration.

GETTING COMPLETES

However much market researchers want to see themselves as scientists delving for insights into the human condition with finely tuned "instruments," market research remains for the most part a commercial enterprise. Surveys need to be completed by respondents. With CATI and CAPI we can train interviewers to get completes, and indeed, employ people who are good at getting completes. An interviewer who presents an interview perfectly but never gets a complete is not much use. Interviewers use their human skills of communication, at whatever level, to get the respondent to complete the survey. That is their job. In a web survey we have lost that "ghost," getting completes becomes a question of motivating the respondents to be accurate and engaged. With interviewers there was unconscious (or conscious often) variation of the interview process to get a complete. This could be in terms of correcting errors, paralinguistic cues or any form of "relationship" building with the respondent. With a web survey we have a static set of screens that do not vary between respondents and do not vary according to how bored respondents are or how many errors they make. Panel attrition is a serious commercial problem. Giving respondents interviews that accelerate their exit from a panel has to be minimized. Interviews may need to be made more variable simply to keep respondents. While this may go against the idea of "instruments" measuring a respondent, it is in more keeping with the history of interviewing. Perhaps we should measure what type of respondent we have when we enroll the respondent in the panel or start the interview. Maybe this should determine interview style, or even check during an interview how a respondent is behaving (Jeavons 2000) and adjust the interview accordingly.

Three questions therefore came to the fore when considering interviewing on the web:

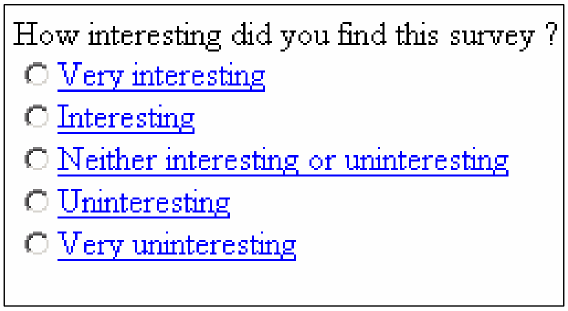
1. The veracity of responses. Can we ask a simple question and tell if we have a respondent who is telling the truth? Can we tell if the respondent simply is not paying attention?
2. Do other psychographic factors, such as personality traits, affect the interview process and the information provided by the respondent?
3. Can we make any suggestions about what, exactly, the oft quoted but never defined "modality" effect in web interviews is?

EXPERIMENTAL SURVEY

An experimental survey was designed with the aim at trying to shed some more light on the questions stated previously.

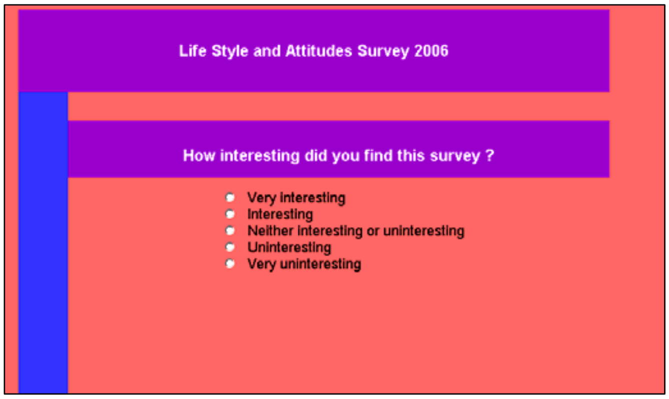
A survey was designed of around 10-15 minutes length based on common questions used in survey research, with the addition of one question derived from the work of Kahneman and Tversky (see Kahneman 2003 for an overview of their work) who have investigated judgments and concepts of personal probability for decades, culminating in the award of the Nobel Prize for economics to Daniel Kahneman in 2003. Also, at the start of the interview, two additional questions were asked. The first asked the respondent's hand preference (right, left or both); the second asked if respondents saw themselves as extroverts or introverts.

Two "environments" for the interviews were also constructed using HTML. The first condition was designated as low stimulation. A sample screen of this condition is to the left.



How interesting did you find this survey ?

○ Very interesting
○ Interesting
○ Neither interesting or uninteresting
○ Uninteresting
○ Very uninteresting



Life Style and Attitudes Survey 2006

How interesting did you find this survey ?

○ Very interesting
○ Interesting
○ Neither interesting or uninteresting
○ Uninteresting
○ Very uninteresting

The second condition was designated as high stimulation, an example screen shot is to the left.

The differences are obviously in the use of color and font/character size. Two independent groups took the surveys (which we will refer to as the low and high groups).

The sample, supplied with the cooperation of Easymail Interactive, was "gen pop," that is balanced across age and gender. For each survey a target quota of 100 completes was set. No other quotas were set. In terms of completion rates (those respondents that completed the survey once they started) the rates were:

Survey Type	**Completion Rate**
Low	93%
High	87%

For gender there were no statistical differences between the two surveys.

The breakdown for gender was:

Q'type	**Male**	**Female**
Low	37.6%	62.4%
High	41%	59%

In general, more females completed the survey than males, but across the two survey types there were no statistically significant differences.

In terms of self-reported extroversion/introversion the results were:

Q'type	**Introvert**	**Extrovert**
Low	51.5%	48.5%
High	58%	42%

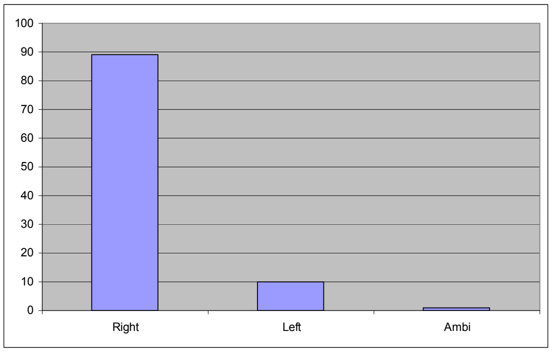
There were no statistically significant differences between the two questionnaire types in terms of reported extroversion/introversion.

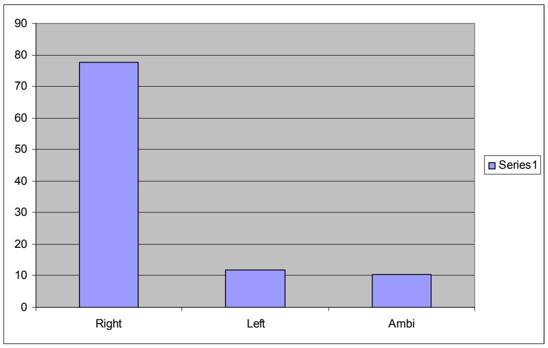
Gender	**Introvert**	**Extrovert**
Male	45.6%	**54.4%**
Female	**60.7%**	39.3%

Here we see that females tend to report being introverted more frequently, and males as being more extroverted. This is significant with p=0.022 using Fishers exact left tail test.

HANDEDNESS

Handedness is a relatively stable psychological characteristic. That is, the ratios of left to right to ambidextrous tend to be stable across populations. The "British Survey of Left" (see graph left) carried out in 1992 revealed that 89% of the British population are right handed, 10% left handed and 1% ambidextrous. Most sources agree that about 10% of the population is left handed, with a very small percentage (1 or 2 at the most) being ambidextrous.





The respondents were asked if they were right handed, left handed or able to use both hands equally well (the responses were randomized). The results across both surveys were:

Hand Preference	Reported
Left	11.94%
Right	77.61%
Ambidextrous	10.45%

The reported incidence of the ambidextrous group is clearly too high. The incidence of left handedness (sinistrality) is within bounds, but it seems that the right handed frequency is depressed. The sinistral population appears to be reporting its hand preference accurately, but within the right handed population it seems reasonable to conclude that they are not reporting the hand preference correctly. We have to conclude therefore that nearly 10% of the respondents are either extremely cognitively challenged (a 9 year old will know their hand preference) or are not replying accurately to the question. The question was the third to be asked in the survey, after gender and extroversion/introversion.

REASONING

Following the demise of behaviorism, cognitive psychology blossomed and a great deal of research has been performed on memory and logical reasoning. Kahneman and Tversky have become well known for their work on personal probability in judgments.

They developed a theory which divided decisions into intuitive and judgmental. The former takes place based on impressions of problems and situations, the latter requires conscious reasoning. Their taxonomy of judgmental processes are used to monitor (or at least should be) intuitive decisions. They referred to this as "cognitive self monitoring."

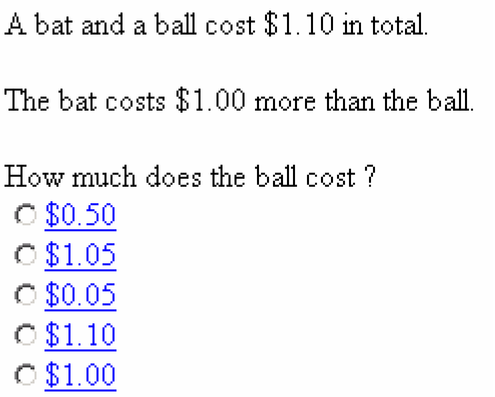
Kahneman and Frederick (2002—see Kahneman 2003) illustrated that this self monitoring may be fairly relaxed. They asked students at Princeton and University of Michigan (UMich) the following puzzle:

"A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?"

They found that 50% of the Princeton students and 56% of the UMich students gave the answer of 10¢. The answer should be 5¢. Their conclusion was that "people were simply not used to thinking hard" in everyday life, and that the self monitoring of intuitive judgments, which will fixate on an accessible aspect of the problem (here 10¢, any easy division of the $1.10 total) was very lax. Interestingly they go on to conclude that:

"…errors in this puzzle and in others of the same type were significant predictors of intolerance of delay and also of cheating behavior"

In the test survey a slightly modified version of this question was asked, in that it was made easier to get the correct answer. In fact it should have been very easy, the problem was presented as:



A bat and a ball cost $1.10 in total.

The bat costs $1.00 more than the ball.

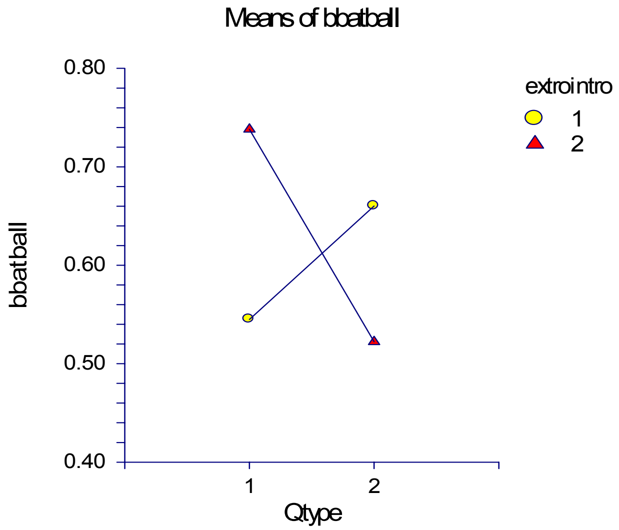
How much does the ball cost ?

○ $0.50
○ $1.05
○ $0.05
○ $1.10
○ $1.00

The "easy" answer, which provokes fixation according the Kahneman and Frederick, was not even presented (10¢), all that was required was some simple mental arithmetic which it is to be hoped that the general population was capable of. Overall the percentage of respondents who got this question wrong was 39.3%. Clearly a significant proportion of the subjects aren't paying attention to their answers, or lack simple division and addition skills. The most common wrong answer was $1.10, by a large proportion—26.37% of respondents gave this answer. This seems to confirm the accessibility theory, as this is a value contained in the question itself. The inability of respondents to do simple mathematics, such as adding numbers to 100 also surfaced in Jeavons (1999). On the other hand 60% correct is high, even with help, when some of the brightest students get it wrong at least 50% of the time. In this comparison our web respondents are holding up well, as a sample of the general population. Indeed we may expect their performance would have been considerably below the rarified sample used by Kahneman and Frederick (2002).

In terms of effects of the questionnaire type, age and extroversion/introversion, this was tested by a 3-way ANOVA using the value of the bat-ball question transformed to a binary value. Overall there were no main effects for gender, questionnaire type (high/low) and personality.

There was however a significant interaction between questionnaire type and personality ($p < 0.02$). Below are the mean plots for the interaction:



Group 1 of "extrointro" are self classified introverts, group 2 are extroverts. The variable bbatball was a recoded version of the responses to the bat-ball question so 0 is the wrong answer and 1 the correct answer. It seems that in the low stimulation questionnaire, introverts do worse than extroverts; however in the high stimulation questionnaire this is reversed. The introverts do better than the extroverts. This doesn't fit a classical activation theory of introversion and extroversion. Using that approach, extroverts should do better in the high simulation environment.

If we use the idea of cognitive self monitoring, the different personality types are showing variations in self monitoring that are produced by the questionnaire environment. For some reason a "personality consonant" questionnaire environment reduces cognitive self monitoring. So for introverts the low stimulation environment produces less accurate responses than the stimulating environment, and for extroverts the reverse is true.

MODE EFFECTS

It isn't clear what the interaction of personality type and questionnaire means—indeed it may not be a real effect. But at worst it calls for some more investigation. What is known is that there are mode effects at work in non-interviewer mediated interviews, WAPI vs. CATI/CAPI. Other mode effects have been reported, such as longer open ends on the web. The question that doesn't seem to be answered is: what is a mode effect? The term is used often but not explained. Let us accept for the sake of argument, that respondents perform the bat-ball task better on the web. Given that top university students do nearly as well as they do (and also accepting we simplified the task slightly), it still doesn't seem to be too much of a stretch to make this assumption.

There is a cognitive shift in respondents completing questionnaires on the web, and it is probably a better term to use than "mode effect." Something changes in the way they perform the tasks when they are on the web. Given the results indicating richer open ends and better reasoning it seems they think more. Is this a good thing? Dijksterhuis *et al.* (2006**)** performed a study where they tried to see if unconscious decisions vs. conscious ones were better. It's arguable that there is no such thing as an unconscious decision. Perhaps Kahnemans' "intuitive" or "low processing decision" is a better term. They also use the term, "deliberation without attention." Dijksterhuis *et al.* (2006**)** found that consumers' purchases of complex products (such as cars) were better when decisions were made without conscious deliberation. How "right" the decision was assessed by asking the respondents after a period of time when they had bought the product and used it. They also found that simple products (such as kitchen accessories) benefited from more thought—that is consumers were more likely to be satisfied with a simple product if they had thought about it more. Maybe what people think doesn't matter sometimes; maybe it depends on the products. When we come to the web, if we have an environment that enhances conscious thought perhaps this may be affecting some results. At the least the environment of the web does seem to affect cognition in a positive way for problem solving.

In market research it may be that Caesers' lament of "…yond Cassius has a lean and hungry look, he thinks too much, such men are dangerous" takes on new relevance.

References

Bedny G, Seglin M "Individual Features of Personality in the Former Soviet Union," Journal of Research in Personality 33, 546–563 (1999)

Chomsky, Noam "A Review of BF Skinners Verbal Behavior," Language 35(1):pp. 26-58, 1959

Dijksterhuis A, Maarten B, Nordgren L, van Baaren R, "Science" 17 February 2006: Vol. 311. no. 5763, pp. 1005 - 1007

Eysenck, H. J. (1967). The Biological Basis of Personality. Springfield, IL: C. C. Thomas.

Jeavons A, "Paradata: Uses in Web Surveying" Paper published in ESOMAR Monograph on Internet Market Research, September 2000.

Jeavons A, "Ethology and the Web: Observing Respondent Behaviour in Web Surveys." ESOMAR Internet Conference February 1999, London.

Kahneman D, "A Perspective on Judgment and Choice Mapping Bounded Rationality" September 2003, American Psychologist 697 Vol. 58, No. 9, 697–720

Koestler, Arthur "The Ghost in the Machine," Arkana Books, 1967

Sherman, R. C., End, C., Kraan, E., Cole, A., Campbell, J., Klausner, J., & Birchmeier, Z. (2001). Metaperception in Cyberspace. Cyberpsychology and Behavior, 4, 123–129. Cited in: Kruger J, Eply N, Parker J, Zhi Wen N "Egocentrism Over E-Mail: Can We Communicate as Well as We Think?" Journal of Personality and Social Psychology 2005, Vol. 89, No. 6, 925–936

Skinner, B. F. "Verbal Behavior" Copley Publishing (reprinted 1991), 1957

Yerkes R, Dodson J, "The Relation of Strength of Stimulus to Rapidity of Habit-Formation (1908) Journal of Comparative Neurology and Psychology," 18, 459-482

SCALABLE PREFERENCE MARKETS

ELY DAHAN
UCLA
ARINA SOUKHOROUKOVA AND MARTIN SPANN
UNIVERSITY OF PASSAU

On what dimensions do markets outperform alternative methods of measuring preferences?

SUMMARY

Scalable Preference Markets work well, and provide five potential benefits: (1) scalability to measure preferences for a virtually unlimited number of concepts, attributes, and aspects, (2) lower cost tests that can be conducted quickly and with fewer respondents, (3) better response rates from respondents most of whom prefer the method, get less fatigued, and are more interested and satisfied, (4) learning between traders who have heterogeneous preferences and trade very differently from each other, and (5) mediation of some respondent biases. Some drawbacks include the availability of aggregate preference data, making segmentation more difficult, and the need for simultaneous data collection.

Motivation

Given that product markets are complex systems fraught with uncertainty, many firms utilize market research to resolve the uncertainty. But even though market research reduces uncertainty, it adds cost and time. And it involves limitations such as consumers' biases and bounded rationality. One possible solution comes from stock markets as they are efficient aggregators of information. In prior work, prediction markets have been shown to work.

The present research seeks to measure whether stock markets measure preferences and overcome biases. Further, preference markets can be scaled up to measure many features and products. So, can preference markets reduce cost and time? And will respondents prefer preference markets over conventional surveys?

Figure 1: A Model of Preference Markets

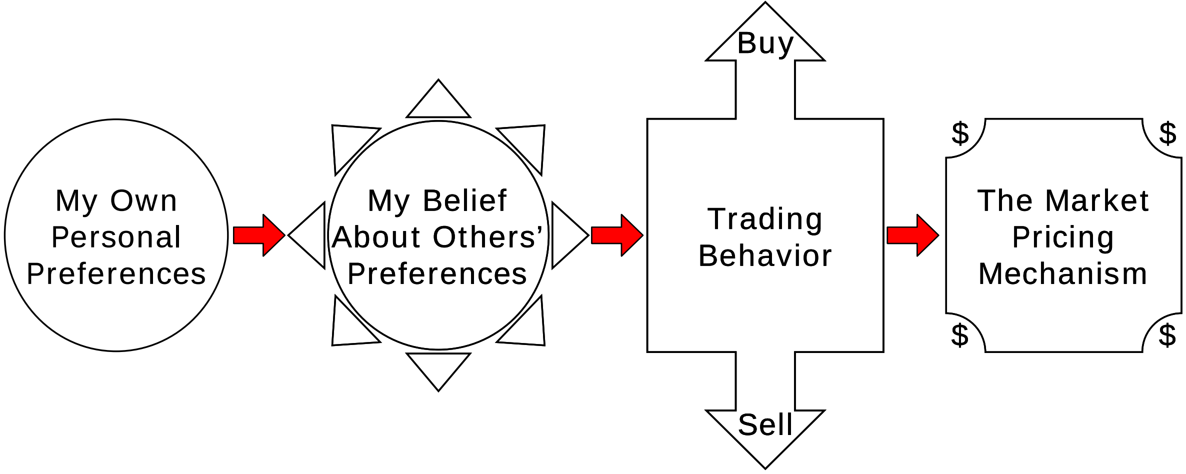
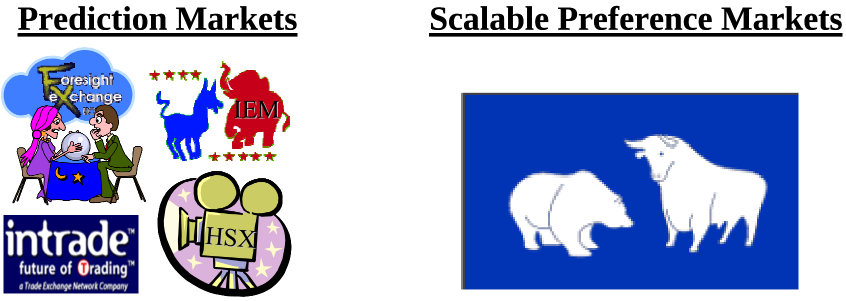

My Own Personal Preferences
My Belief About Others' Preferences
Buy
Trading Behavior
Sell
$
$
The Market Pricing Mechanism
$
$

Figure 1 illustrates the relationship between personal preferences, individual trading behavior, and equilibrium market prices. In Figure 2, we distinguish scalable preference markets, which measure preferences in under an hour, from prediction or information markets, which forecast actual outcomes over days, weeks, and longer.

Figure 2: Distinctions between Prediction Markets and Preference Markets

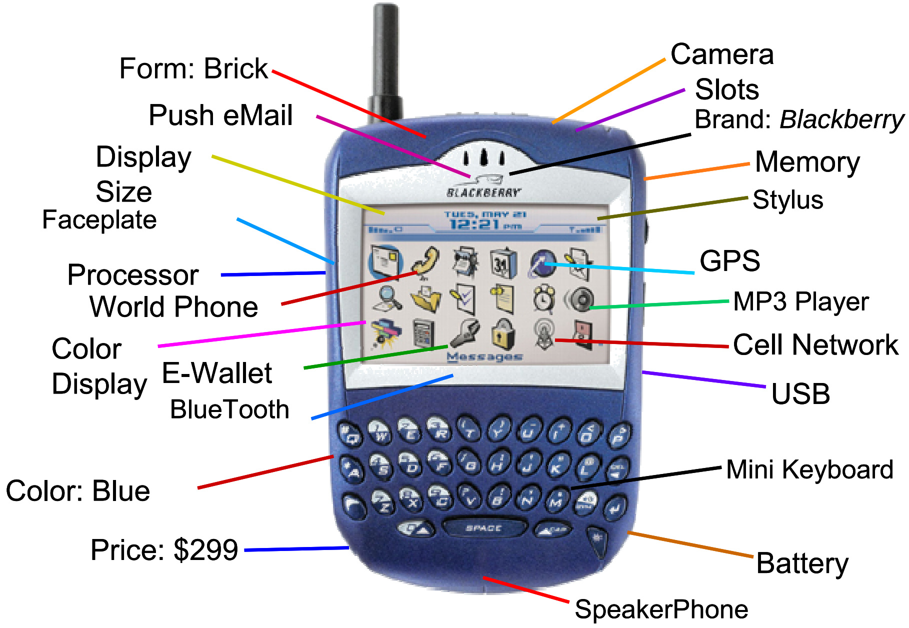


- Predict *actual* future outcomes
- Don't necessarily identify underlying reasons
- Games last days, weeks, months

- Measure preferences
- Measure preferences for concepts, features, aspects
- Games last minutes

To test the effectiveness and scalability of the method, we used the example of smart phones, including multiple product concepts and attributes, examples of which are shown in Figure 3.

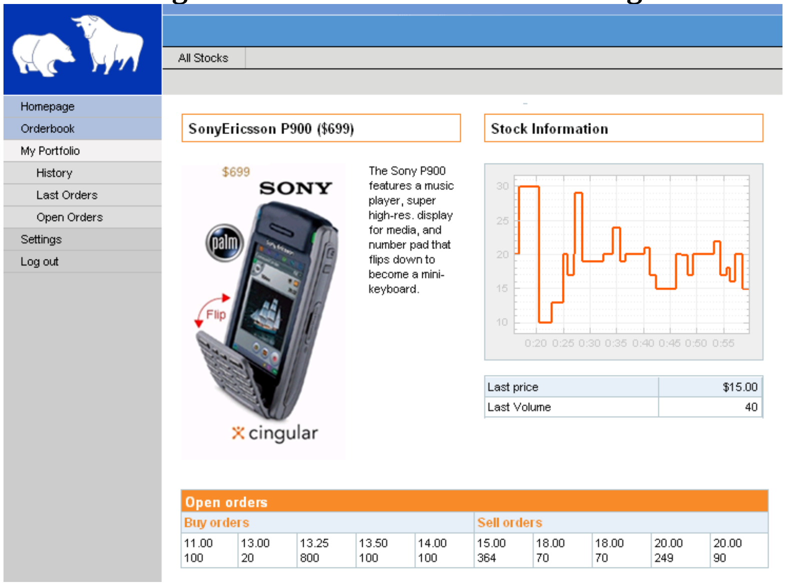
Figure 3: Sample Smart Phone Features



Relying on prior studies and pretests we selected 50 different product attributes out of a broad range of colors, form factors, software applications, and other feature categories. Nineteen (19) mutually exclusive product attributes (e.g. Cell Network Providers, form factors or colors) and thirty-one (31) “binary” options (e.g. FM Tuner or WiFi) were researched. In addition, six (6) state-of-the-art smart phone concepts were selected to provide a common anchor between the sub-markets. In all, fifty-six (56) “stocks” were created to describe the attributes and concepts.

Participants trade their expectations of the percentage of users who would choose a specific attribute for a given price (e.g., if 15% of potential buyers would select the SonyEricsson P900 at a price of $699, then its stock price should equilibrate to $15). Figure 4 shows the user interface.

Figure 4: User Interface for Trading



SonyEricsson P900 ($699)

The Sony P900 features a music player, super high-res. display for media, and number pad that flips down to become a mini-keyboard.

Stock Information

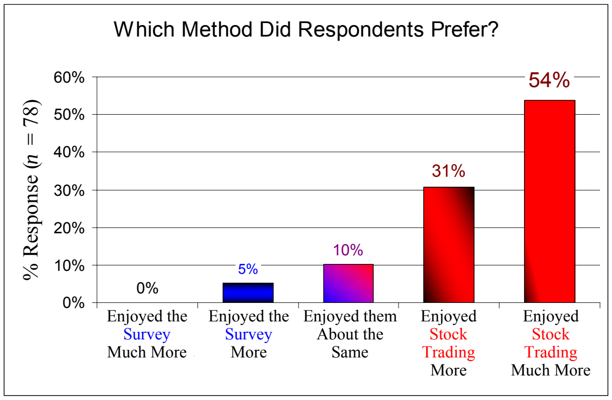
Last price	$15.00
Last Volume	40

Open orders

Buy orders					Sell orders				
11.00 100	13.00 20	13.25 800	13.50 100	14.00 100	15.00 364	18.00 70	18.00 70	20.00 249	20.00 90

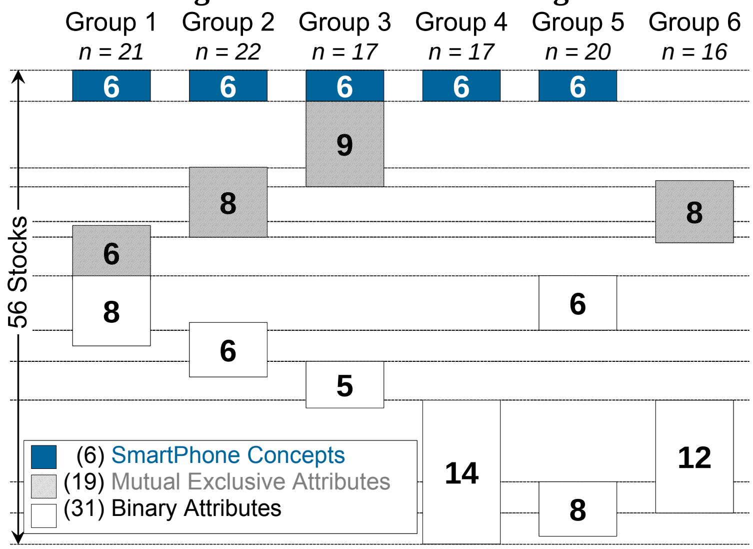
One potential advantage of scalable preference markets over traditional market research is a reduction in respondent and recruiting costs due to: (a) smaller sample size needed due to multiple transactions per respondent and interactions between respondents, (b) higher attention from respondents because of task enjoyment, (c) lower recruiting costs due to the fun factor of playing a game in competition, and (d) higher response rates for the same reason. We have some survey data in support of this conclusion.

Figure 5: Surveys versus Stock Trading Game



In our research, we compare volume-weighted average stock prices against individual survey results for 113 traders / survey respondents. The traders are organized into six subgroups of 16 to 21 traders each, with each trader group trading 20 of the 56 possible stocks, as illustrated in Figure 6. Note that some stocks are common to multiple trading groups.

Figure 6: Stock Market Design

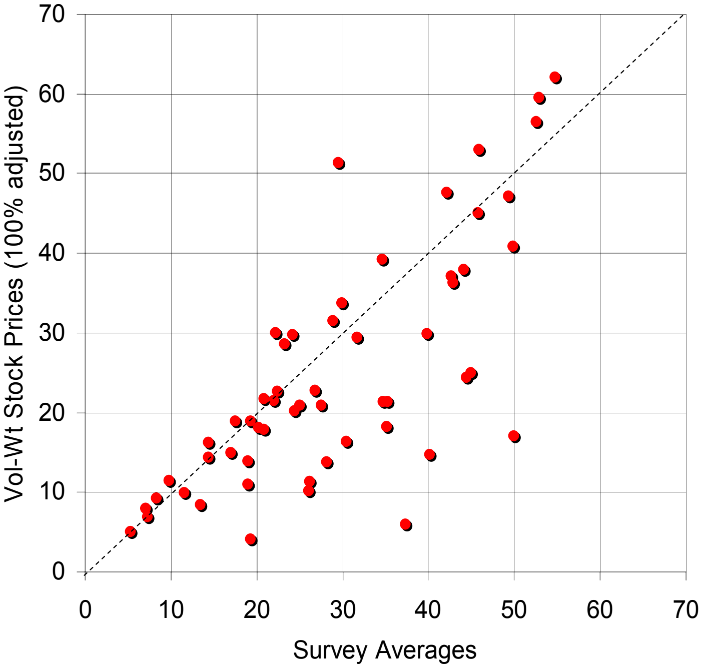


This experimental design allows all traders to directly or indirectly trade with each other, while only having to focus on a reasonably sized (size 20) subset of stocks.

Results of the trading game and survey are shown in Figure 7

Figure 7: Stock Trading Game Results

Stock Prices versus Survey Results (r^2 = 0.56)



The results show that the stock trading captured the key survey results reasonably well. Further, data was presented that these results were consistent with other prior studies done on the same product concepts and attributes, providing a reasonably high degree of external validity.

CONCLUSIONS

The present empirical research reveals that scalable preference markets, as developed in this paper, have the potential to lower respondent costs for many market research studies. The lower recruiting and respondent compensation costs derive from the desirability, speed, and efficiency afforded by scalable preference markets.

A well-attended stock market simulation quickly collects data multiple times per respondent (i.e. the same security can be traded by a single "trader" multiple times), and from many respondents simultaneously, typically in under an hour of trading. Thus the stock game method of data collection reduces the sample size of people required per question, since the number of data points typically exceeds the number of respondents, and the quality of the data is maintained through the trading mechanism (i.e. spurious noise is filtered out by the fact that widely divergent bid- and ask prices in the order book do not result in executed trades).

Future research will focus on the accuracy and differences between question types in order to provide suggestions for the optimal market design, and metrics for market liquidity (participants, number of stocks in subgroups, etc.). Additionally, the ideal market design and respondent incentives for participation can be analyzed as a tradeoff between cost and accuracy. Finally, it remains to be seen how to best recruit potential participants, and the relative importance of insight about the product category being studied versus expertise in stock trading itself, as far as the impact on gathering accurate preference information.

REFERENCES

Dahan, E., A. Lo, A., Poggio, T., Chan N. T., & Kim, A. W. (2006). Securities Trading of Concepts (STOC). Working Paper.

Dahan, E. & Hauser, J. R. (2002). The Virtual Customer. Journal of Product Innovation Management, 19 (5), 332-353.

Dubofsky, D. A. (1991). Volatility Increases Subsequent to NYSE and AMEX Stock Splits. Journal of Finance, 46 (1), 421-431.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance, 25, 383-417.

Fama, E. F. (1991). Efficient Capital Markets: II. Journal of Finance, 46 (5), 1575-1617.

Forsythe, R., Rietz, T. A. & Ross, T. W. (1999). Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets. Journal of Economic Behavior & Organization, 39, 83-110.

Hayek, F. A. v. (1945). The Use of Knowledge in Society. American Economic Review, 35 (4), 519-530.

Ohlson, J. A. & Penman, S. H. (1985). Volatility Increases Subsequent to Stock Splits - An Empirical Aberration. Journal of Financial Economics, 14 (2), 251-266.

Oliven, K. and Rietz, T. A. (2004). Suckers Are Born but Markets Are Made: Individual Rationality, Arbitrage, and Market Efficiency on an Electronic Futures Market. Management Science, 50 (3), 336-351.

Roll, R. (1984). Orange Juice and Weather. American Economic Review, 74 (5), 861-880.

Sawhney, M., Verona, G. and Prandelli, E. (2005). Collaborating to Create: The Internet as a Platform for Customer Engagement in Product Innovation. Journal of Interactive Marketing, 19 (4), 4-17.

Schwert, W. G. (1989). Why Does Stock Market Volatility Change Over Time? Journal of Finance, 44 (5), 1115-1153.

Shugan, S. M. (1980). The Cost of Thinking. Journal of Consumer Research, 7 (2), 99-111.

Smith, V. L. (1982). Microeconomic Systems as an Experimental Science. American Economic Review, 72 (5), 923-955.

Spann, M. & Skiera, B. (2003). Internet-Based Virtual Stock Markets for Business Forecasting. Management Science, 49 (10), 1310-1326.

Tversky, A. and Kahneman, D. (1974). Judgment and Uncertainty: Heuristics and Biases. Science, 185 (4157), 1124-1131.

Wind, J., Green, P. E., Shifflet, D. & Scarbrough, M. (1989). Courtyard by Marriott: Designing a Hotel Facility with Consumer-Based Marketing Models. Interfaces, 19 (1), 25-47.

JOHN FIEDLER'S MEASURE OF CLUSTER INTEGRITY: SUMMARY AND COMMENT

RICH JOHNSON,
SAWTOOTH SOFTWARE

At the 2006 Sawtooth Software Conference, John Fiedler proposed a new measure of "goodness of clustering." Unexpected family responsibilities made it impossible for him to provide a paper in time for these Proceedings. I have prepared this note because his presentation was valuable for understanding cluster analysis and also because it improved upon some earlier work of my own. John is a skillful and entertaining presenter, and his slides are provided as an appendix to this comment.

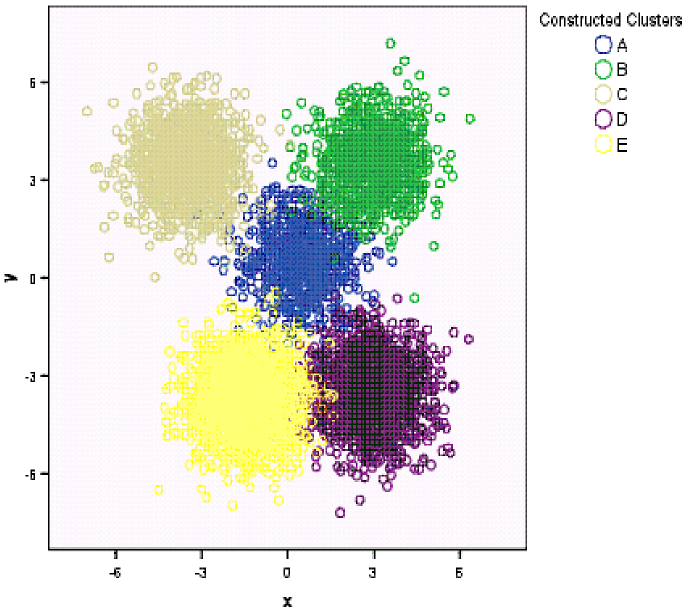
BACKGROUND

Cluster Analysis is often used to define market segments. For example, if survey respondents answer questions about the desirability of several product features, cluster analysis might be used to divide respondents into groups having similar desires.

Among the most widely used clustering techniques are the "k means" methods. The researcher specifies **k**, the desired number of groups. These methods start by dividing objects (for example, survey respondents), into **k** groups, perhaps randomly, and then proceeding through many iterations in which each object is reclassified into the newly modified group to which it is most similar. Similarity between an object and a group is often measured as a function of the sum of squared differences between the object's values and the group's means for the variables of interest. The process stops when each object is most similar to the group of which it is a member.

These methods always produce groups, whether or not those groups conform in any way to our intuitive notions of "clusters." For example, imagine a single variable (Fiedler uses IQ as an example) on which objects are distributed normally. If such a clustering method is asked to produce two clusters, it will divide objects into two groups: those with values above the mean and those with values below the mean. Such groups may be useful for many purposes, but they do not conform to what most of us have in mind when we speak of "clusters."

Suppose there are two variables rather than one, so that objects' values can be used to plot the location of each object in a two-dimensional space. Fiedler gives such an example, using five artificially-constructed clusters, reproduced below:



What are the characteristics of this distribution of objects (apart from different colors or shades of grey) which suggest the presence of clusters? My early conjecture (Johnson, 1972), was:

> …in a truly satisfying clustering there should be relatively few points occupying the region between clusters, and relatively many points lying near the respective cluster means.

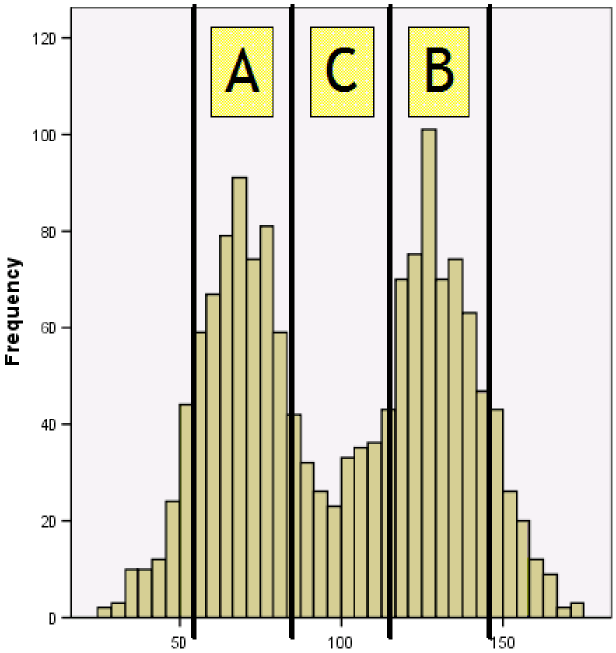
In that attempt to provide an operational definition of "goodness of clustering," I counted the number of objects within a particular distance of each cluster's mean, and also the number of objects within the same distance of the midpoint between each pair of clusters. It seemed that if points were denser within clusters than between clusters, that might provide evidence that the clusters conformed to common intuition of the way clusters ought to be.

Unfortunately, this idea didn't work out very well in practice. In most real applications there are many more than two dimensions. In the two-dimensional example above, the total area shown in the diagram is about 18^2 and the area within one unit of any point is roughly one percent of the total area. But with n variables scaled similarly, the total volume of the space would be about 18^n, a very large number, so the percentage of the volume within one unit of any point would be very small. When dealing with many variables it turned out that few objects were within a reasonable distance of any particular point, so all densities approached zero and it was not useful to compare them.

Fiedler's Idea

Building on this work, Fiedler has proposed a measure that seems much more useful. Rather than considering an n-dimensional space, he evaluated each pair of clusters using only a single dimension, consisting of that linear combination of variables which distinguished most clearly between those two clusters, as provided by a two-group discriminant analysis. No matter how many variables are used in clustering, the separation of each pair of clusters is evaluated using a single dimension.

Suppose two groups, when scored on the single dimension best separating them, appear as in the diagram below (also from Fiedler's slides).



Consider three areas, A, B, and C, with equal width. Area A contains objects within some distance of one cluster's mean. Area B contains objects within the same distance of the other cluster's mean. Area C contains objects within the same distance of the midpoint between clusters. For the three areas to be adjacent, the common limiting distance must be equal to ¼ of the distance between means.

Quoting Fiedler,

> Cluster Integrity is concerned with the density in the mid-area "C" compared to what we would predict solely on the basis of the [densities] of the two adjacent cluster centers, "A" and "B."

Fiedler's measure of Cluster Integrity (CI) for these two clusters is 1 – c / sqrt(a*b) where a, b, and c are counts of objects within the respective regions. CI of zero would mean that the middle area is as dense as the (geometric) mean of the two cluster areas. A negative value would mean that the middle area is more dense, as is typically the case with a single-peaked distribution. A positive value of CI would mean that the middle area is less dense than the geometric average of the two cluster areas, suggesting that the groups are relatively distinct. In this example the counts of objects in each region are

a = 508, b = 507, and c = 237, for CI = .533.

If there are more than two clusters, then an overall CI score can be obtained by averaging the pairwise CI scores. For the artificial five-group data set shown above, average CI was .74, indicating quite strong separation among clusters. However, the CI measure was not completely successful for this data set. Fiedler examined solutions with 2 through 8 clusters, finding that the 5 and 8-cluster solutions were tied, with the 6 and 7-cluster solutions also having CI values nearly as high.

He also considered two other measures often used to evaluate cluster solutions. One is "Reproducibility," the percentage of objects classified identically in repeated clusterings from random starting points. Another is "Discrimination," the average F ratio when each variable is used to compare variances between groups with variances within groups. Although it is recognized that F ratios obtained in this way cannot be used to test significance of differences between groups, they do provide a convenient metric to measure differences between groups, relative to the within-group variability. For this data set the 4 and 5-cluster solutions had highest Reproducibility, and the 5-cluster solution had uniquely high Discrimination, although there was little difference between the 4 through 8-cluster solutions.

FIEDLER'S ANALYSES

Fiedler also analyzed several data sets more typical of those found in practice.

Artificial Data Sets: He used four artificial data sets, each with 25 variables and 5,000 objects. Each data set contained four groups, of sizes 500, 1,000, 1,500, and 2,000 objects. Each group had a different vector of means with groups differing from one another by various amounts, ranging from no separation to separation so clear that no clustering procedure should be misled. The within-group variances and covariances were identical to those found in a data set of similar size from real respondents.

For the data set with clearest separation among clusters, the 2 and 4-cluster solutions were tied with greatest CI. For the data set with slightly less separation among clusters, the 4 cluster solution was identified correctly as having greatest CI. For the other two data sets all CI measures indicated less structure in the data, and no clear determination of which was the best solution.

For the Discrimination measure, there was a clear difference among data sets, with discrimination higher for data sets with more separation among groups, but in none of them was there a clear indication of which was the best number of clusters.

For the Reproducibility measure, the 4-cluster solution was correctly identified as best for the data set with greatest separation, but not in the other data sets. Disturbingly, Reproducibility was high for some solutions even when there were no true differences among clusters.

From these analyses Fiedler concluded tentatively that CI is somewhat more useful than Reproducibility in data sets with well-defined clusters, and that it reveals the amount of structure in the data in a way similar to Discrimination. However, he also noted that none of the measures consistently identifies the correct solution when there is only a moderate amount of structure in the data.

Three Commercial Data Sets: Fiedler also examined three large data sets that had been used commercially:

> **Study A:** This was an Internet-based survey with 46 *very lengthy* (emphasis Fiedler's) attitude statements and 1600 individuals responding on a 10-point agreement scale.

Study B: Also an Internet-based survey, with 37 attitude statements, and 3750 individuals responding on a 5-point fully-anchored agreement scale.

Study C: General Social Survey consisting of in-home personal interviews by NORC of more than 8,000 individuals, with 94 demographic and attitudinal variables, using standardized scaling of variables.

Study A had lowest CI and least Discrimination for nearly all numbers of clusters, suggesting that Study A may have had data of inferior quality. Disturbingly, Study A had high Reproducibility for several numbers of clusters, suggesting that Reproducibility is an undependable way of choosing the best solution.

Discrimination showed clear differences among data sets, but for no data set was there a clear indication of a "best" number of clusters. This suggests that discrimination may be useful for assessing the quality of a data set in general, but of less use in choosing the best solution.

For all three studies there appeared to be specific numbers of clusters with especially high CI values. Also, Study A had lower CI for nearly all numbers of clusters than studies B or C. This may confirm that CI can not only suggest the "best" number of clusters, but also reveal the relative amount of structure in the data set.

FIEDLER'S CONCLUSIONS

Although we do not know the "correct" cluster solutions for the commercial data sets, we can inquire whether the various measures of clustering success tell similar or different stories. In these analyses they gave conflicting answers. If each is useful in some way, then using all of them together is likely to produce a better result than any separately. Fiedler concluded that CI provides valuable insight beyond the other two measures.

Having compared several real commercial data sets to several artificial data sets with known properties, Fiedler was also able to generalize about the cluster structure in data sets encountered in the real world. He examined 55 different clusterings, finding that CI and Reproducibility were uncorrelated. By classifying as "excellent" the 22 clusterings with highest Discrimination and CI, he found that 20 of them were from artificial data sets and only two were from real data sets. Expanding the number of solutions to include an additional 14 "good" clusterings, 26 were from artificial data sets and only ten were from real data sets. Examining the remaining clusterings, which he classified as "poor," he found that many analyses using commercial survey data have CI comparable to random data, and some of those have near perfect Reproducibility. These findings suggest that real data sets may often have less clear cluster structure than had been supposed, based on the measures previously available.

INTERPRETATION

I believe that Fiedler's work provides a valuable new way to assess the results of cluster analysis. Since cluster analysis *always* produces clusters, irrespective of the amount of structure present in the data, the mere existence of a clustering solution tells us nothing about its quality.

It may be useful to divide even a unimodal continuous distribution into segments, such as heavy vs. light users of a product, or satisfied vs. unsatisfied customers. Dividing such a distribution into two groups will always produce "high" and "low" groups, so Reproducibility should be excellent. But since there are more objects near the midpoint between clusters than near the cluster means, CI should be negative. This was seen in many of the solutions Fiedler classified as "poor."

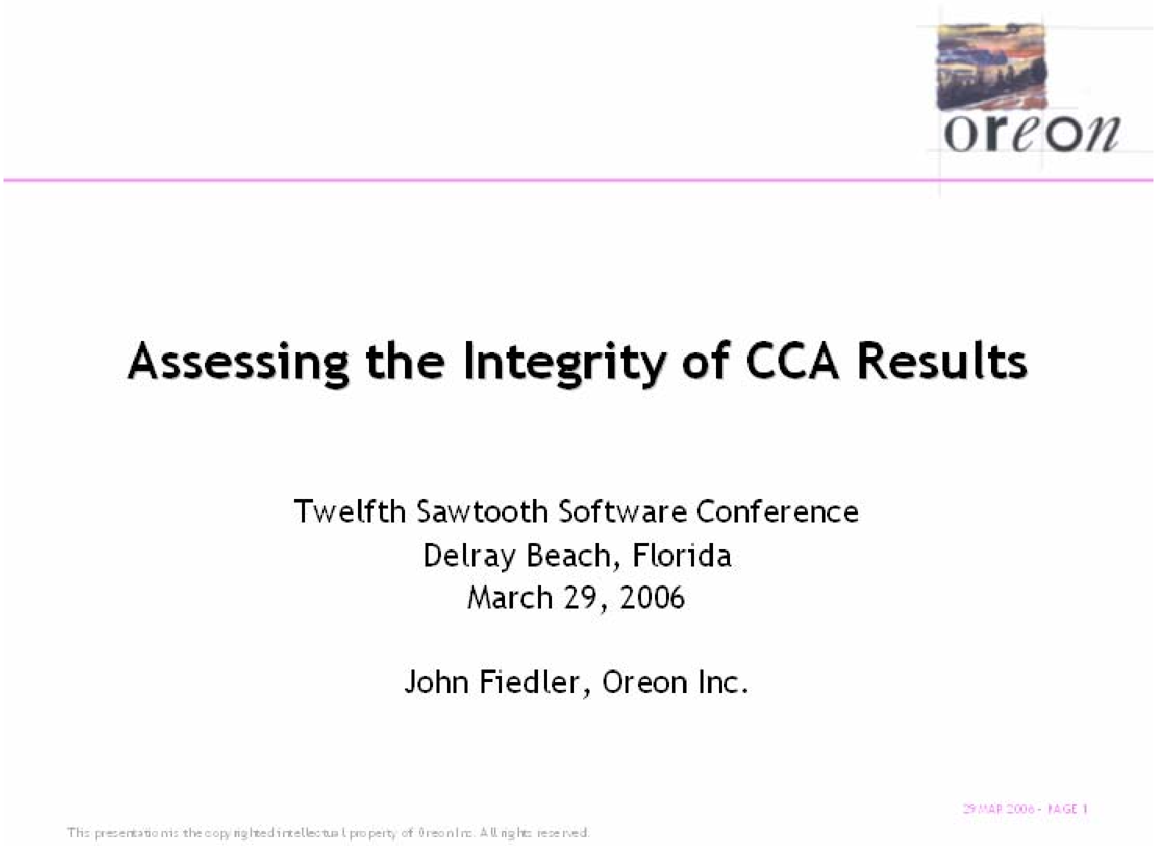
Many applications of market segmentation depend for their usefulness on the assumption that there are more individuals near the centers of their segments than near the boundaries between segments. For example, sometimes companies design products to appeal to specific segments. If a product is positioned to appeal to an individual near the average for a segment, this must be with the hope there are more individuals near that segment's center than elsewhere. But suppose we have divided a continuous single-peaked distribution into two segments. In that case the best place for a product is clearly at the boundary between segments rather than at the center of either segment. I fear this error has been made many times by market researchers who have used cluster analysis to define market segments.

There is another issue of importance to marketing researchers for which Fiedler's work also has strong implications. Imagine a distribution of individuals in a multidimensional space, according to their desires for various product features. What is the shape of this distribution? Is it typically a single-peaked distribution similar to a multivariate normal distribution? Or does it have a more "lumpy" shape, similar to a collection of multivariate normals? To be more graphic, is it shaped more like a watermelon, or like a bag full of cantaloupes? This is an important question because its answer should determine the methods we use in studying respondent heterogeneity. If segments are locally dense, with relatively empty space between them, then we should be using corresponding methods to obtain segments, such as latent class or cluster analysis. If the distribution is more nearly continuous, with density greatest near its center and diminishing as we move from the center in any direction, then cluster analysis is not an appropriate method. In that case it is probably best to attempt individual-level analysis using something like hierarchical Bayes with a normally distributed prior.

Fiedler has shown a promising direction for further research which can tell us more about the nature of the solutions produced by cluster analysis. Clusters should have high Cluster Integrity. It remains to be seen whether they do.

REFERENCES

Johnson, Richard M. "How Can You Tell if Things Are Really Clustered?" Working paper available at http://www.oreon.net/tech_papers/johnson_clustering_1972.pdf



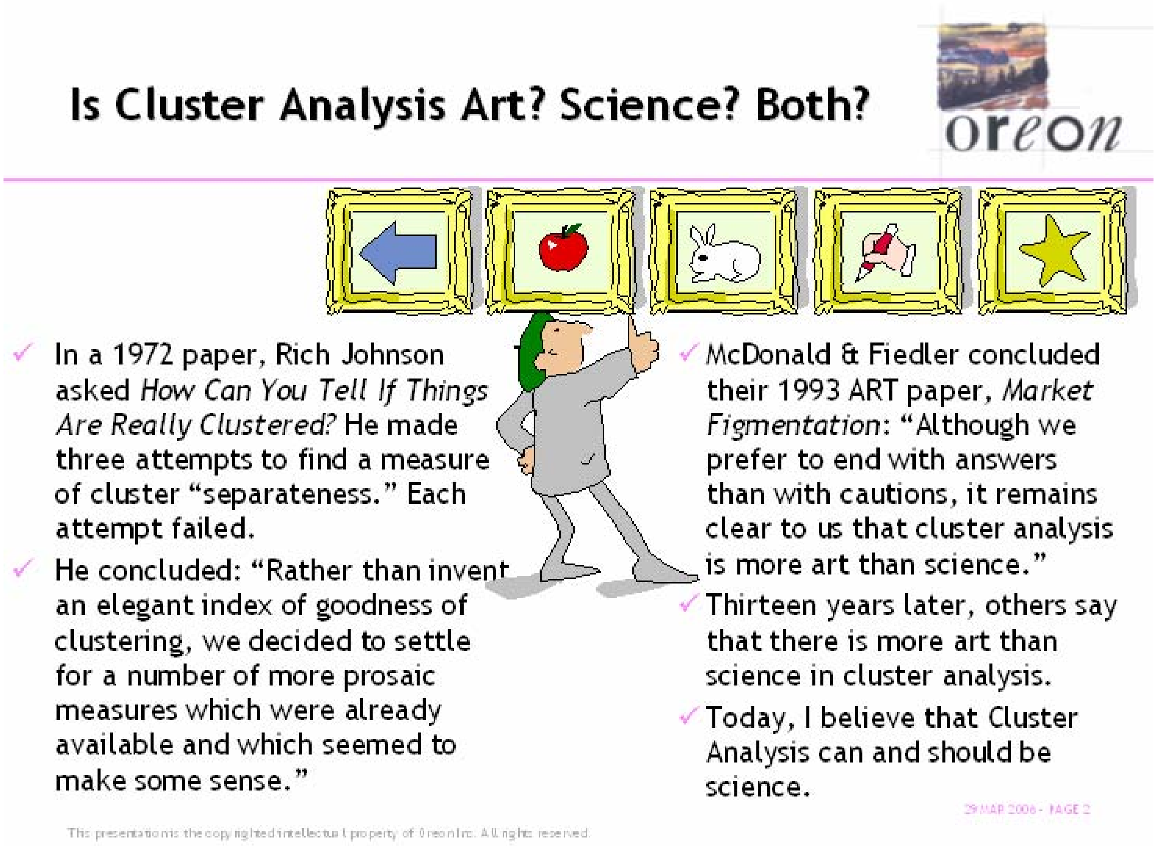
oreon

Assessing the Integrity of CCA Results

Twelfth Sawtooth Software Conference
Delray Beach, Florida
March 29, 2006

John Fiedler, Oreon Inc.

29 MAR 2006 - PAGE 1

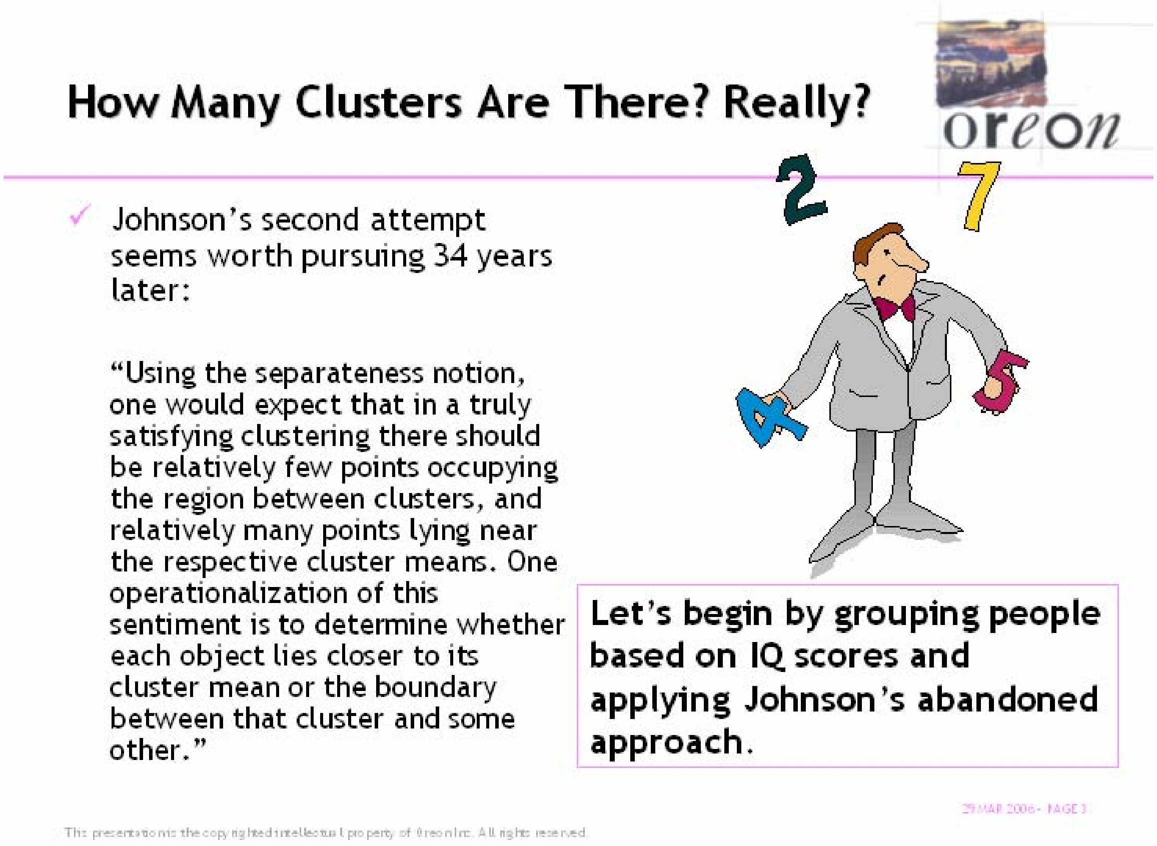


Is Cluster Analysis Art? Science? Both?

oreon

- In a 1972 paper, Rich Johnson asked *How Can You Tell If Things Are Really Clustered?* He made three attempts to find a measure of cluster "separateness." Each attempt failed.
- He concluded: "Rather than invent an elegant index of goodness of clustering, we decided to settle for a number of more prosaic measures which were already available and which seemed to make some sense."
- McDonald & Fiedler concluded their 1993 ART paper, *Market Figmentation*: "Although we prefer to end with answers than with cautions, it remains clear to us that cluster analysis is more art than science."
- Thirteen years later, others say that there is more art than science in cluster analysis.
- Today, I believe that Cluster Analysis can and should be science.

29 MAR 2006 - PAGE 2



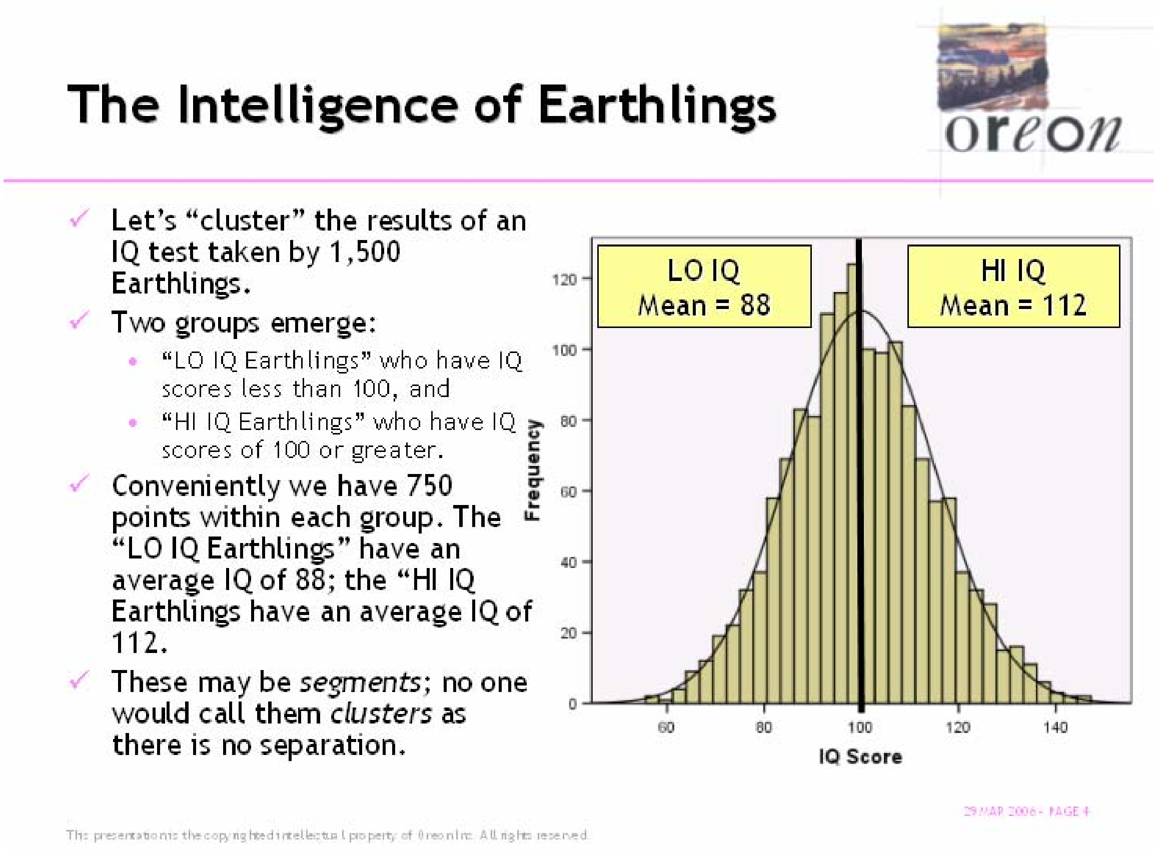
How Many Clusters Are There? Really?

- ✓ Johnson's second attempt seems worth pursuing 34 years later:

  "Using the separateness notion, one would expect that in a truly satisfying clustering there should be relatively few points occupying the region between clusters, and relatively many points lying near the respective cluster means. One operationalization of this sentiment is to determine whether each object lies closer to its cluster mean or the boundary between that cluster and some other."

Let's begin by grouping people based on IQ scores and applying Johnson's abandoned approach.

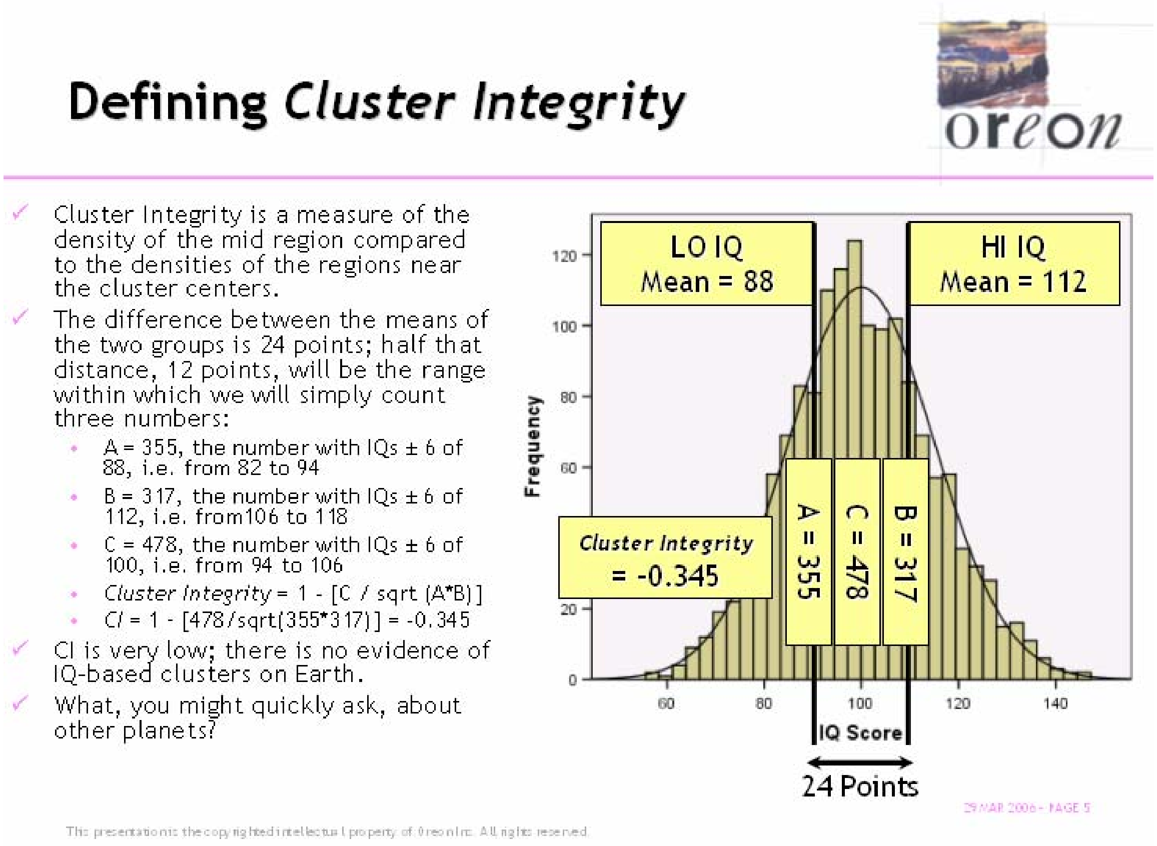
29 MAR 2006 - PAGE 3



The Intelligence of Earthlings

- ✓ Let's "cluster" the results of an IQ test taken by 1,500 Earthlings.
- ✓ Two groups emerge:
  - "LO IQ Earthlings" who have IQ scores less than 100, and
  - "HI IQ Earthlings" who have IQ scores of 100 or greater.
- ✓ Conveniently we have 750 points within each group. The "LO IQ Earthlings" have an average IQ of 88; the "HI IQ Earthlings have an average IQ of 112.
- ✓ These may be *segments*; no one would call them *clusters* as there is no separation.

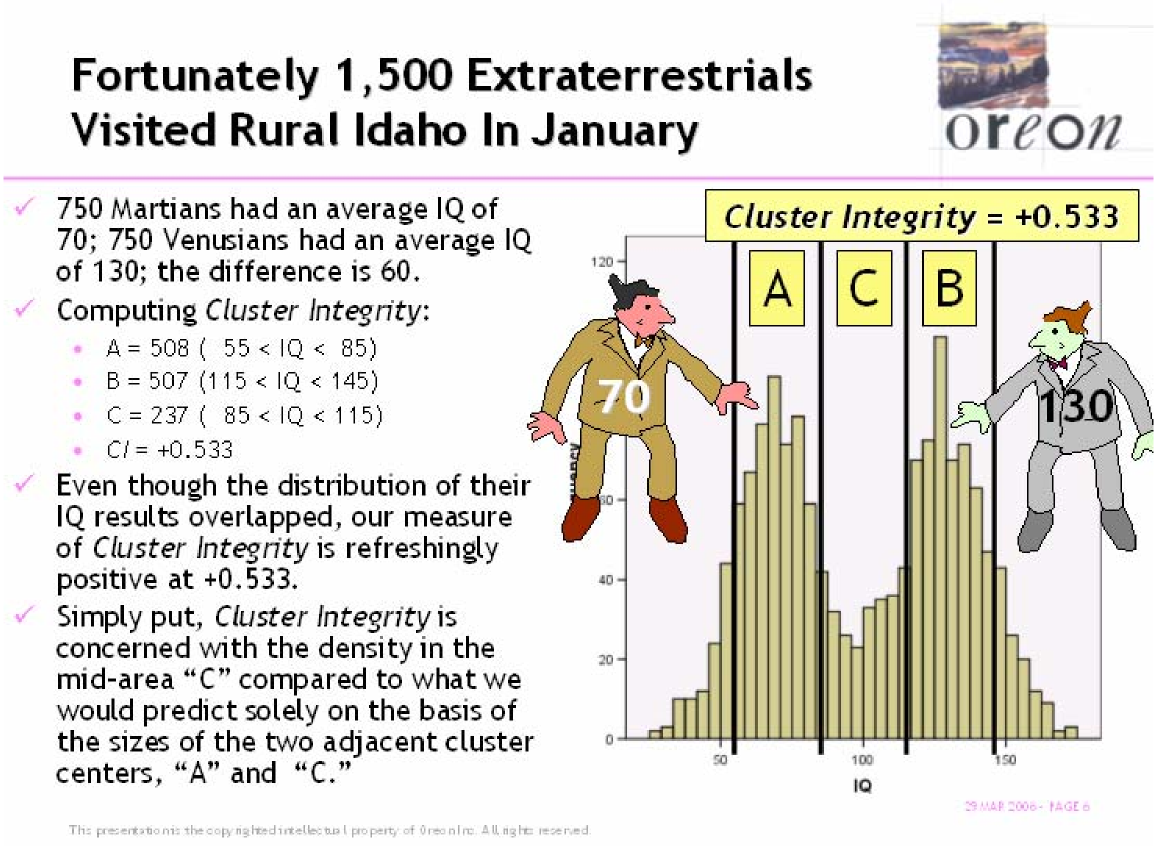
29 MAR 2006 - PAGE 4



Defining *Cluster Integrity*

- Cluster Integrity is a measure of the density of the mid region compared to the densities of the regions near the cluster centers.
- The difference between the means of the two groups is 24 points; half that distance, 12 points, will be the range within which we will simply count three numbers:
  - A = 355, the number with IQs ± 6 of 88, i.e. from 82 to 94
  - B = 317, the number with IQs ± 6 of 112, i.e. from106 to 118
  - C = 478, the number with IQs ± 6 of 100, i.e. from 94 to 106
  - *Cluster Integrity* = 1 - [C / sqrt (A*B)]
  - *CI* = 1 - [478/sqrt(355*317)] = -0.345
- CI is very low; there is no evidence of IQ-based clusters on Earth.
- What, you might quickly ask, about other planets?

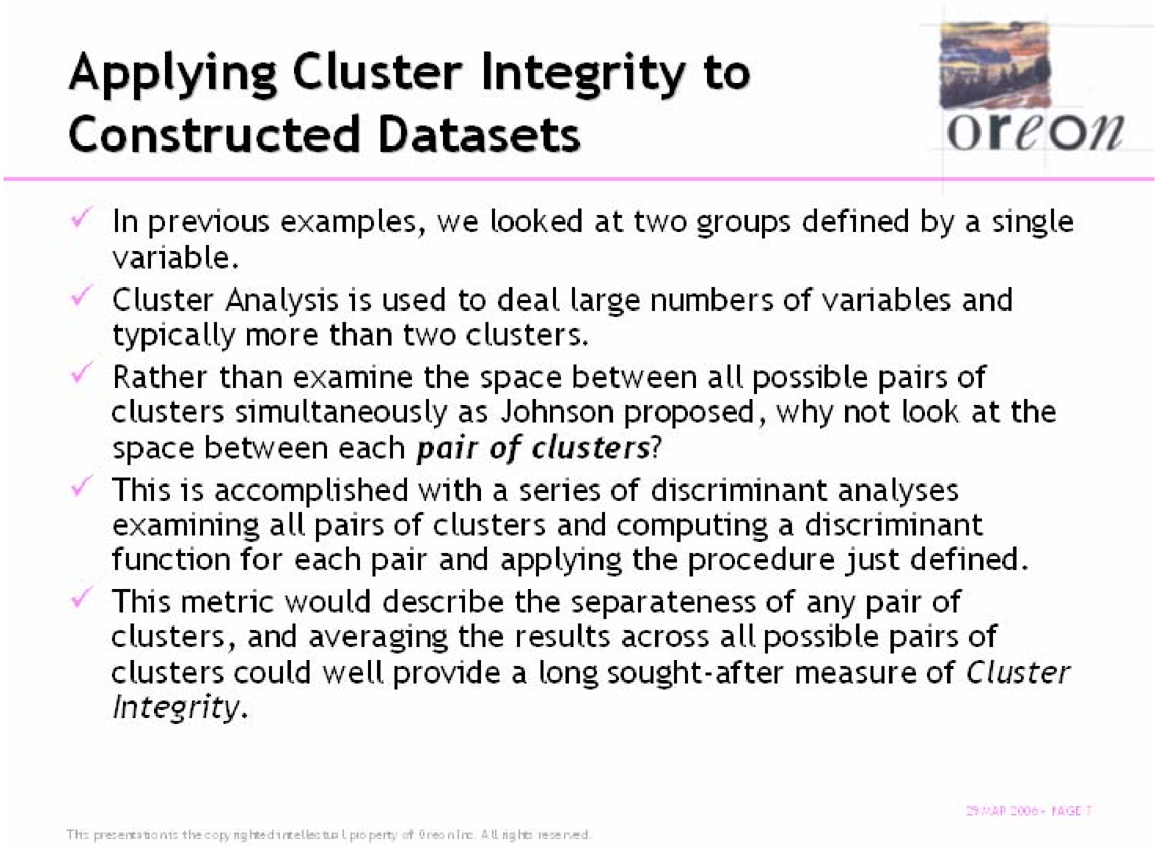
29 MAR 2006 - PAGE 5



Fortunately 1,500 Extraterrestrials Visited Rural Idaho In January

- 750 Martians had an average IQ of 70; 750 Venusians had an average IQ of 130; the difference is 60.
- Computing *Cluster Integrity*:
  - A = 508 (55 < IQ < 85)
  - B = 507 (115 < IQ < 145)
  - C = 237 (85 < IQ < 115)
  - *CI* = +0.533
- Even though the distribution of their IQ results overlapped, our measure of *Cluster Integrity* is refreshingly positive at +0.533.
- Simply put, *Cluster Integrity* is concerned with the density in the mid-area "C" compared to what we would predict solely on the basis of the sizes of the two adjacent cluster centers, "A" and "C."

29 MAR 2006 - PAGE 6

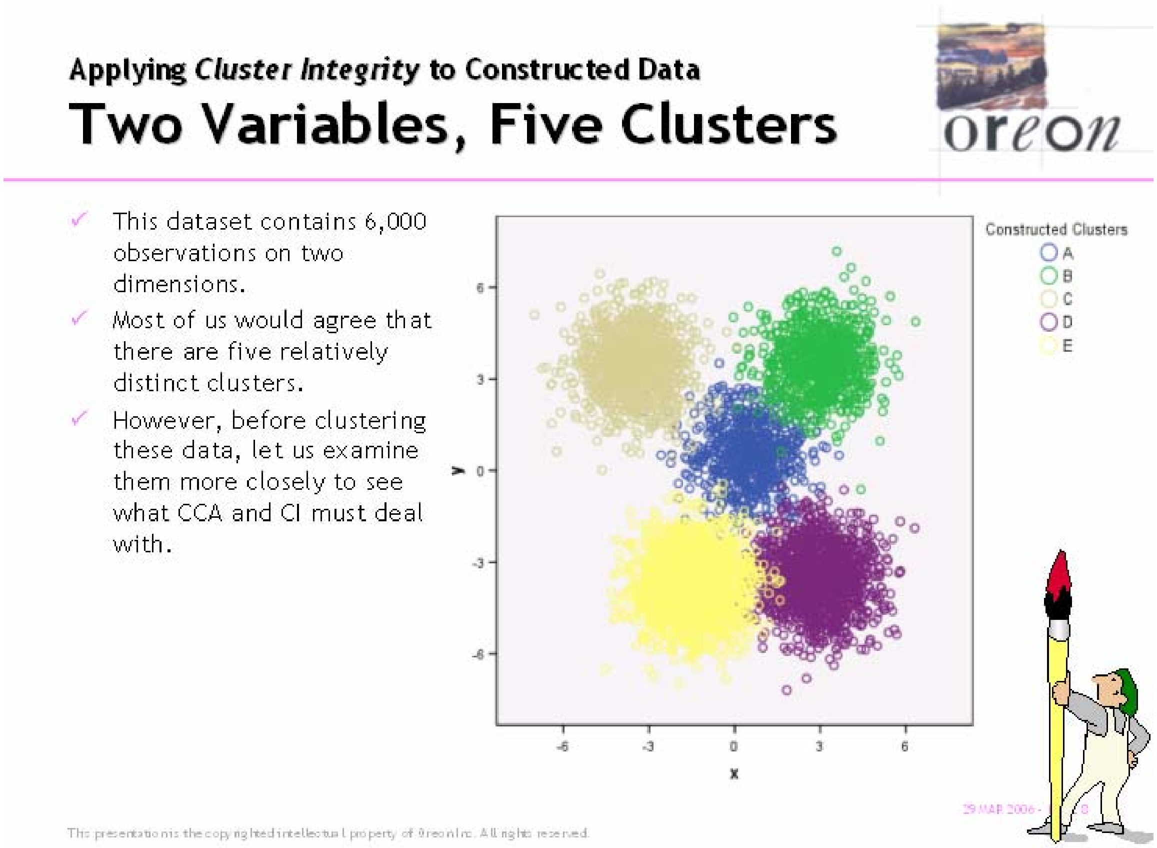


Applying Cluster Integrity to Constructed Datasets

- ✓ In previous examples, we looked at two groups defined by a single variable.
- ✓ Cluster Analysis is used to deal large numbers of variables and typically more than two clusters.
- ✓ Rather than examine the space between all possible pairs of clusters simultaneously as Johnson proposed, why not look at the space between each ***pair of clusters***?
- ✓ This is accomplished with a series of discriminant analyses examining all pairs of clusters and computing a discriminant function for each pair and applying the procedure just defined.
- ✓ This metric would describe the separateness of any pair of clusters, and averaging the results across all possible pairs of clusters could well provide a long sought-after measure of *Cluster Integrity*.

29 MAR 2006 - PAGE 7

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



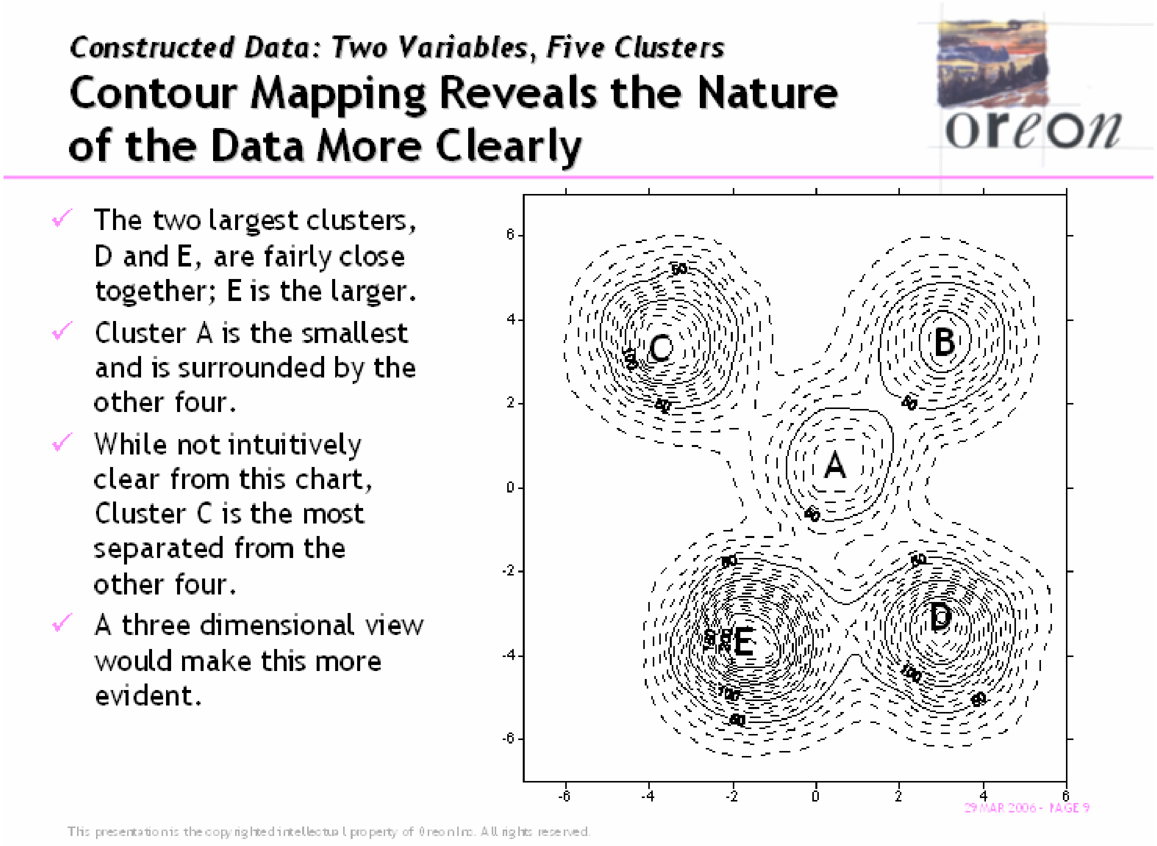
Applying *Cluster Integrity* to Constructed Data

Two Variables, Five Clusters

- ✓ This dataset contains 6,000 observations on two dimensions.
- ✓ Most of us would agree that there are five relatively distinct clusters.
- ✓ However, before clustering these data, let us examine them more closely to see what CCA and CI must deal with.

29 MAR 2006 - 8

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



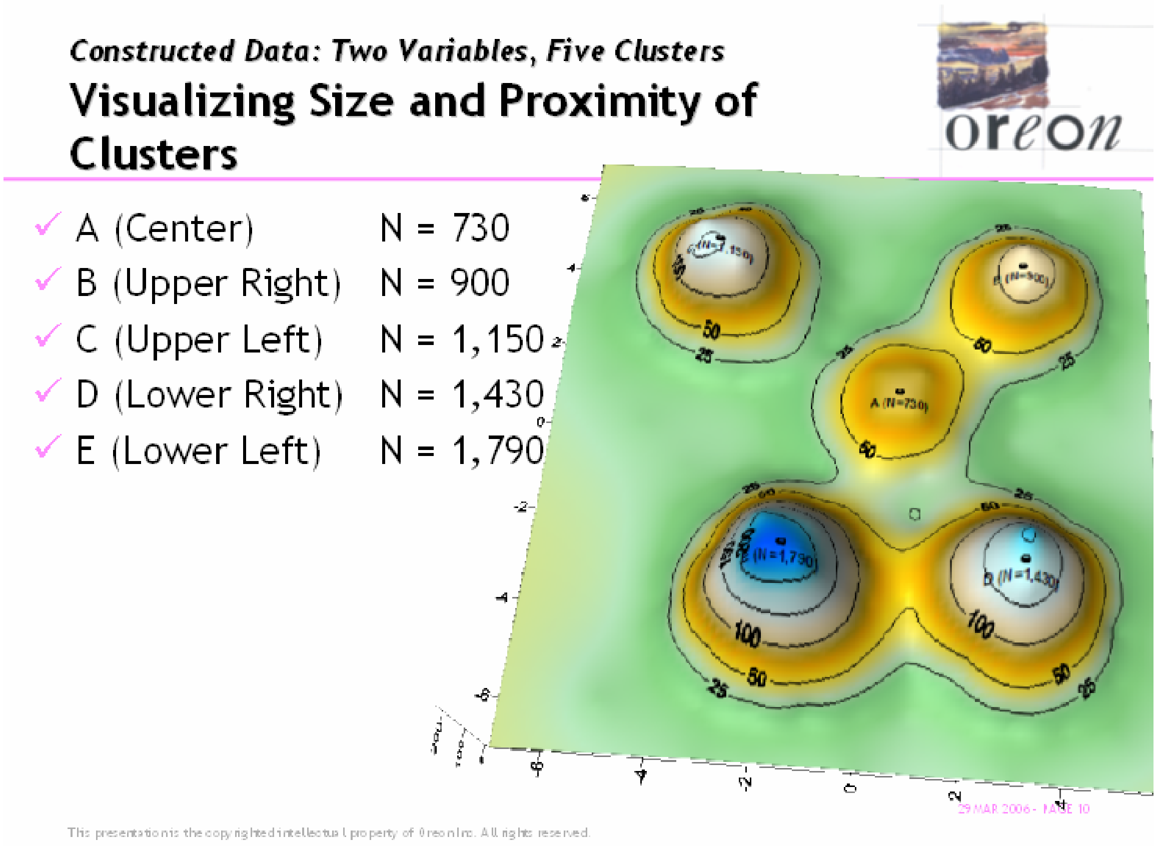
Constructed Data: Two Variables, Five Clusters

Contour Mapping Reveals the Nature of the Data More Clearly

- ✓ The two largest clusters, D and E, are fairly close together; E is the larger.
- ✓ Cluster A is the smallest and is surrounded by the other four.
- ✓ While not intuitively clear from this chart, Cluster C is the most separated from the other four.
- ✓ A three dimensional view would make this more evident.

29 MAR 2006 - PAGE 9

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



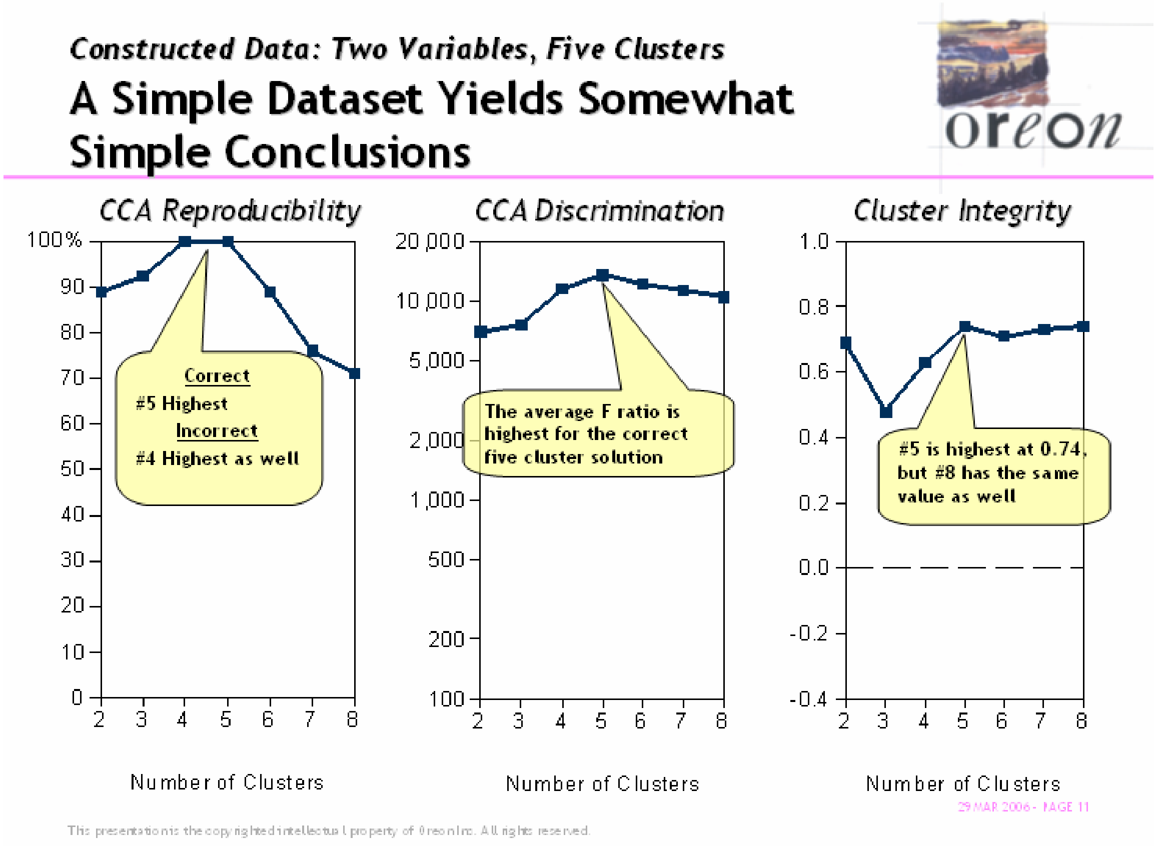
Constructed Data: Two Variables, Five Clusters

Visualizing Size and Proximity of Clusters

✓ A (Center)	N = 730
✓ B (Upper Right)	N = 900
✓ C (Upper Left)	N = 1,150
✓ D (Lower Right)	N = 1,430
✓ E (Lower Left)	N = 1,790

29 MAR 2006 - PAGE 10

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



Constructed Data: Two Variables, Five Clusters

A Simple Dataset Yields Somewhat Simple Conclusions

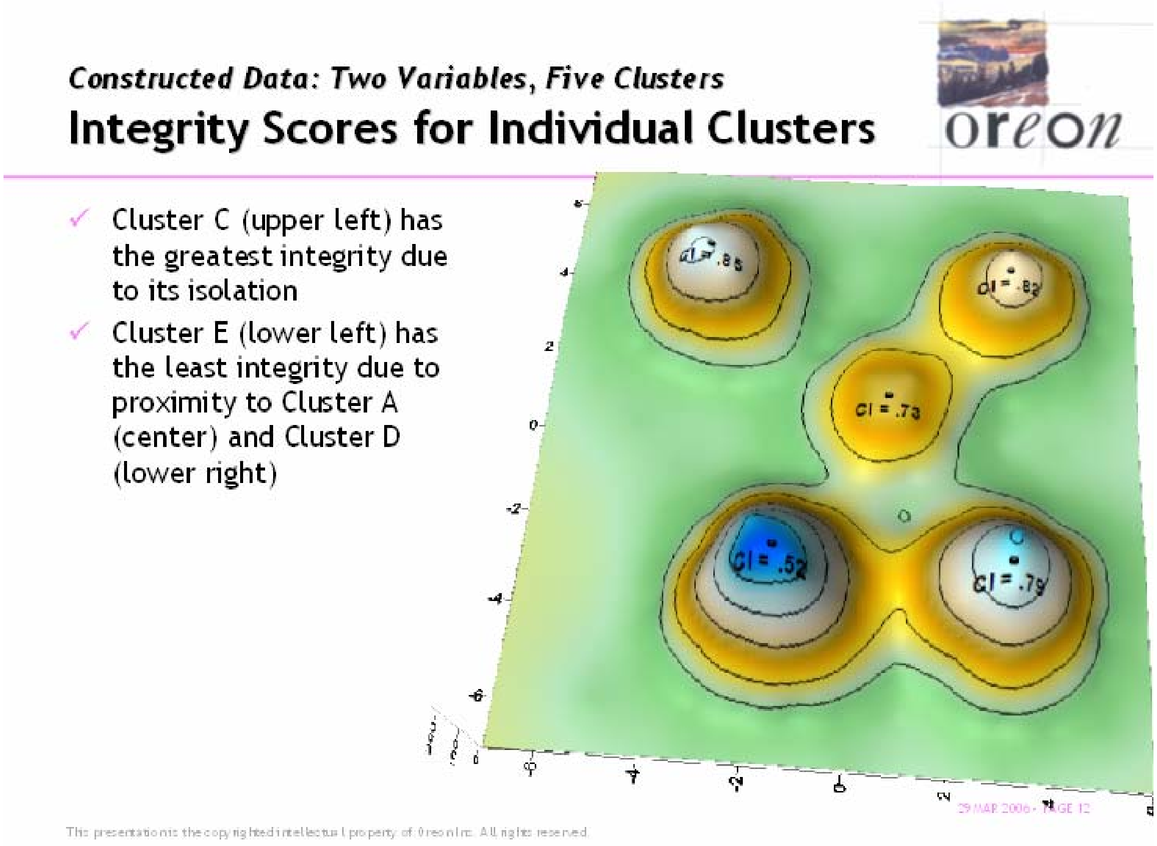
Correct
#5 Highest
Incorrect
#4 Highest as well

The average F ratio is highest for the correct five cluster solution

#5 is highest at 0.74, but #8 has the same value as well

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

29 MAR 2006 - PAGE 11



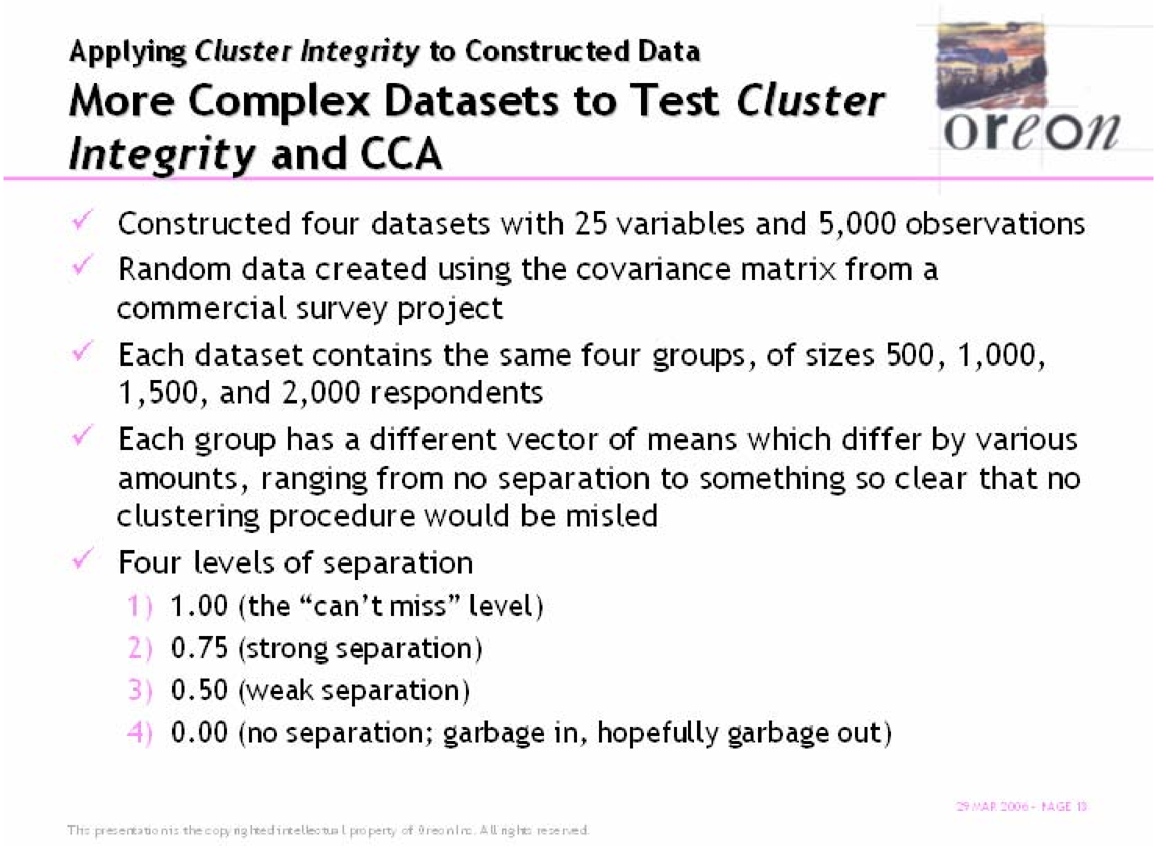
Constructed Data: Two Variables, Five Clusters

Integrity Scores for Individual Clusters

- ✓ Cluster C (upper left) has the greatest integrity due to its isolation
- ✓ Cluster E (lower left) has the least integrity due to proximity to Cluster A (center) and Cluster D (lower right)

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

29 MAR 2006 - PAGE 12

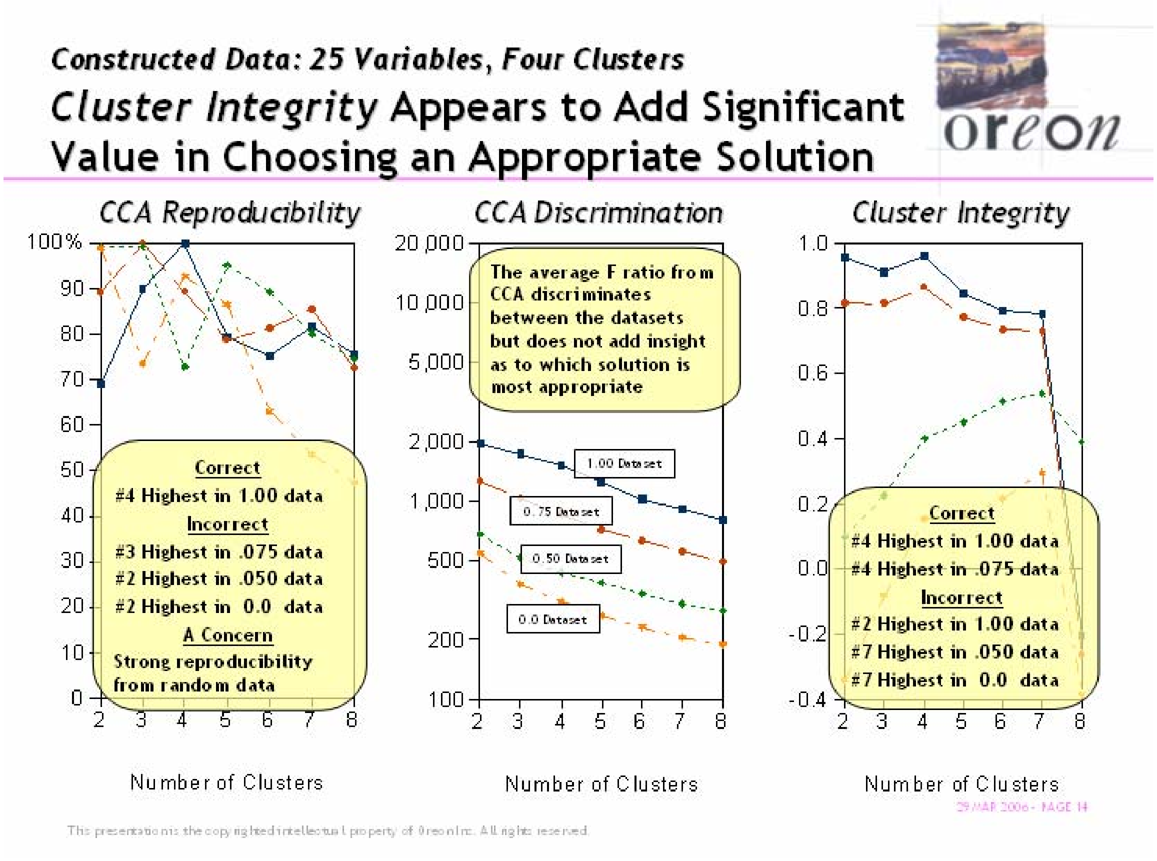


Applying *Cluster Integrity* to Constructed Data

More Complex Datasets to Test *Cluster Integrity* and CCA

- ✓ Constructed four datasets with 25 variables and 5,000 observations
- ✓ Random data created using the covariance matrix from a commercial survey project
- ✓ Each dataset contains the same four groups, of sizes 500, 1,000, 1,500, and 2,000 respondents
- ✓ Each group has a different vector of means which differ by various amounts, ranging from no separation to something so clear that no clustering procedure would be misled
- ✓ Four levels of separation
  1) 1.00 (the "can't miss" level)
  2) 0.75 (strong separation)
  3) 0.50 (weak separation)
  4) 0.00 (no separation; garbage in, hopefully garbage out)

29 MAR 2006 - PAGE 13



Constructed Data: 25 Variables, Four Clusters

Cluster Integrity Appears to Add Significant Value in Choosing an Appropriate Solution

Correct

#4 Highest in 1.00 data

Incorrect

#3 Highest in .075 data
#2 Highest in .050 data
#2 Highest in 0.0 data

A Concern

Strong reproducibility from random data

The average F ratio from CCA discriminates between the datasets but does not add insight as to which solution is most appropriate

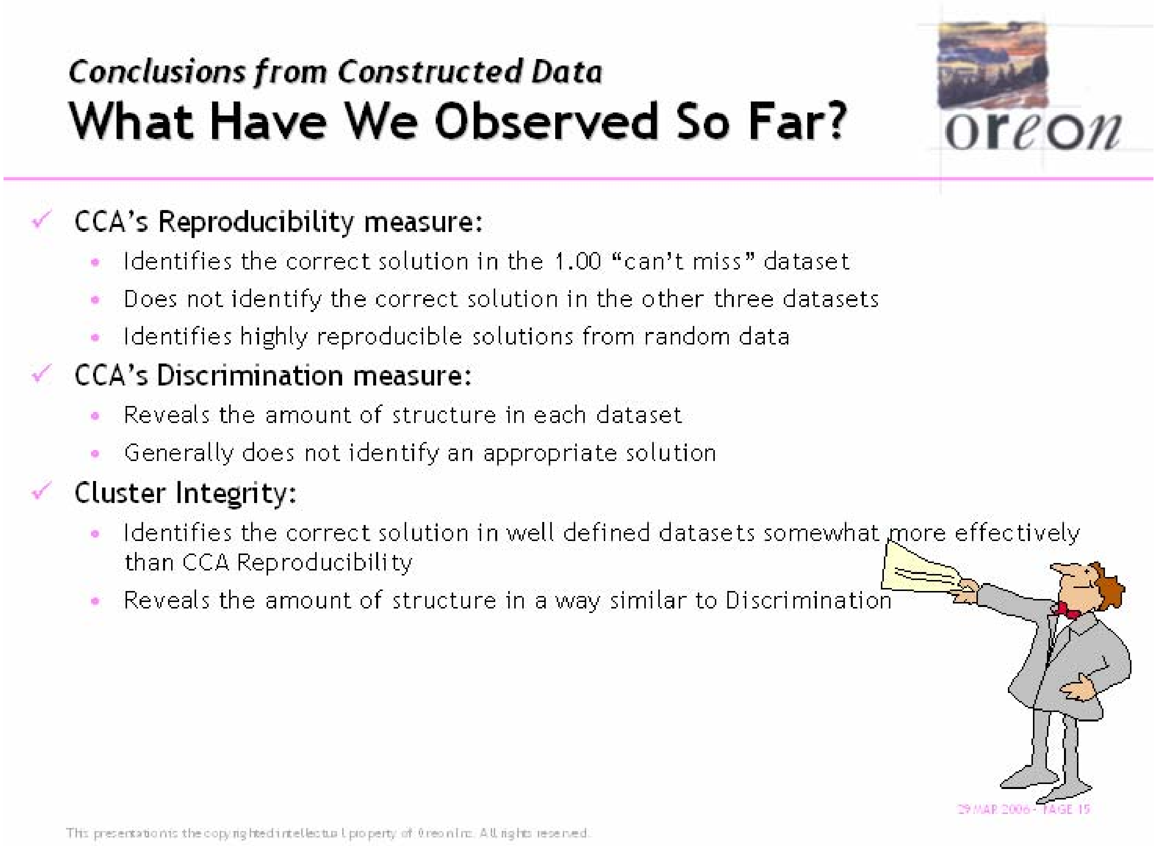
Correct

#4 Highest in 1.00 data
#4 Highest in .075 data

Incorrect

#2 Highest in 1.00 data
#7 Highest in .050 data
#7 Highest in 0.0 data

29 MAR 2006 - PAGE 14



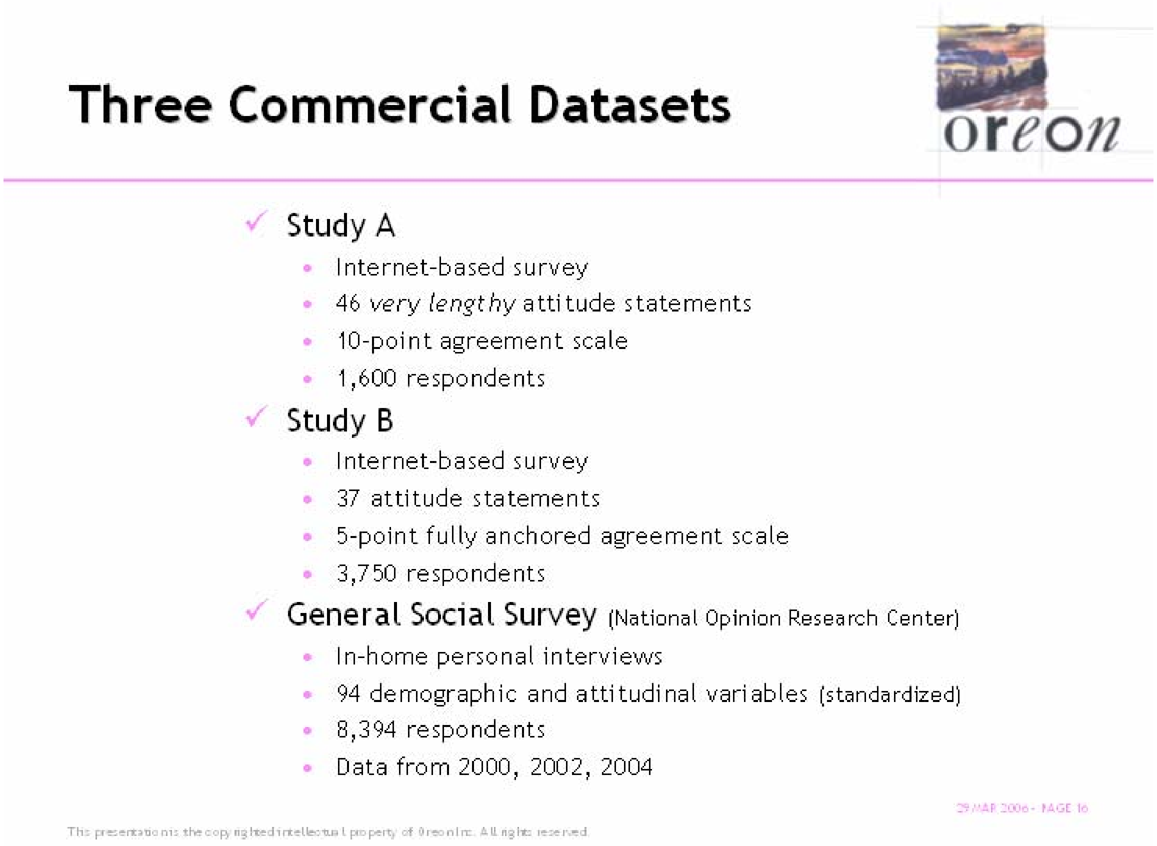
Conclusions from Constructed Data

What Have We Observed So Far?

- ✓ CCA's Reproducibility measure:
  - Identifies the correct solution in the 1.00 "can't miss" dataset
  - Does not identify the correct solution in the other three datasets
  - Identifies highly reproducible solutions from random data
- ✓ CCA's Discrimination measure:
  - Reveals the amount of structure in each dataset
  - Generally does not identify an appropriate solution
- ✓ Cluster Integrity:
  - Identifies the correct solution in well defined datasets somewhat more effectively than CCA Reproducibility
  - Reveals the amount of structure in a way similar to Discrimination

29 MAR 2006 - PAGE 15

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

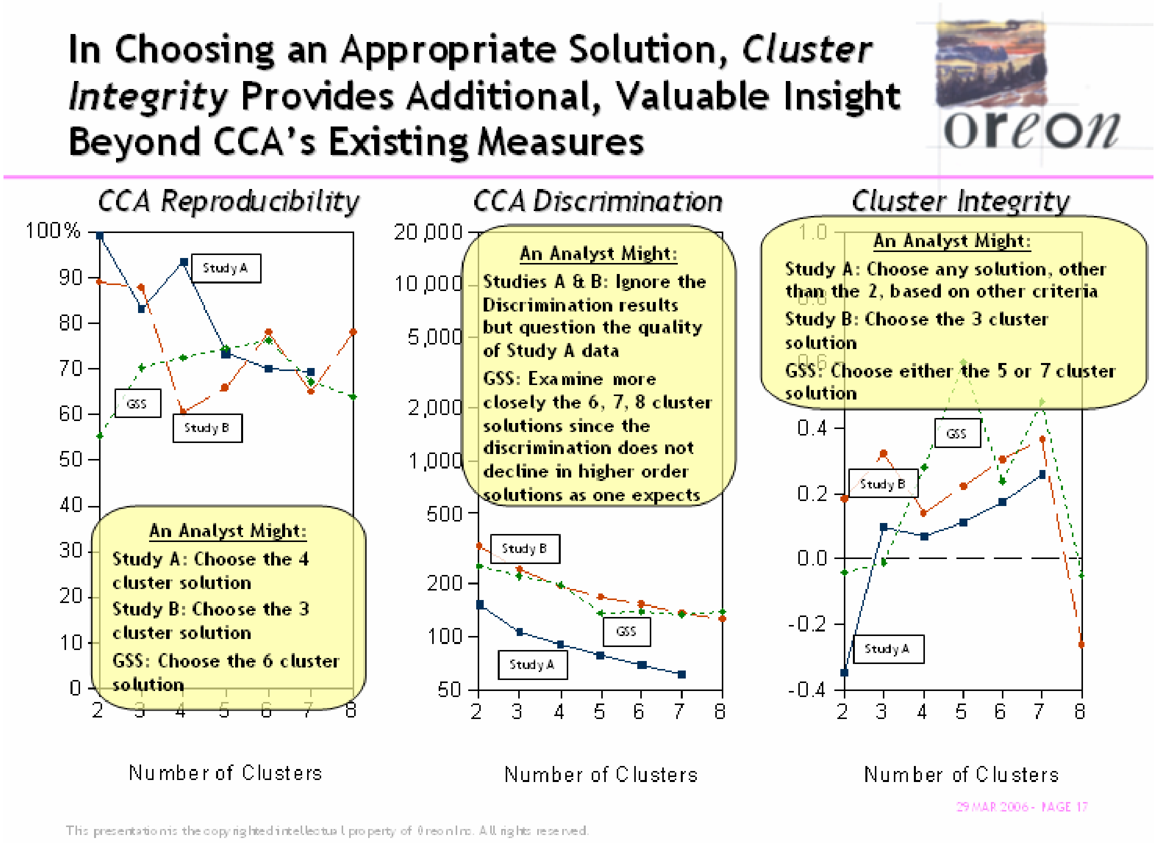


Three Commercial Datasets

- ✓ Study A
  - Internet-based survey
  - 46 *very lengthy* attitude statements
  - 10-point agreement scale
  - 1,600 respondents
- ✓ Study B
  - Internet-based survey
  - 37 attitude statements
  - 5-point fully anchored agreement scale
  - 3,750 respondents
- ✓ General Social Survey (National Opinion Research Center)
  - In-home personal interviews
  - 94 demographic and attitudinal variables (standardized)
  - 8,394 respondents
  - Data from 2000, 2002, 2004

29 MAR 2006 - PAGE 16

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



In Choosing an Appropriate Solution, *Cluster Integrity* Provides Additional, Valuable Insight Beyond CCA's Existing Measures

An Analyst Might:

Study A: Choose the 4 cluster solution

Study B: Choose the 3 cluster solution

GSS: Choose the 6 cluster solution

An Analyst Might:

Studies A & B: Ignore the Discrimination results but question the quality of Study A data

GSS: Examine more closely the 6, 7, 8 cluster solutions since the discrimination does not decline in higher order solutions as one expects

An Analyst Might:

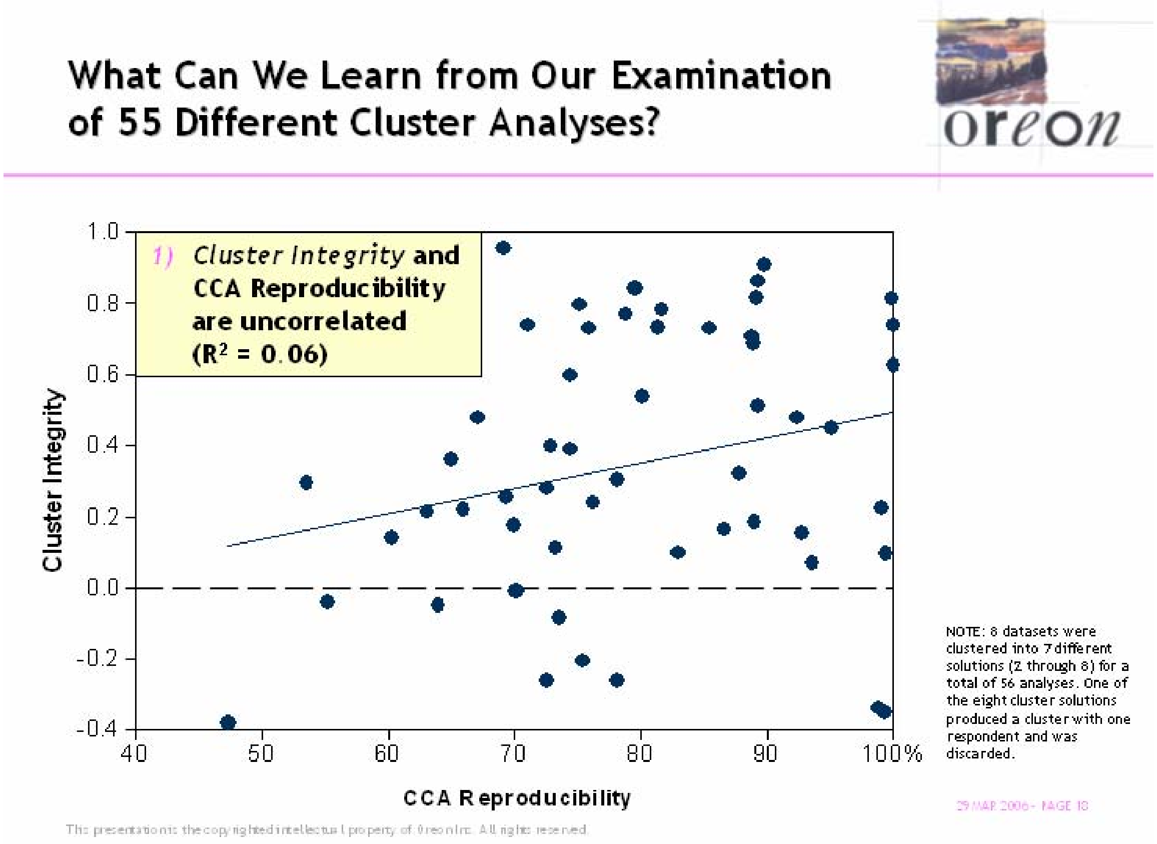
Study A: Choose any solution, other than the 2, based on other criteria

Study B: Choose the 3 cluster solution

GSS: Choose either the 5 or 7 cluster solution

29 MAR 2006 - PAGE 17

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.



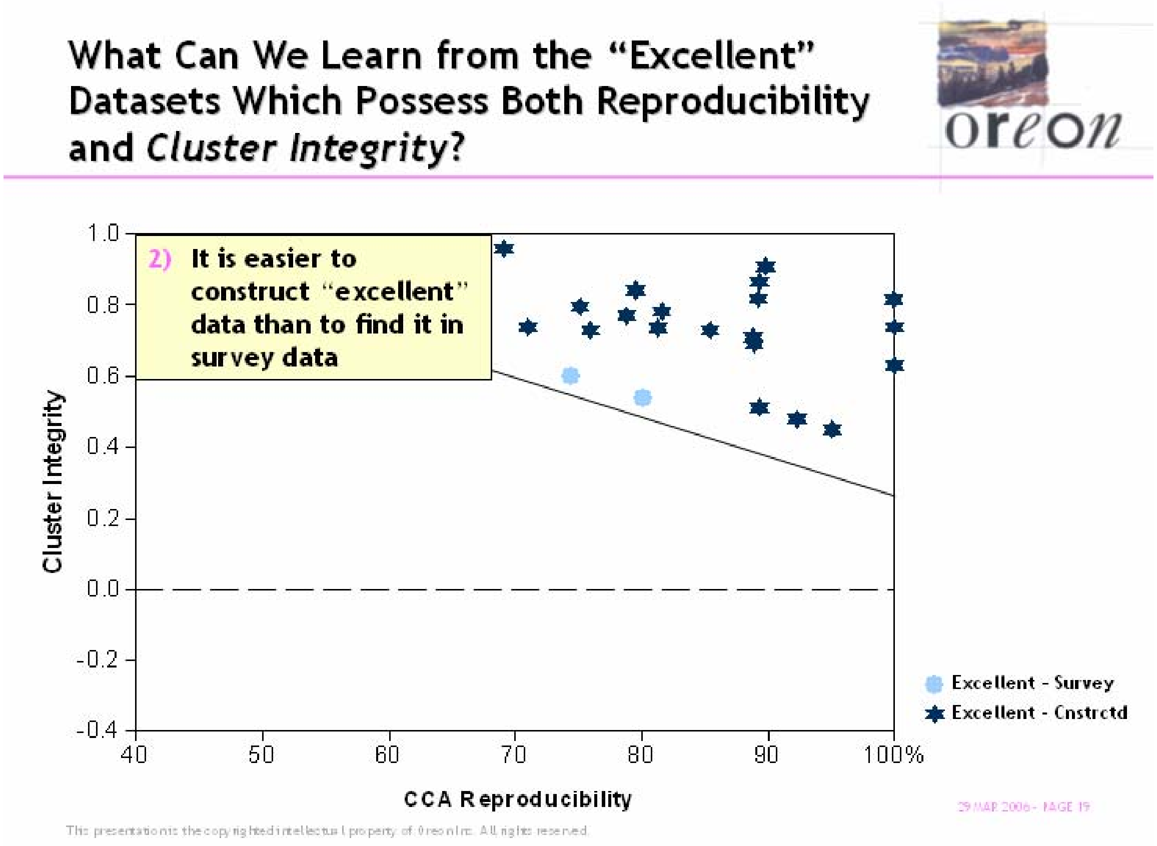
What Can We Learn from Our Examination of 55 Different Cluster Analyses?

1) *Cluster Integrity* and CCA Reproducibility are uncorrelated (R^2 = 0.06)

NOTE: 8 datasets were clustered into 7 different solutions (2 through 8) for a total of 56 analyses. One of the eight cluster solutions produced a cluster with one respondent and was discarded.

29 MAR 2006 - PAGE 18

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

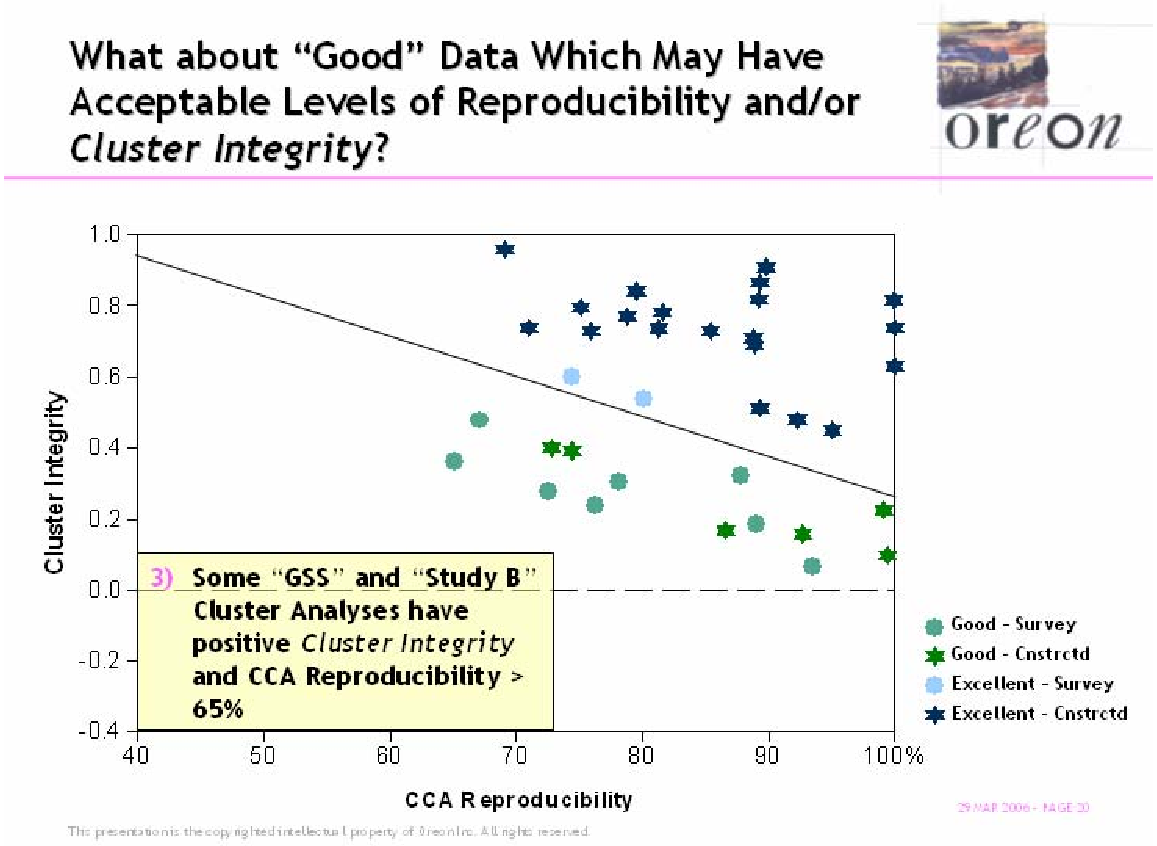


What Can We Learn from the "Excellent" Datasets Which Possess Both Reproducibility and *Cluster Integrity*?

2) It is easier to construct "excellent" data than to find it in survey data

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

29 MAR 2006 - PAGE 19



What about "Good" Data Which May Have Acceptable Levels of Reproducibility and/or *Cluster Integrity*?

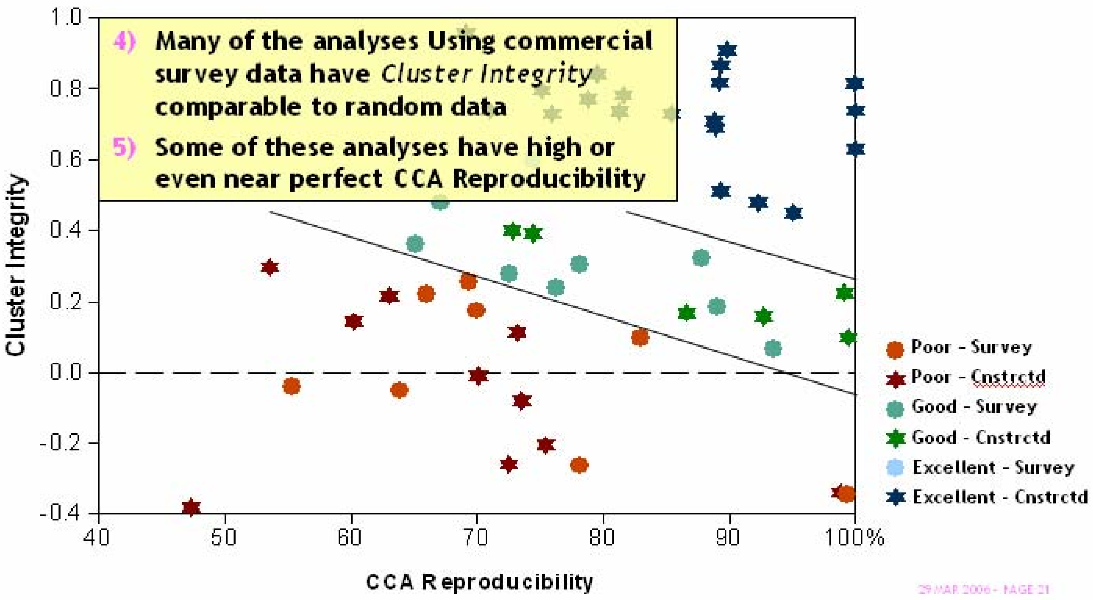
3) Some "GSS" and "Study B" Cluster Analyses have positive *Cluster Integrity* and CCA Reproducibility > 65%

This presentation is the copyrighted intellectual property of Oreon Inc. All rights reserved.

29 MAR 2006 - PAGE 20

What Can We Learn from the Analyses of "Poor" Data Not on the Previous Two Slides?

oreon



4) Many of the analyses Using commercial survey data have *Cluster Integrity* comparable to random data

5) Some of these analyses have high or even near perfect CCA Reproducibility

Conclusions (for Now)

oreon

- ✓ *Cluster Integrity* adds significant, new insight not only into choosing the best solution, but an excellent one as well, while recognizing that *CI* needs extensive, rigorous, further validation effort.
- ✓ The usefulness of the *CI* measure suggests developing a clustering algorithm that seeks to maximize *CI*.
- ✓ Encourage SSI to update the only piece of its arsenal that has not been touched for a decade while acknowledging that CCA's multiple replications with various starting point procedures still offer significant user benefits.
- ✓ We all agree to stop saying that "Cluster Analysis is more art than science."
  - If it is art, then it is not science, not even marketing science;
  - If we continue to say it is art, then we are offering an excuse for ignoring science.

IDENTIFICATION OF SEGMENTS DETERMINED THROUGH NON-SCALAR METHODOLOGIES

JOHN PEMBERTON AND JODY POWLETT
INSIGHT EXPRESS

ABSTRACT

Recent research in segmentation techniques has focused on the application of non-scalar research methodologies for the creation of segmentation bases that offer increased validity, reliability and substantiality when used within typical clustering algorithms. A variety of authors have proposed various forms of non-scalar data collection methodologies augmented by simulation estimation techniques to accomplish this task (Allenby & Fennell, 2002; Cohen, 2003; Chrzan, 2005; Pinnell and Fridley, 2005). The findings from this research appear promising towards the establishment of improved segmentation bases for practical applications.

A key issue in practical application that is not addressed in this research is the identification of resulting segments for future research involving new samples of respondents. Traditional segmentation research involving clustering of data collected via Likert scales is typically accompanied by the development of a "Typing Tool" that is often based upon discriminant analysis or tree-based algorithms (Chaid or CART). In segmentation schemes determined by non-scalar methodologies, identification with such tools is likely to be troublesome and detracts from the benefits of pursuing such techniques in the first place.

The authors propose an examination of a class of typing algorithms based on non-scalar data collection techniques. The exact tasks are determined by exhaustive combinatorial searches of the basis measures translated into popular sorting and selection tasks. Logic built upon these tasks guides the assignment of respondents into segments. This procedure is compared to traditional typing techniques and is evaluated on the ability to reproduce segments obtained via non-scalar techniques and traditional Likert questioning.

The authors would like to acknowledge useful comments made on preliminary drafts of the associated presentation offered by Jeff Dietrich and Enrico Rodriguez. Sara Stuteville helped refine the qualitative input into an appropriate attitudinal battery. Valuable assistance in the fielding of this study was provided by Rus Kehoe.

INTRODUCTION

The recent popularity of non-scalar data collection methodologies paired with Bayesian estimation techniques have provided the marketing researcher new bases for segmentation research. These segmentation bases do not exhibit the artifacts of those collected with traditional Likert scales. Previous research has emphasized the promise that these techniques bring to the researcher's tool set.

Within a marketing corporation the findings of a well executed segmentation study can take on a life of their own, with product development and marketing efforts tailored to appeal to target segments identified in the segmentation research. In order to test targeted concepts or to track marketing efforts within targeted segments, typing tools are customarily developed as part of the original segmentation to place new respondents into the established segments.

When segmentation studies are based on the clustering of items collected on a Likert scale, typing algorithms are usually based on a discriminant function. These typing tools involve a subset of the original segmentation attributes selected such that the discriminant function accurately predicts correct segment assignment above some given threshold within the original sample base or holdout sample.

For segmentation schemes involving a basis of utility simulations from a non-scalar methodology, there is no comparable approach for creating typing algorithms. Respondents in the original study indicate their preferences to either designed or randomized choice sets. The preference utilities extracted from these tasks are not directly observed for the original respondents and certainly not for subsequent respondents. Yet these preference utilities are the basis for the segmentation. Developing a typing algorithm for this methodology must result in a survey instrument where respondents reveal their preferences in a reliable way that can be used to assign segment membership.

The purpose of this paper is to formulate and test techniques that will allow segmentation schemes built with modern segmentation bases to be used with new sets of respondents.

METHODOLOGY

The subject selected for this research is related to attitudes concerning desserts. This topic was chosen to engage respondents with the hope of holding their interest throughout the course of the questionnaire. A short qualitative effort resulted in an initial list of attitudes which was refined down into 25 tested statements. A complete list of these 25 statements is included as appendix material.

The survey was fielded on-line using two different respondent audiences. The first audience was recruited via InsightExpress' proprietary e-RDD methodology. E-RDD seeks to identify a sample representative of the on-line community by using banner ads on a wide variety of websites to recruit prospective respondents to a brief demographic survey. Responses to this survey direct respondents through a queue to an appropriate survey.

The second audience was recruited from InsightExpress' community of pre-recruited respondents. This pool of respondents was originally identified through e-RDD intercepts and participants subsequently agreed to receive invitations for future survey requests.

The 25 attitudinal statements concerning desserts were presented to respondents with three different question types:

- **Traditional 5-point Likert agreement scale**. Attributes were randomized for each respondent and presented five on a screen across five survey screens.
- **Ranking tasks**. Attributes were randomized for each respondent and placed into five groups of five. Each group of five was presented on its own screen and respondents were asked to rank the items 1 to 5 by agreement from "Most agree with" to "Least agree with".

- **MaxDiff tasks**. Attributes were randomized for each respondent and placed into five groups of five. Each group of five was presented on its own screen and respondents were asked to identify the statement they "Most agree with" and the statement they "Least agree with".

Respondents were randomly exposed to two of the three question types described above. In each of the two exposures, respondents saw the same ordering of attributes across screens as the attribute randomization occurred only once. Additional usage and demographic measures were captured within the survey.

Across the two sample groups a total of 2252 respondents completed the survey. Across the three question types the following distribution was obtained:

Sample Universe	Likert Scales	Ranking Tasks	Max Diff Tasks
IX – e-RDD	752	716	700
IX – Pre Recruit	727	731	741
Total p/Question Type	1479	1447	1441

The respondents from both sample sources for each question type define the three segmentation sample groups.

UTILITY EXTRACTION

After close of field, the bases for the three segmentations needed to be created. For the Likert cell, this constituted screening out respondents who provided data with little or no variation across the scale points or that provided responses for less than 60% of the attributes. Missing data were imputed using a clustering algorithm.

For the ranking and MaxDiff cells, preference utilities needed to be extracted. Data from these tasks were coded to represent choice tasks. Using these tasks a HB-MCMC simulation utilizing Gibbs Sampling created the final preference utilities. The ranking simulation created utilities summarized by a Percent Certainty of .892 and RLH of .928. The MaxDiff simulation created utilities summarized by a Percent Certainty of .925 and RLH of .950. It may seem that the better model was obtained for the MaxDiff exercises, but it must be remembered that the MaxDiff model is only responsible for identifying the top and bottom attributes within each task. However, the ranking model is responsible for the exact placement of all attributes within a task.

As an aside, it is probably a good time to discuss what the educated reader might identify as a shortcoming in the design of the MaxDiff and Ranking tasks. In the design, each attribute was evaluated within a task once and only once. Traditional MaxDiff studies evaluate attributes multiple times within a fixed design.

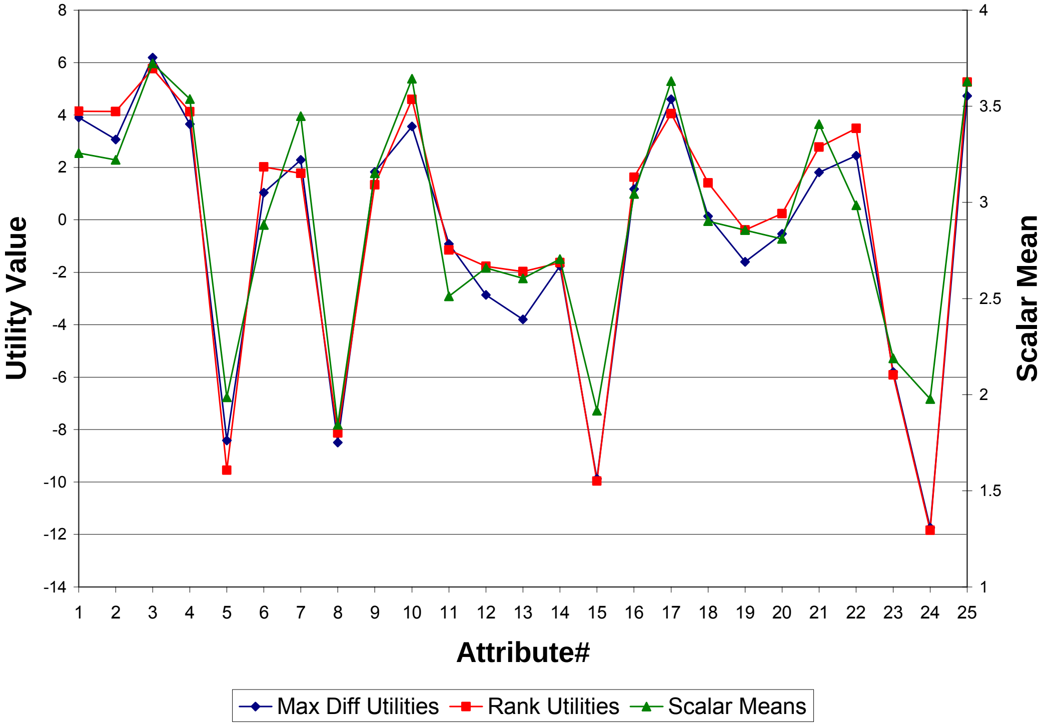
While it does not relate to the specific goal of this paper, part of the impetus of this research was to identify whether the respondent burden could be dramatically shortened with minimal impact on the extracted utilities. A planned design of 20 attributes typically involves five sets of four repeated in four to five rotations of the attributes. Overall, a respondent evaluates 20 to 25 sets of attributes. When compared to the Likert task of scales ratings on 20 attributes, we would have increased the burden on respondents right as the industry transitions to a methodology where respondent cooperation rates drop dramatically after 15 minutes of survey time. Our single exposure design is an effort to see how far we can push the envelope towards respondent convenience while still recovering reliable information.

Though these statements remain empirically untested, we posit that two factors may be on our side in creating viable utilities. First, we employed a complete randomization strategy for each respondent. Thus, exposures of preferred to preferred attributes vs. preferred to un-preferred attributes are evened out across the entire sample upon which each individual's utilities are indirectly based. Second, the questions used are an attitudinal battery. Practical researchers are accustomed to the idea that a structural correlation structure may exist between the attributes. The 25 attributes are very likely only measuring 5 – 10 distinct latent constructs. This correlation structure could potentially be recovered from the aggregate estimation of the covariance structure upon which individual estimates are drawn. So a design such as ours that has no repeated measures on the attribute level may in fact have quite a bit of redundancy on the latent construct level, should one investigate its magnitude and existence. Interesting research might center on the question as to whether correlation between tested attributes is a good proxy for repeated exposures in design.

Regardless of where opinion or scientific investigation nets out on this question, the subject of this paper remains unaffected. The typing procedure that is outlined below requires a respondent-level table of utilities. The technique can be applied regardless of the original derivation of the utilities themselves.

Table 1 summarizes the pattern in attribute preference across the three survey types at the total sample level. The visual pattern complemented by correlations above .95 for Likert scales compared to both the MaxDiff and ranking tasks indicate that the three methodologies are reliably placing the attributes in relative order of agreement.

Table 1
Aggregate Measurement Similarities



On first pass, reviewing reliability patterns across respondents at a disaggregated level is less optimistic. The two non-scalar methodologies have an average correlation across all attributes and respondents of .525; Likert and Ranking .495; Likert vs. MaxDiff .492. It would seem that the single exposure strategy may have pushed the envelope too far towards respondent convenience. To provide some context on the implication of these correlations, identical research was fielded with the same set of respondents at a subsequent date to the first study. Within the sample of respondents for whom the attributes were collected via Likert scales it was found that the wave to wave correlations for individual respondents across all attributes actually did worse, a correlation of .392.

SEGMENT CREATION

After the bases are created, the next step in the process is to create the cluster solutions. Consistent with commonly used protocol, hierarchical clusters are created in order to create seeds for the final K-means clustering algorithm. The final solution selected would be based on a mixture of statistical and interpretational considerations.

In the current study, this approach was altered slightly. The segmentation process occurred independently for each of the three bases and samples. The decision of which particular solutions were chosen to incorporate into the typing exercise was coordinated across the three cells. Solutions for each basis needed to be of equivalent magnitude (# of clusters). Otherwise, typing tools for a basis where a smaller solution was selected may have performed better than a basis that utilized a larger solution.

Across all three methodologies, six cluster solutions were found to fit the data fairly well, based both on the Cubic Clustering Criterion and based on a cross validated discriminant model that utilized all available attributes. At this point, with segmentation schemes in place, the typing can begin.

THE NON-SCALAR TYPING ALGORITHMS

A The Pairs

The strategy for typing the non-scalar segmentation bases is built upon the premise that the instrument must force respondents to indicate their preferences across a subset of attributes. Within the basis itself, pairwise comparisons of attributes can be made. Based on the magnitude of the utilities involved, a prediction as to the likely choice can be made. If the attribute pairs which are most reliable in predicting segment membership can be identified, then the typing tool can be constructed based on pairwise preference questions fielded to future respondents.

The first step in the process is to identify all possible pairs of attribute utilities within the Ranking and Max/Diff segmentation bases. There were 25*24/2 = 300 possible pairings of utilities. Each was numerically indexed to represent when X was compared to Y. From this comparison of utilities if X > Y then the prediction variable was coded "1" if X < Y then the prediction variable was coded "0."

These 300 pairwise preference indicators mimic the choices respondents would make if they were presented with the actual task of indicating a preference between two attributes. The typing tools that will be outlined shortly represent two different strategies for selecting the particular pairs which best predict segment membership and the functional form required to implement them in practice.

B The Discriminant Approach

The first proposed approach for identifying the specific pairs is similar to the commonly applied approach for scalar segmentation work. A stepwise linear discriminant analysis is run where the predicted variable is cluster membership and the predictors are the pairwise preference indicators.

The stepwise algorithm is run to identify the first "Z" pairs for consideration. Then the Z pairs are used to predict membership. The relative success can be measured through classification rates in a learning/test division of the sample or by cross validation. Then the same step occurs for the first Z-1 pairs, then the first Z-2 pairs, etc.

C The Bayesian Approach

If we are going to utilize a Bayesian specification to establish the non-scalar bases then we might as well go all out and use a Bayesian updating procedure to identify the best pairs to use also. The discriminant approach outlined below utilizes a linear functional form. It stands to reason that this form could well serve as a restriction. Whereas an approach that is less reliant on a specified form could well classify better.

Since both the probabilistic variable and the observed predictors are nominal with finite categories, updating with Bayes rule becomes a possibility. Though somewhat cumbersome to apply, its less restrictive specification may allow for a better prediction percentage.

It is well known that the specification of Bayes' rule is as follows:

$$P(Si \mid Pj) = \frac{P(Pj \mid Si)}{\sum_{i=1}^{N} P(Pj \mid Si)}$$

Where Si represents the ith segment, N equals the number of segments and Pj indicates the outcome of the jth pair of all possible pairs. The steps to implement this rule as a search algorithm are as follows:

1. Create initial probability P(Si) for each segment. In this case equal probability of membership for each segment was assumed.
2. Create crosstabs for each preference indicating pair variable and segment membership variable (col = pair choice (Pj), row = segment (Si)). By saving the row percentages, P(Pj|Si), we have created a table of all conditional probabilities used in the updating algorithm.
3. Use the conditional probabilities and Bayes theorem to update the prior segment probabilities for every possible preference indicating pair.
4. Assign segment membership based on the P(Si|Pj) with the highest probability. This decision places a respondent into a predicted segment.
5. Compare the predicted segment to the actual segment membership and create a misclassification variable, indicating correct with a "1" and incorrect "0."

6. Summarize the performance of each pair by averaging the misclassification variable for each of the 300 pairs. Using the attribute pair that scored the highest correct classification percentage, update the previous prior segment probabilities with that pair's conditional probabilities.

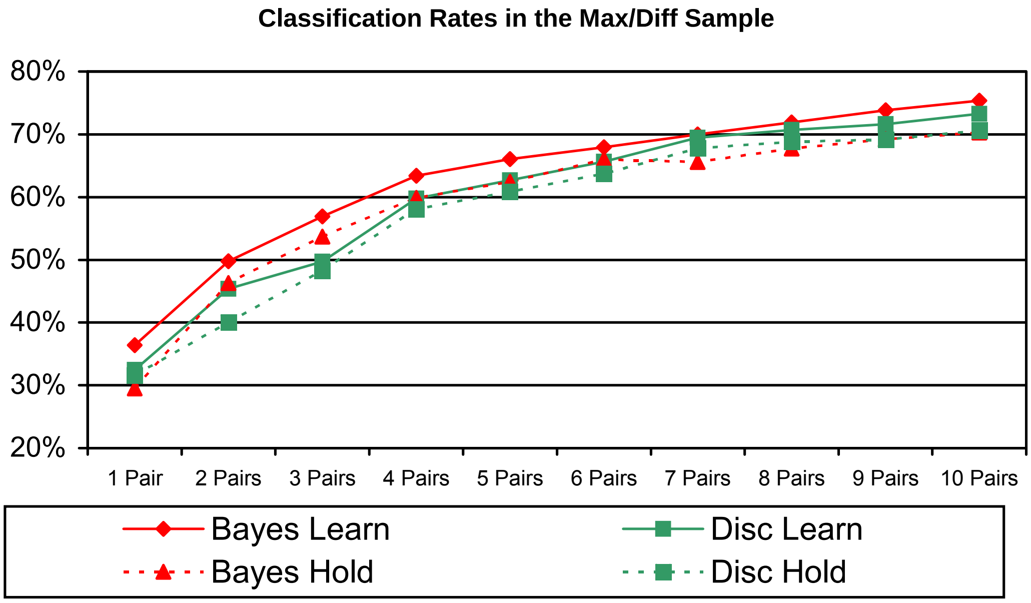
These updated probabilities are the new prior probabilities for segment membership, and steps 3 – 6 can now be repeated until "Z" pairs have been identified for inclusion into the typing algorithm.

Performance of the Algorithms

In order to ascertain the comparative performance of the two tested algorithms, the samples were split into learning and holdout samples. Comparisons were made within the MaxDiff and Ranking task samples of respondents. Tools ranging in size from one to ten pairs of attributes were estimated.

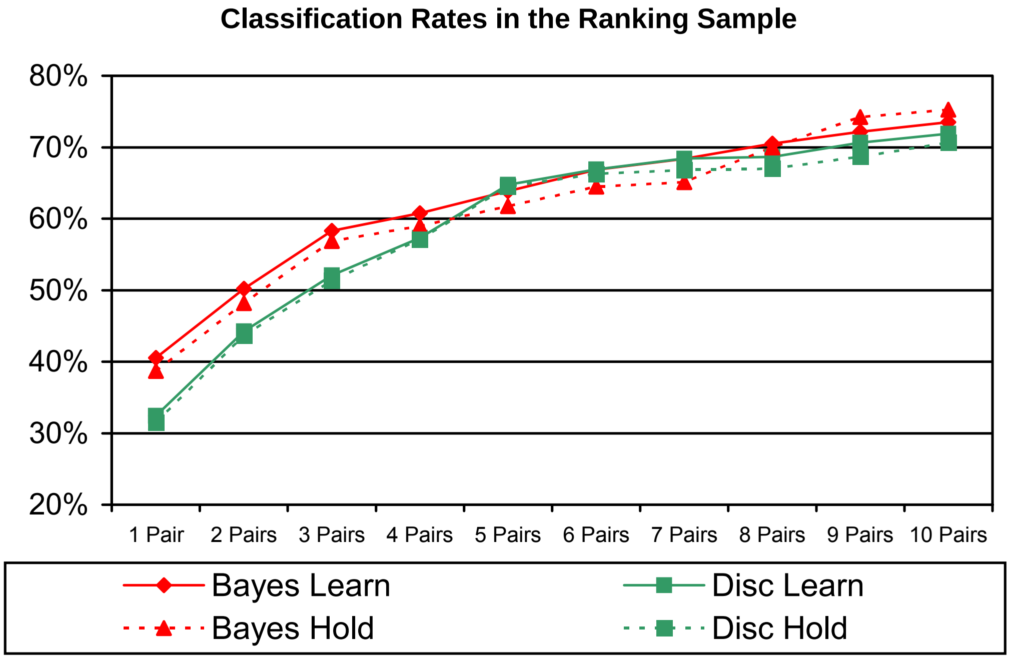
For the samples tested, the largest gains in prediction accuracy are gained with the addition of the first five pairs into the tool. The returns on prediction percent diminish thereafter. These diminishing returns will be common to all similar typing tools, though the shapes of the curves themselves will vary depending on the size and complexity of the segmentation involved.

Early within both tools (roughly four pairs or fewer) the Bayesian tools outperform the discriminant tools. As the number of pairs added increases, this performance gap diminishes in the learning samples and often disappears in the holdout samples for both the MaxDiff and Ranking task samples.



For the MaxDiff sample specifically, within the learning sample, as the size of the typing tool increases, the Bayesian algorithm consistently outperforms the discriminant, from a 7.2% gap for three pairs diminishing to a gap of .5% for seven pairs. Within the holdout sample performance for both tools was "choppy" as the number of pairs increased. At three pairs the widest gap was observed with the Bayesian tool out performing the discriminant tool by 6.3%. From six pairs on, there were no significant differences in the holdout sample between the two methodologies.

Similarly for the Ranking Tasks Sample, within the learning sample, the Bayesian algorithm outperforms the Discriminant algorithm for smaller typing tools (four or fewer pairs). Performance is comparable for five or more pairs used. Within the holdout sample, the Bayesian algorithm outperformed the discriminant algorithm for smaller numbers of pairs and for larger numbers of pairs used. From five pairs on the two algorithms performed nearly identically.



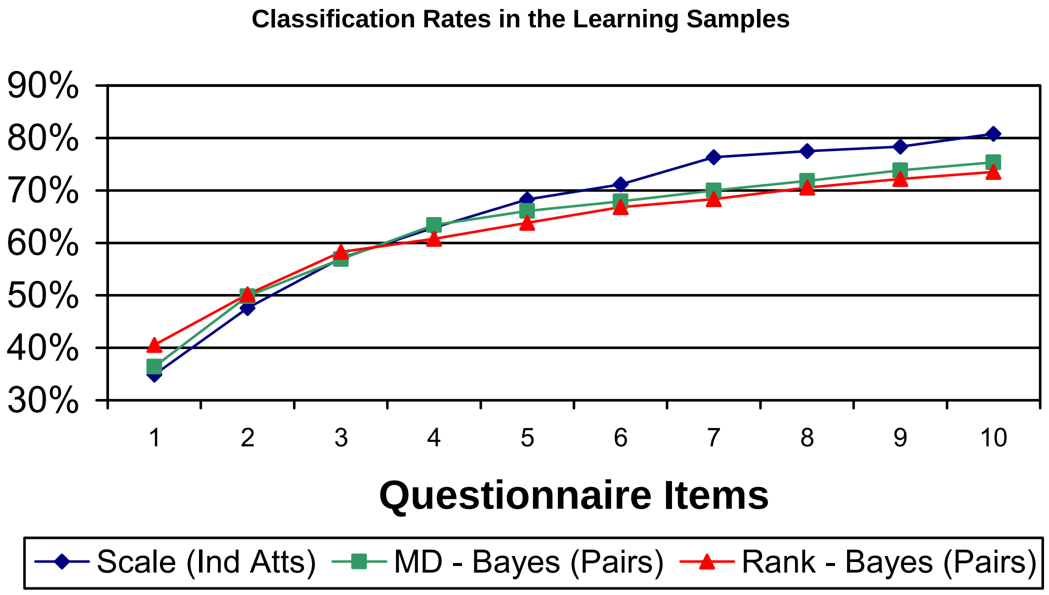
It would appear that the superiority of the Bayesian tool in smaller tools compensates for an overall lack of information available to the tool. As the amount of information increases, key knowledge gaps disappear and the performance differences due to functional form are overwhelmed by the availability of information. Given the differences in computational complexity, in most circumstances the gains from the Bayesian tool do not recover their computational costs.

Comparisons to Traditional Segmentation and Typing

This paper has shown that typing algorithms for segmentation schemes involving non-scalar bases can be created. The question remains as to how they perform compared to traditional segmentations based on Likert scales and discriminant analysis.

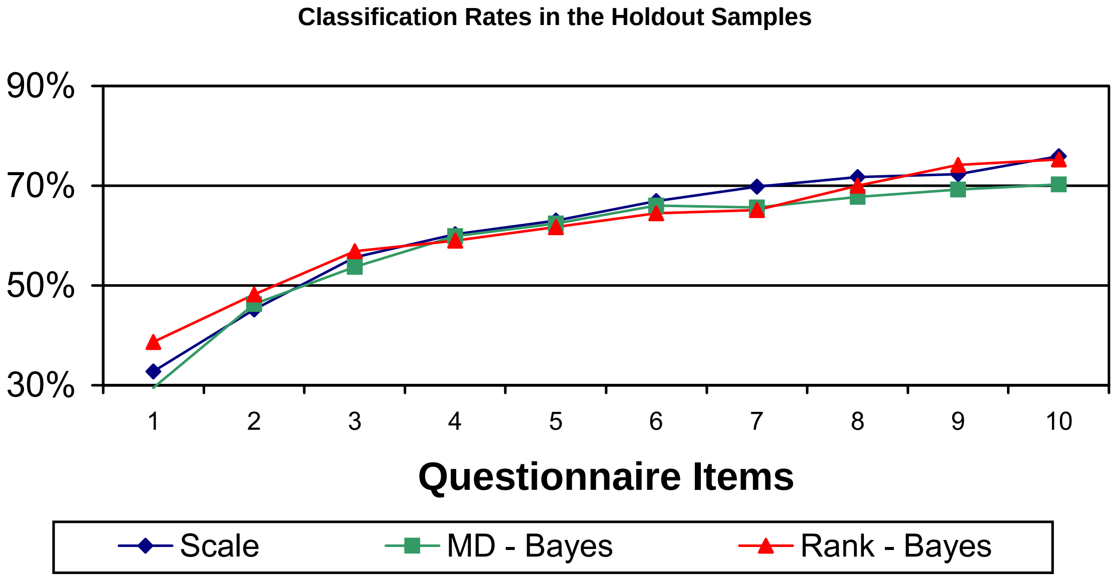
To answer this question, the Bayesian tools for the non-scalar segmentations are compared to the discriminant tool from the scalar segmentation.

The magnitude of the tools involved and compared will be defined by the number of items fielded to respondents. For a traditional scalar tool this refers to the number of attributes a respondent evaluates on the rating scale. For the non-scalar tools this refers to the number of attribute pairs a respondent must evaluate.



In comparing the top performing algorithms (Bayesian) from the two non-scalar segmentations to a traditional scalar segmentation typed by a discriminant function using individual attributes, we find that the non-scalar algorithms compare well for smaller tools. The traditional scalar approach opens a gap when the number of items in the typing tool increases in magnitude. In these examples, a significant gap opens for tools larger than seven.

A review of the holdout sample yields a different story. For the holdout sample all tools seem to perform similarly, with no meaningful performance gaps across the magnitude of typing tools tested. This may suggest that the traditional typing of scalar based segmentations with discriminant models could be over fitting the learning sample. When respondents not included in the creation of the tool are included, the tools perform similarly.



Implementing the Algorithms in Practice

Developing a tool for practical use would begin as identified above, creating the pairs outlined above and then using the search routine for either the Discriminant or Bayesian specification. The number, Z, of pairs would be determined based on the accuracy demanded for the application of the tool.

For future respondents, these pairs would be presented to respondents as pairwise choice tasks with wording that closely matched the spirit of the initial tasks in the segmentation study. On the back end, the responses would be scored by the discriminant function with the most likely segment identified, or the responses would be used to update segment membership probabilities using Bayes' rule.

While we have not empirically tested this in any fashion, it might be assumed that a typing tool of this form may be more readily transferable from sample group to sample group and possibly across cultures. With the Scale being removed and replaced by the less culturally sensitive task of choosing between two statements, it could very well become much easier to assess the changes in segment incidence across cultures revealing more reliable inferences on the values, attitudes or criteria in multinational research efforts.

Conclusions

Within this paper the issue of typing future respondents into segmentation schemes established with non-scalar bases has been explored. Two potential candidate tools were identified, both related to presenting respondents pairwise comparisons of attributes initially identified using respondent level utilities from the initial segmentation basis. Using these pairs, tools based on a linear discriminant function and Bayesian updating of segment membership probabilities were described.

In testing these tools, it was determined that when ample information (pairs) were included into the typing algorithm, the prediction accuracy gap that initially existed between the Bayesian and Discriminant tools closed and became indifferent. When comparing the Bayesian tool for the non-scalar segmentation schemes to a traditional segmentation and discriminant based typing tool, it was discovered that within the holdout test sample the non-scalar typing tools predicted with similar accuracy as did the traditional approach.

References

Allenby, Greg and Geraldine Fennel, "Conceptualizing and Measuring User Wants: Understanding the Source of Brand Preference," 13th Annual AMA Advanced Research Techniques Forum, June, 2002, Vail, Colorado.

Cohen, Steve, "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," 2003 Sawtooth Software Conference, San Antonio, Texas.

Chrzan, Keith, "An Empirical Test of Three Attitude Scaling Techniques," 16th Annual AMA Advanced Research Techniques Forum, June 2005, Coeur d'Alene, Idaho.

Pinnell, Jon and Lisa Fridley, "Measurement and Scaling Methods Independent of Response Style," Papers and Proceedings, Sixtieth Annual AAPOR Conference Program.

APPENDIX

Attitudinal battery fielded:

- I eat desserts to splurge or reward myself
- Ice cream is the best choice for dessert
- Desserts are more enjoyable when shared with friends
- I like the mixture of hot/cold in desserts, like pie a la mode
- Desserts could be eliminated from the menu entirely
- I only have dessert on special occasions
- Custard type desserts, such as crème brulee, are delicious
- I will not have dessert if I can't have a cup of coffee with it
- I usually skip dessert
- I like cake for dessert
- I have dessert at restaurants more often than at home
- In order to be health conscious, I like to order fruit for dessert
- A meal isn't complete without dessert
- Sometimes I like dessert before dinner as a snack
- An after-dinner drink is my dessert
- Due to nutritional reasons, I limit the amount of desserts I consume
- Restaurants should offer smaller size desserts for sampling
- Desserts served cold are more tasty than hot desserts
- Desserts should be less sweet and more savory
- I sometimes limit how much dinner I eat to make sure I can include a dessert
- I like desserts with a fruity taste to them
- Almost all the desserts I eat are Chocolate in flavor
- An after-dinner Coffee or Tea is all the dessert I need for dessert
- I like after dinner Drinks like Port, Brandy or sweet Alsatian Dessert wines
- Dessert puts me in a good mood

REVERSE SEGMENTATION: AN ALTERNATIVE APPROACH

URSZULA JONES, CURTIS L. FRAZIER, CHRISTOPHER MURPHY,
MILLWARD BROWN
JOHN WURST
SDR/UNIVERSITY OF GEORGIA

I. BACKGROUND

Most segmentation work segments respondents based on attitudes and behaviors, or segments them on demographics or firmographics. Both approaches have important problems regarding actionability.

While segmentation on attitudes and behaviors results in distinct and meaningful segments in terms of attitudes, those segments tend to be difficult, if not impossible, to identify in the market or databases. From attitudinal segmentation results we know how to message the product/service, but we don't know who to message it to.

On the other hand, segmentation based on demographics or firmographics results in easy to identify segments that are not distinct in ways that are actionable to marketers. Often times, resulting segments are not differentiated enough in terms of attitudes and behaviors. Therefore, segments are reachable, but it is not clear how to communicate with them.

When undertaking a segmentation project, our key objective is to provide our clients with meaningful segments that are actionable to their marketing team. Our clients demand segments that are not only different in terms of attitudes and behaviors, but also are identifiable in the market place. Often times, our clients want to use our segmentation scheme as basis for creation of specific marketing programs that reach segments of interest. As the result, we have been working on a unique approach, reverse segmentation, which significantly increases the potential to find identifiable and actionable segments that can be also linked to our client's internal database.

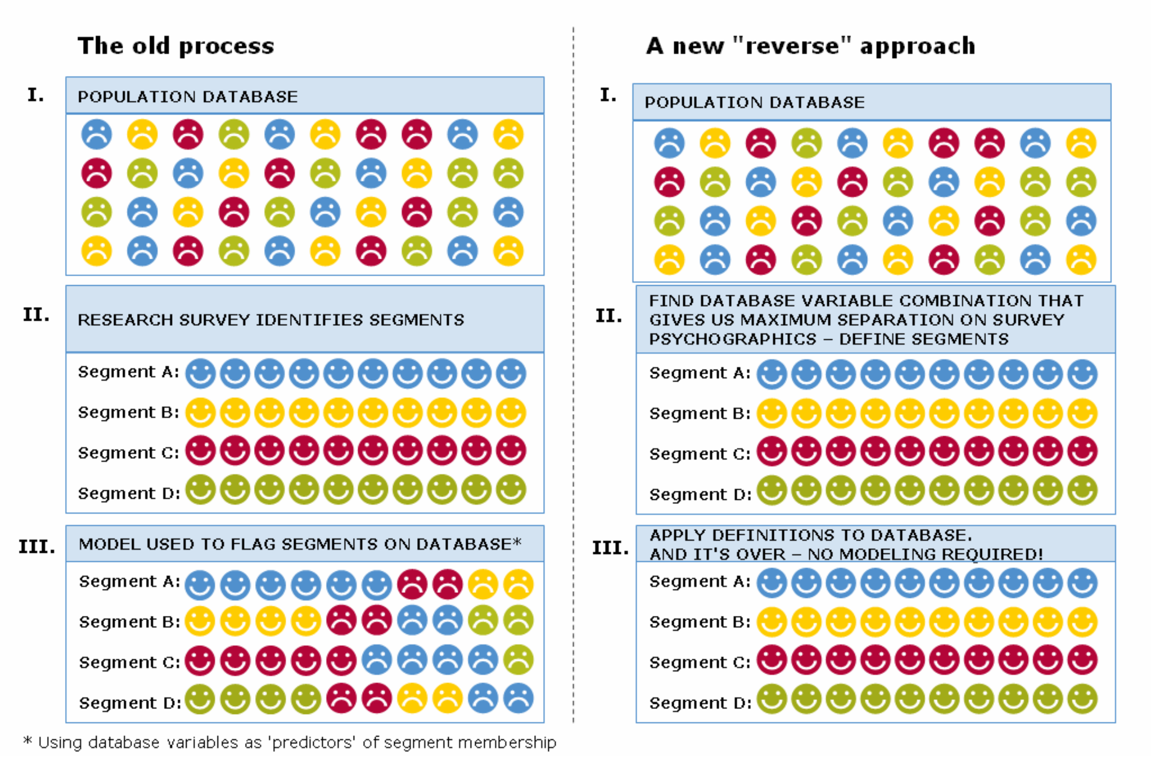
II. WHAT IS REVERSE SEGMENTATION?

Reverse segmentation is an approach that creates perfectly identifiable groups that are differentiated in terms of attitudes and behaviors. In other words, it connects differences in behavior and attitudes to known targetable attributes such as demographics/ firmographics, media usage, or channel usage.

III. WHY AND WHEN REVERSE SEGMENTATION CAN BE USED, AND HOW CAN CLIENTS BENEFIT FROM IT?

We’ve often found high misclassification rates when applying scoring algorithms using database variables to traditional segmentation results. When the reverse segmentation approach is implemented there is no misclassification since the methodology creates perfectly identifiable segments. (see Figure 1). When conducting a reverse segmentation we typically focus on demographics (firmographics) that are currently within the client's database to maximize actionability.

Figure 1


The old process
I.
POPULATION DATABASE
II.
RESEARCH SURVEY IDENTIFIES SEGMENTS
Segment A:
Segment B:
Segment C:
Segment D:
III.
MODEL USED TO FLAG SEGMENTS ON DATABASE*
Segment A:
Segment B:
Segment C:
Segment D:
* Using database variables as 'predictors' of segment membership
A new "reverse" approach
I.
POPULATION DATABASE
II.
FIND DATABASE VARIABLE COMBINATION THAT GIVES US MAXIMUM SEPARATION ON SURVEY PSYCHOGRAPHICS – DEFINE SEGMENTS
Segment A:
Segment B:
Segment C:
Segment D:
III.
APPLY DEFINITIONS TO DATABASE. AND IT'S OVER – NO MODELING REQUIRED!
Segment A:
Segment B:
Segment C:
Segment D:

Reverse segmentation is not intended to replace the traditional approaches. It serves a different purpose and can even co-exist with other strategic segmentation schemes. As with traditional approaches, it can be used in any consumer or B2B study. However, it is particularly useful when a client is interested in flagging his/her database, and/or is interested in developing specific marketing programs. Linking attitudinal and behavioral data with demographics, firmographics, media usage, or shopping channel, enables accurate segment targeting. Again, the ultimate payoff is eliminating segment misclassification.

IV. How Does Reverse Segmentation Work?

Reverse segmentation creates a collection of “objects.” These objects represent different combinations of the client's targetable attributes.

Consider the following example. Suppose that the client's database or demographics of interest include four variables—education (college:yes/no), income (high/low), gender, and children (yes/no). Based on those four variables, there are sixteen possible combinations, or objects (see Figure 2).

Figure 2

	College	Income	Gender	Children
Object 1	No	Low	Female	No
Object 2	No	Low	Female	Yes
Object 3	No	Low	Male	No
Object 4	No	Low	Male	Yes
Object 5	No	High	Female	No
Object 6	No	High	Female	Yes
Object 7	No	High	Male	No
Object 8	No	High	Male	Yes
Object 9	Yes	Low	Female	No
Object 10	Yes	Low	Female	Yes
Object 11	Yes	Low	Male	No
Object 12	Yes	Low	Male	Yes
Object 13	Yes	High	Female	No
Object 14	Yes	High	Female	Yes
Object 15	Yes	High	Male	No
Object 16	Yes	High	Male	Yes

The way reverse segmentation is different from traditional approaches is that instead of clustering respondents on attitudinal and/or behavioral data, we cluster objects on attitudinal and/or behavioral data. In order to create our final segments, we combine objects in different ways so significant differentiation on attitudinal and behavior attributes is achieved. This two-step process ensures that our segments are identifiable (since the structure from step 1 is based on demographics or firmographics), and that they are as meaningful as possible (based on combining the objects in step 2).

V. How to Choose Targetable Attributes That Will Make up the Objects?

Typically, we identify targetable attributes that are either available in the client's database or are thought to be actionable by the client's marketing team. However, some of those targetable attributes are not going to provide us differentiation on attitudinal and/or behavioral variables. Therefore, we use standard ANOVA tests to determine which behavioral/attitudinal questions are significantly different by various targetable attributes. This also allows us to reduce the list considerably, limiting it to targetable attributes that are best discriminators for the behavioral/attitudinal questions.

VI. How Are the Segments Derived?

Consider all ways of putting objects into 2, 3, 4,… segments. The number of possible solutions can be very large when objects consist of several targetable attributes with several levels within the attribute. Segmentation by hand would be impractical and oftentimes impossible. The way we have been tackling this problem is by data summarization. We summarize the basis variables and use standard clustering software on the summarized file. The way this works is as follows. For each object, we compute the basis variable means and cluster the file of objects using the means as basis variable values.

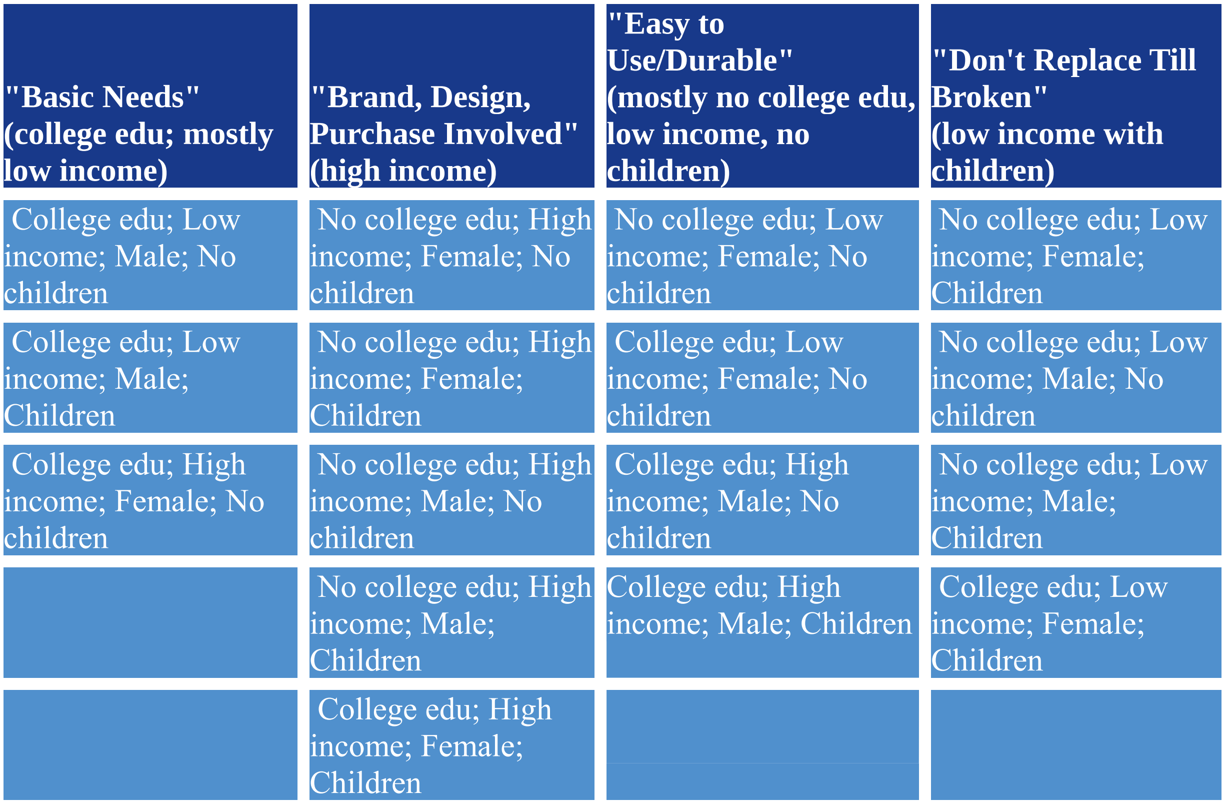
VII. What Is the Outcome?

The initial outcome of reverse segmentation is slightly different than the outcome of the traditional approach. Initial segments consist of objects instead of respondents. In order to look at the data at the respondent level, respondents have to be assigned to segments based on the object that they belong to. After that additional step is accomplished, the results of reverse segmentation should be evaluated using the same criteria for evaluating traditional segmentation approaches. Segmentation schemes should be evaluated in terms of the following:

- Are the segments identifiable and meaningful? Segments resulting from reverse segmentation are identifiable, but are they meaningful? Sometimes objects that group together may not make intuitive sense, therefore an analyst running reverse segmentation must choose the solution that is not only statistically sound, but also the one that makes most sense.
- Are segments substantial? Are the segments “big” enough to justify targeted marketing efforts?
- Are segments accessible? Can you reach them, either by advertising or more direct sales approaches?
- How are the segments in terms of responsiveness? Will these segments respond differently to our marketing efforts?
- Is the segmentation scheme actionable? Given your goals and competencies, will you be able to act upon the segmentation?

Figure 3 illustrates how different objects are segmented. The data comes from a segmentation study for durables. 3024 people participated in the study.

Figure 3



"Basic Needs" (college edu; mostly low income)	"Brand, Design, Purchase Involved" (high income)	"Easy to Use/Durable" (mostly no college edu, low income, no children)	"Don't Replace Till Broken" (low income with children)
College edu; Low income; Male; No children	No college edu; High income; Female; No children	No college edu; Low income; Female; No children	No college edu; Low income; Female; Children
College edu; Low income; Male; Children	No college edu; High income; Female; Children	College edu; Low income; Female; No children	No college edu; Low income; Male; No children
College edu; High income; Female; No children	No college edu; High income; Male; No children	College edu; High income; Male; No children	No college edu; Low income; Male; Children
	No college edu; High income; Male; Children	College edu; High income; Male; Children	College edu; Low income; Female; Children
	College edu; High income; Female; Children		

"Basic Needs" segment consists of college graduates of which the majority has low income. Those individuals are minimalists and will only consider the cheapest products. They display little interest in technology, design and innovation in part because they only need their products to serve basic functions.

"Brand, Design, Purchase Involved" segment is made up of individuals with high income. These consumers look for products that provide advanced technology and beautiful design. They like to entertain and feel that the products they own say a lot about themselves. They are very involved in the purchase process.

"Easy to Use/Durable" segment represents people, who for the most part have no college education, have low income, and have no children. Dependability and ease of use is the name of the game for this group. Products have to be long-lasting and straightforward in usage. They equate technology and feature-laden products with hassle and complications.

"Don't Replace Till Broken" segment is comprised of people with low income and with children. Those people put little weight on the product. They don't care about product design or brand. They also don't intend to replace the product until the old one is broken.

VIII. Validation

When validating a segmentation scheme the most important criterion is checking whether the selected segmentation solution makes intuitive sense. It does not matter how statistically sound a solution is, if segments are not meaningful.

Another important validation measure is checking whether the segmentation scheme is projectable to the population. In order to validate results in terms of projectability, we recommend randomly splitting the data into test and holdout samples. Test segments should be compared to holdout segments and tested for significance. For a solution to be valid, very few statistical differences should be found in behavioral/attitudinal data of two samples.

IX. CLIENT DELIVERABLES

One of the great benefits of this method is ability to score the client's database and/or assign survey respondents with accuracy to the appropriate segment. Segment assignment is based on the respondent's demographic, firmographic, media, or channel usage profile and is automatic, no modeling required.

X. CASE STUDY

In order to compare reverse segmentation to traditional approaches, we conducted a segmentation study using three methodologies: reverse segmentation, segmentation on attitudinal data, and segmentation on demographics. We used the durability study data that was mentioned earlier.

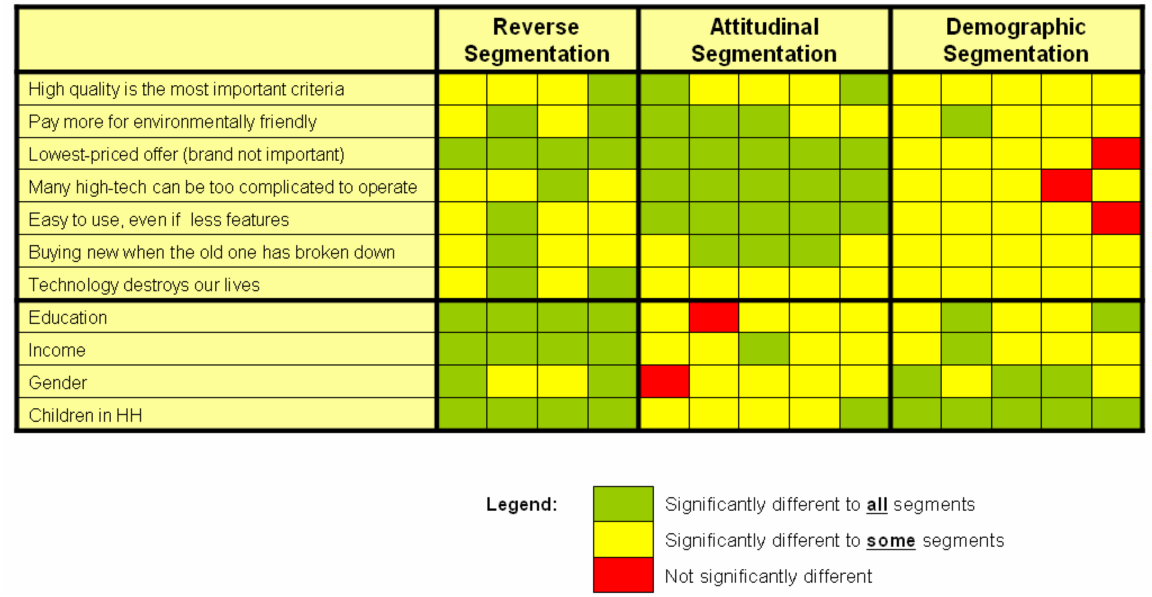
Figure 4 demonstrates the amount of differentiation for the attitudinal data and for the demographics in all three approaches. Interpretation of the table is straightforward. Each column within a segmentation approach represents a segment. Reverse segmentation generated a four segment solution, while both traditional approaches generated five segment solutions. Color coding represents the outcome of significance test.

Segments marked as green signify that they are significantly different from all segments within a particular segmentation scheme. For example, in reverse segmentation, segment 4 is significantly different from all other segments on the following attitudinal statements: "high quality is the most important criteria," "pay more for environmentally friendly," "lowest-priced offer," "technology destroys our lives."

Segments marked as yellow indicate that they are significantly different from some of the segments. For example, in reverse segmentation, segment 4 is significantly different than some of the segments on the following attitudinal statements: "many high-tech can be too complicated to operate," "easy to use, even if less features," "Buying new when the old one has broken down."

Segments marked in red indicate that a particular segment is not significantly different than the other segments. For example, in the attitudinal segmentation, segment one is not significantly different in terms of gender split.

Figure 4



	Reverse Segmentation				Attitudinal Segmentation					Demographic Segmentation				
High quality is the most important criteria	🟨	🟨	🟨	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟨	🟨
Pay more for environmentally friendly	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟨	🟨
Lowest-priced offer (brand not important)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟥
Many high-tech can be too complicated to operate	🟨	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟥	🟨
Easy to use, even if less features	🟨	🟩	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟥
Buying new when the old one has broken down	🟨	🟩	🟨	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟨	🟨
Technology destroys our lives	🟨	🟩	🟨	🟩	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Education	🟩	🟩	🟩	🟩	🟨	🟥	🟨	🟨	🟨	🟨	🟩	🟨	🟨	🟩
Income	🟩	🟩	🟩	🟩	🟨	🟨	🟩	🟨	🟨	🟨	🟩	🟨	🟨	🟨
Gender	🟩	🟨	🟨	🟩	🟥	🟨	🟨	🟨	🟨	🟩	🟨	🟩	🟩	🟨
Children in HH	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩

Legend:

🟩	Significantly different to **all** segments
🟨	Significantly different to **some** segments
🟥	Not significantly different

As you probably already noticed, reverse segmentation gives us good segment separation in terms of attitudinal statements, and gives us excellent separation in terms of demographics.

On the other hand, attitudinal segmentation outperforms reverse segmentation in terms of segment separation on attitudes. However, it gives us very little separation in terms of demographics. Since little separation is found in terms of demographics, this brings up the problem of not being able to pinpoint the people we want to message to.

In demographic segmentation, there is good separation in terms of demographics, but very little separation in terms of attitudes. Here we are dealing with the "what do we say to those people" problem.

XI. Problems/Issues with the Approach

While reverse segmentation offers some important advantages, there are also some associated problems/issues with the approach. Reverse segmentation is very time consuming. It requires some back and forth steps to determine the best set of object definitions. It is also restricted in terms of solutions it can generate, since instead of segmenting respondents, we are segmenting objects. Additionally, since the approach is fairly new it requires more validation work. Reverse segmentation should be tested for the quality of the solution in terms of marketing decisions. There should be also more work which tests how well the solution holds up when compared to holdout data.

XII. Conclusions: Giving Clients Fact-Based Options

Because reverse segmentation is still an experimental technique, we advise our clients to take parallel analytic paths with the segmentation dataset. We recommend taking the "reverse" track and the traditional track. At the end of the analytic process, the client can then choose which path makes more sense for them. So far, it has come down to the following decision for our clients:

- Would we rather get diluted attitudinal/behavioral separation via the definitional/reverse approach? If so, how much?
- How much classification error do we get with the traditional approach? Do we get better separation on attitudes/behaviors going down this path? If so, is it enough to warrant accepting the classification error?

The answers to those questions will largely be driven by the application—and there are instances in which both results could, and in fact, have, been utilized.

TESTING FOR THE OPTIMAL NUMBER OF ATTRIBUTES IN MAXDIFF QUESTIONS

KEITH CHRZAN
MARITZ RESEARCH
MICHAEL PATTERSON
PROBIT RESEARCH

INTRODUCTION

Maximum difference ("maxdiff") scaling (Finn and Louviere 1992) appeals to applied researchers for a number of good reasons:

- It presents respondents with a simple and theoretically appealing task
- By constraining responses, it prevents scale use bias, which makes it ideal for supporting segmentation and allowing cross-cultural comparisons (Cohen 2003)
- It produces sensitive and discriminating measures (Chrzan and Golovashkina 2006)
- Available commercial software makes design and analysis of maxdiff experiments easy (Sawtooth Software 2005).

If one uses the Sawtooth Software maxdiff software to make a design rather than a balanced incomplete block design, one must decide how many maxdiff items (attributes) to include in each choice question. Bryan Orme (2005) addressed this question with an analysis of synthetic data. He found that a substantial increase in predictive accuracy occurs in maxdiff experiments with 5 items/set rather than 3 items/set. A smaller improvement occurs when moving from 5 to 7 items/set, however.

We seek to augment Orme's findings with analysis of data from real respondents. For our empirical analyses we draw data from three commercial studies, in each of which we systematically manipulated the number of items per question across cells of otherwise similar respondents.

PLANNED COMPARISONS

Dropout Rates

We use χ^2 tests to assess whether respondents quit surveys at different rates when maxdiff questions have more or fewer items per set. Because the dropout rates we measured were for the whole survey, and not just for the maxdiff section, we expect this to be a weak test.

Task Length

Timers at the beginning and end of the maxdiff portion of each questionnaire allow us to measure the duration of the maxdiff questions. Using a median test (because some respondents pause the survey during the maxdiff section) we test for the effect of the number of items per set on task length. With a regression analysis we quantify the contribution of the number of maxdiff questions and the number of items per question on task length.

Positional Bias

Order or position effects may be more pronounced as the number of items per set increases or decreases, and we track these to check.

Parameter Equivalence

Using the Swait and Louviere (1993) procedure for separating the scale from substantive parameters, we test whether maxdiff questions with different numbers of items per set result in substantively different model parameters, or in models with different amounts of response error (scale).

Predictive Validity

Taking Elrod's (2001) advice against relying on hit rates to validate our maxdiff models, we supplement hit rate analysis with more meaningful out of sample validation tests.

EMPIRICAL STUDIES

Study 1

As part of a 16 minute interview about vacationers' travel preferences, 884 respondents evaluated the appeal of 17 activities in maxdiff questions. Each respondent completed 17 maxdiff questions, as follows:

- 220 had maxdiff questions containing 4 activity items
- 223 had maxdiff questions containing 5 activity items
- 220 had maxdiff questions containing 6 activity items
- 221 had maxdiff questions containing 7 activity items

For a validation holdout task we had respondents rank six of the 17 objects known from prior research to span the spectrum from low to high preference.

Study 2

Like Study 1, this study of 19 activities formed part of a larger study of vacation preferences. A total of 1,236 respondents divided randomly and evenly into three treatments:

- 19 questions, 3 items/question
- 19 questions, 5 items/question
- 19 questions, 8 items/question

Again respondents ranked 6 of the activities that spanned the preference range.

Study 3

This smaller study tested just 12 activities, again as part of a larger piece of travel research. A total of 904 respondents completed 12 maxdiff questions as follows:

- 302 had 3 items in each maxdiff question
- 300 had 5 items/maxdiff question
- 302 had 7 items/question

In this case the client cancelled the holdout question, so we split the sample in half and use the maxdiff model from one half to predict best and worst choices in the other half, and vice versa, as our holdout strategy.

RESULTS

Dropout Rate

In study 1, more items per question produced higher dropout rates, results strong enough for a marginally significant linear trend (p<.11). In Study 2, maxdiff questionnaires with 3 items per set had lower dropout rates than those with either 5 or 9 items per question. Study 3 however, had much lower dropout among respondents receiving 5 items per question than among those receiving 3 or 7.

	% Dropout		
# Items/set	Study 1	Study 2	Study 3
3	-	2.6	3.2
4	10.0	-	-
5	13.5	6.8	0.7
6	13.6	-	-
7	14.9	-	3.8
8	-	5.5	-

As expected the dropout rate differences only weakly point away from using larger numbers of items per question.

Task Length

In all three studies, highly significant (p<.001) χ^2 tests for median test time showed longer task length as the number of items per questions (and the number of questions) increased:

	Interview Length (seconds)		
# Items/set	Study 1	Study 2	Study 3
3	-	207	95
4	220	-	-
5	248	292	139
6	279	-	-
7	390	-	138
8	-	361.5	-

A handy regression equation summarizes the impact of number of questions and number of items per question on task length:

Length (sec) = 9.4(number of questions) + 17.5 number of items/ question)

Positional Bias

Very slight position effects seem to occur, but they do not differ based on the number of items per question. For example, in Study 3, a slightly greater number of bests seem to occur in the first position, and a slightly greater number of worsts seem to occur in the last two positions, but (a) the effect is minimal, and (b) it is similar regardless of the number of items per question:

Position	3 items		5 items		7 items	
	b	w	b	w	b	w
1	36%	33%	24%	16%	16%	13%
2	36	33	19	21	16	15
3	29	34	22	22	15	13
4	-	-	18	24	13	13
5	-	-	16	18	15	14
6	-	-	-	-	15	17
7	-	-	-	-	11	15

Parameter Equivalence

Earlier work assessing the number of questions to include in partial profile choice questions (Patterson and Chrzan 2003) found equivalent substantive parameters across partial profile questions of different sizes, but more response error (a smaller scale factor) for questions with more attributes per profile. In the case of maxdiff scaling, we expected to find this but we did not: In Study 1 each of the 6 pairs of models had significantly different substantive parameters. Because the Swait and Louviere (1993) procedure is sequential and allows a test of scale only if the test for substantive parameter differences is non-significant, this rendered us unable to test scale parameters. These scale effects were almost negligibly small even if we could not test them for significance. What at first appeared odd became the norm when the same outcome occurred in Studies 2 and 3.

A glance at the aggregate MNL parameters from Study 3, however, reveals that while the substantive parameter vectors may be significantly different, the practical differences among them are small, and would not affect decisions made on the basis of the research:

Item	3 items/set	5 items/set	7 items/set
1	.60	.73	.74
2	-.69	-1.01	-.92
3	.14	.29	.09
4	.87	.99	1.02
5	.04	-.12	-.16
6	-.30	-.30	-.21
7	.07	.08	.11
8	-.63	-.61	-.53
9	-.63	-1.03	-.99
10	.59	.63	.57
11	.53	.49	.54

Studies 1 and 2 also had similar, but statistically significantly different, parameter vectors.

Predictive Validity

Hit rates are a poor man's measure of predictive validity (Elrod 2001), but we report them for their familiarity. Neither prediction of first choice nor of full rank orders differ significantly in quality depending on the number of items per question in Study 1:

Items/set	First Choice %	Rank Order %
4	63.6%	49.9%
5	68.6	54.2
6	60.0	52.0
7	62.4	52.0

In Study 2, first choice hit rates improve with increasing numbers of items per question (χ^2 = 13.6, p < .01) but this pattern disappears for the prediction of the full rank order:

Items/set	First Choice %	Rank Order %
3	64.1	47.8
5	72.1	49.8
8	75.5	49.4

Root mean square error (RMSE) is a measure of how well predictions of shares match actual shares: higher RMSE is worse than smaller. Both for holdout samples of a rank order (Studies 1 and 2) and of best and worst choices (Study 3) RMSE shows no significant differences by number of items per question, though there is a hint that 3 items per question may produce slightly worse predictions than questions with more items:

Items/set	Study 1	Study 2	Study 3
3	-	4.1	13.1
4	3.1	-	-
5	3.8	2.9	12.2
6	5.5	-	-
7	4.7	-	11.1
8	-	4.7	-

DISCUSSION

No qualitative differences in results emerge from these analyses as a function of the number of items per question. Given Orme's (2005) findings about increasing accuracy with larger numbers of maxdiff questions, and our finding that task length decreases with number of items per question, we recommend a larger number of questions with smaller numbers of items per question. Given the slight evidence of poorer hit rates and poorer out-of-sample for 3 items per question we recommend using 4 or 5 items per question in maxdiff experiments. For example, in the same amount of time, respondents could rate 20 maxdiff questions with 4 items per question or 15 with 7 items per question. With mixed evidence about the number of items per set (except that smaller questions are shorter questions) and Orme's evidence of greater accuracy with more questions, the design having respondents complete 20 questions with 4 items per question makes more sense than the design with 15 questions and 7 items per question.

References

Chrzan, Keith and Natalia Golovashkina (2006) "An Empirical Test of Six Stated Importance Measures," International Journal of Market Research, in press.

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

Elrod, Terry (2001) "Recommendations for Validation of Choice Models," 2001 Sawtooth Software Conference Proceedings, Sawtooth Software, 225-43.

Finn, Adam and Jordan Louviere (1993) "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Environmental Economics and Management 29, 228-37.

Orme, Bryan (2005) "Accuracy of HB Estimation in MaxDiff Experiments," Sawtooth Software Research Paper, http://www.sawtoothsoftware.com/download/techpap/maxdacc.pdf

Patterson, Mike and Keith Chrzan (2003) "Partial Profile Discrete Choice: What's the Optimal Number of Attributes," 2003 Sawtooth Software Conference Proceedings, Sawtooth Software, 173-85.

Swait and Louviere (1993) "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," Journal of Marketing Research, 30, 305-14.

PRODUCT LINE OPTIMIZATION THROUGH MAXIMUM DIFFERENCE SCALING

KAREN BUROS
DATA DEVELOPMENT WORLDWIDE

ABSTRACT

Faced with the task of identifying an optimal line of products (flavors, SKU's, fragrances, etc.) from a large array of potential alternatives, researchers traditionally rate or sort the alternatives on liking or purchase interest scales. Then, utilizing top box or top two box ratings, the researcher will attempt to identify the line with broadest appeal across the sample (TURF approach).

Unfortunately, it is difficult for consumers to express meaningful differentiation using Likert scales across a large number of alternatives. This leads to many line optimization problems….

In popular item categories, adding additional items to a line of 4 or 5 items adds only marginally to "reach" of the line. Using additional criteria—such as depth of interest in the line—will often produce too many alternatives that could "round out" a line of 10 to 15 items. All of this makes it difficult to identify the optimal total size of the product line.

This paper illustrates how Maximum Difference Scaling produces greater discrimination across items by using choices from partial item sets, avoiding the problems of scaling. The results additionally offer greater flexibility in the line optimization criteria…

- Optimization on "reach" – like at least one item in the line
- Optimization on "depth" – the number of items liked in the line
- Optimization on "most preferred" item – whether or not the line includes the "favorite" item
- Optimization on "share of requirements" – what proportion of total potential volume of use is tapped by the line of items.

The data presented illustrate the approach for optimization within a brand or brand family, irrespective of competition, and when optimizing versus competitive product lines.

Using blinded data from actual studies, the paper outlines the steps undertaken for an optimization study using Maximum Difference Scaling—questionnaire requirements, utility estimation using HBA, optimization alternatives and user-friendly simulation options.

LINE OPTIMIZATION APPLICATIONS

Line Optimization covers a broad range of common research problems. Typically, one thinks of consumer package goods applications—the number of flavors or fragrances in a product line, for example. However, line optimization approaches apply to a wide range of product and service businesses. In essence, whenever the need arises to create a limited line offering or a bundled product or service, a line optimization approach can prove an efficient way of tackling the problem. Some examples of Line Optimization applications are:

- The number and type of kitchen gadgets or tools on a display board
- SKU"s on a store shelf
- Drinks in a cooler
- Rewards in a frequent flyer program
- Desserts on a display cart
- Main courses on a menu
- Prizes in a game
- Services on a mobile phone
- Games on a website

Experience across a broad range of product and service categories has led to the conclusion that Maximum Difference Scaling produces more actionable results than ratings on Likert scales.

REVIEW OF TURF USING LIKERT SCALES

Research issues, such as those described above, have typically been handled by asking the respondent to rate "interest in" or "likelihood of purchase" for a series of 20 plus alternatives. The analysis proceeds to determine the number of people who give a "top box" or "top two box" score to each item. Then, for a set number of items, the analyst will determine the number of people falling into a "top box" or "top two box" category for at least one item in the line. This is the "Reach" component of TURF (Total Unduplicated Reach and Frequency).

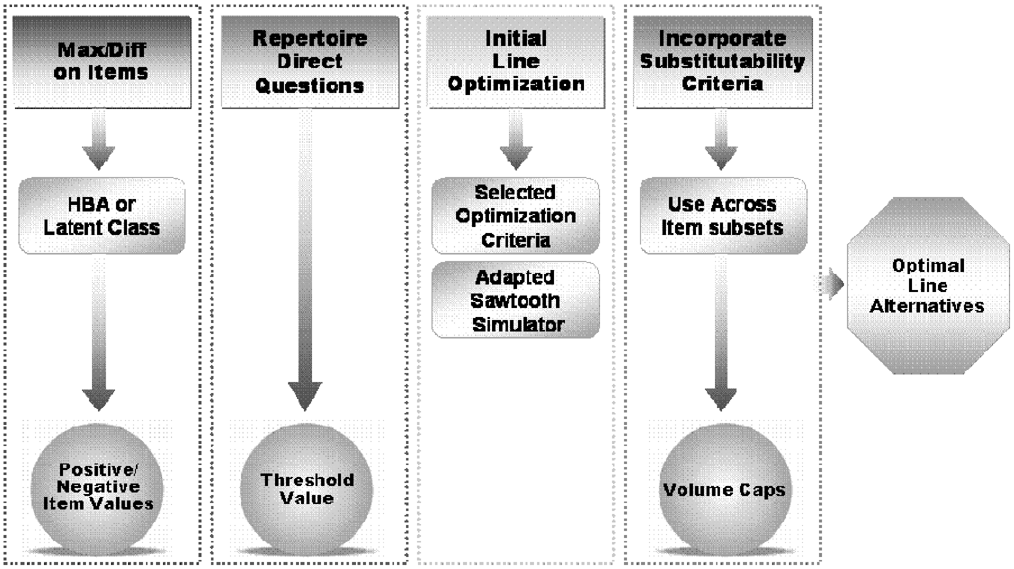
At this point two problems often occur. In many product categories, use of "Top Box" and "Top Two Box" interest is high. Near total "Reach" can often be achieved with limited number items. Thus, achieving an optimal line configuration becomes problematic. In other instances, the scale does not differentiate sufficiently across items to identify an optimal line. One is left to wonder whether the approach is identifying an optimal line or the preferences of "high raters."

In the first instance, the analyst can then proceed to the second stage of "Frequency" referenced in the TURF algorithm—the number of items rated "Top Box" or "Top Two Box" across the highest "Reach" alternatives. The concept here is simplistic—if the respondent likes more than one item in the line, he/she will buy more often and a number of items from the line. This does not address the issue of the frequency of choosing each item.

INCORPORATING MAXDIFF INTO THE TURF ANALYSIS

The following diagram provides an overview of the steps and issues involved when incorporating the power of Maximum Difference Scaling into TURF analysis resulting in greater precision and flexibility determining the optimal product line.

The Process of Line Optimization


Max/Diff on Items
HBA or Latent Class
Positive/ Negative Item Values
Repertoire Direct Questions
Threshold Value
Initial Line Optimization
Selected Optimization Criteria
Adapted Sawtooth Simulator
Incorporate Substitutability Criteria
Use Across Item subsets
Volume Caps
Optimal Line Alternatives

MAXIMUM DIFFERENCE QUESTIONS

This paper will not attempt to serve as a full review of the Maximum Difference Scaling approach which has been covered extensively in previous Sawtooth Software conferences and technical papers. In this application, the MaxDiff approach replaces the Likert scale ratings. Rather than ask "how likely" the respondent is to purchase/use each item, the respondent is asked to choose the most/least likely to buy/use from groups of 5 items.

In this approach, the guideline of representing each item at least three times in the design is used. The design is generated using Sawtooth Software's MaxDiff Designer and the results are calculated using Sawtooth Software's CBC/HB calculation program with specified MaxDiff settings.

The HBA calculation produces respondent-level values which are positively/negatively scaled. However, positive and negative values are not reflective of buy/not buy scaling, only of rank and magnitude of preference.

ITEM REPERTOIRE

It becomes necessary to identify the threshold value for each respondent within the MaxDiff values that represents the point at which an item should not be included in the repertoire of "consideration." This is accomplished using a direct question appropriate to the category. In beverages, for example, one might ask respondents which beverages would be essential to have or which beverages they would never consider drinking.

This can be used to establish a respondent-specific "cut-off" point for re-scaling the MaxDiff values in such a way that the non-considered items are zeroed-out and the remaining items retain magnitude of relative preference.

OPTIMIZATION: MULTIPLE CRITERIA

As stated earlier, traditional TURF analysis focuses primarily on "**Reach**" —the number of respondents who would find at least one item in the line "appealing." Incorporating MaxDiff scaling allows a number of additional options:

Variety: The number of different items in a line

Favorite: The number of respondents who find their most preferred item in the line

Total Requirements: Estimate share of all preferences satisfied by the line

These four measures, as appropriate to the problem, can be incorporated into either Sawtooth Software's Advanced Simulation module (with a customized simulator modification) or built into an Excel- based simulator.

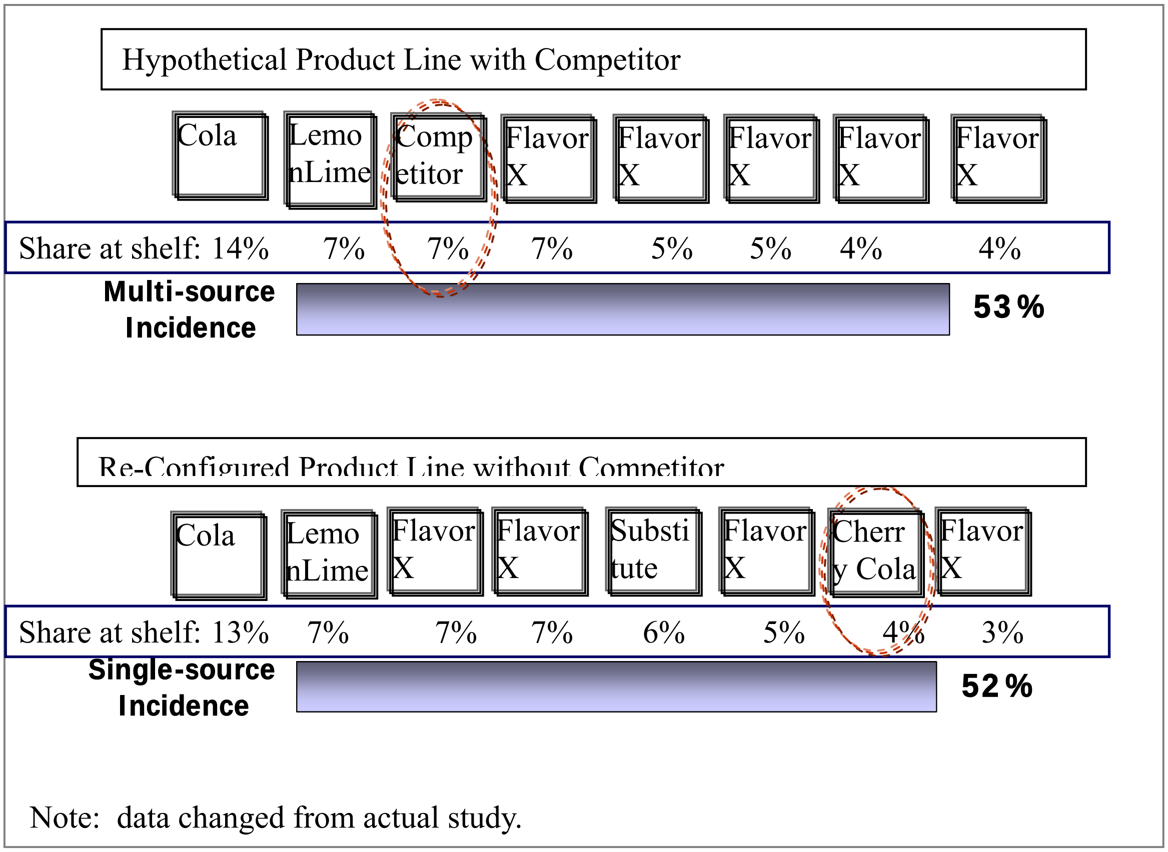
With either option, the measures for total item lines of varying sizes can be readily accommodated. The Sawtooth Advanced Simulation module, with use of a genetic algorithm optimization approach, can be used to generate the optimal product mix. The Excel-based option is a point-and-click option for on-the-fly simulations. Since the Excel version takes minimal training to use, end-users of the information can readily simulate the effect of changes in their product lines.

Case Study #1:
Optimization of Product Line within a Competitive Context

In this study, the client company's "shelf space" was threatened by a popular competitor. Specifically, retailers were trying to remove shelf space to make room for the competitor. The company wanted to demonstrate that an "optimal" line of their products would do as well or better than their smaller line with a competitor.

As shown in the figure below, while the competitor's one item did well, the company's product line could be re-configured to perform as well overall without the competitor—53% "share of requirements" satisfied with the competitor and 52% "share of requirements" satisfied without the competitor. In this example, near total reach was attained with a line of four items so a more sales-oriented metric was more relevant.

Note also that each item fulfills a "share of requirements." The values generated from the Hierarchical Bayesian analysis are expressive of the respondent's degree of preference for the item. Rescaling these values across the items remaining in the respondent's set after removal of the non-considered items provides a "share of requirements" for each item proportionate to preference.



Hypothetical Product Line with Competitor

Share at shelf: 14% 7% 7% 7% 5% 5% 4% 4%

Re-Configured Product Line without Competitor

Share at shelf: 13% 7% 7% 7% 6% 5% 4% 3%

Note: data changed from actual study.

Is "Share of Requirements" Realistic?

The following chart compares the "Share of Requirements" data from this study to sales data in the category.

Although the sample and beverages in the study do not exactly replicate the universe from which the sales data are drawn, the data line up to a fair degree.

	Simulator Share	Category Share		Simulator Share	Category Share
Cola A	19%	19%	Other	3%	1%
Cola B	14%	18%	Other	2%	1%
Lemon-Lime A	10%	16%	Other	6%	1%
Non- Cola B	14%	9%	Other	0%	1%
Competitor	7%	8%	Other	2%	1%
Other	5%	7%	Other	0%	1%
Other	3%	6%	Other	1%	1%
Other	0%	3%	Other	0%	1%
Other	3%	2%	Other	2%	1%
Other	2%	1%	Other	2%	1%
Other	2%	1%	Other	1%	1%
Other	1%	1%	Other	1%	1%

Case Study #2: Product Line with Substitution

Multiple item use within a line can be attributed to different consumer motivations.

- The consumer may be indifferent between two items. For example, the consumer may find both Brand A and Brand B to be equally acceptable. If one is not available, they are not unhappy to buy the other.
- The consumer may select one item to fill a need/desire on a certain occasion. The consumer may "like" another item as well as the first, but not for that particular occasion. For example, a consumer may "like" both hot tea and hot coffee equally, consuming them in equal proportions. He/she may have a cup of coffee in the morning and a cup of tea in the afternoon. When the customer wants coffee, he/she does not want tea.

One can argue that these two situations must be modeled differently in development of a product line of coffees and teas. In the first instance, any item can substitute for another in fulfilling reach, depth and share of requirements optimization criteria. In the second instance, coffees can substitute for each other, as can teas, but not one for the other.

Thus, when removing a favorite item from the line, one should determine whether the second or third favored item is truly a substitute.

ALTERNATIVE APPROACHES TO SUBSTITUTION

Two possible approaches to constraining substitution show feasibility. The first is *a priori* and the second is consumer-generated. The case illustrated here assumes an *a priori* constraint by separating items of different types into separate exercises. The consumer-generated approach is created by allowing the respondent to sort items into groups within which the items are substitutable from their point of view.

In both approaches a volumetric measure of the respondent's use of each grouping is needed to place a constraint on the number of items of that type within the line.

Ideally, one would also have a proximity measure of items within the group to determine the degree to which each is substitutable for other items in the group.

MARKETING ISSUE

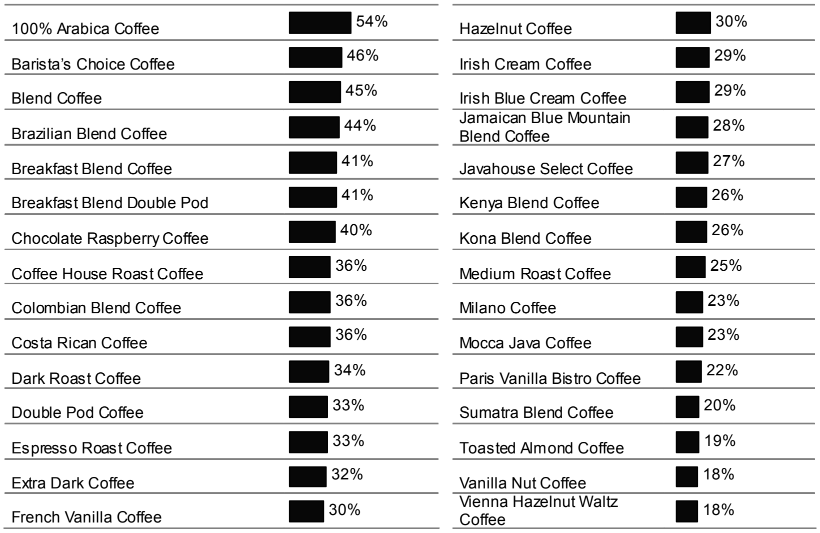
The company desires an optimal menu of multiple flavors of coffees, teas and other hot beverages. There are three issues at hand:

1. The optimal overall size of the menu line
2. The number of items of each type—coffee, tea and so forth
3. Within each type, the mix of items

In this instance, the company was comfortable using the pre-designated types as *a priori* groupings. Thus, the approach was to optimize within each type, determine the number of items allocated to each type based on volume consumption (of each respondent) of each type and finally to determine the impact of increasing the size of the overall line based on fulfillment of "share of requirements."

SHARE OF REQUIREMENTS SATISFIED BY EACH ITEM

Analytically, the first step is to examine the relative preference for each potential menu offering. As illustrated below, establishing a threshold value below which the respondent is not interested in the item, rescaling the HBA "utilities" for each respondent and aggregating preferences across respondents results in a "share" for each potential item on the menu.



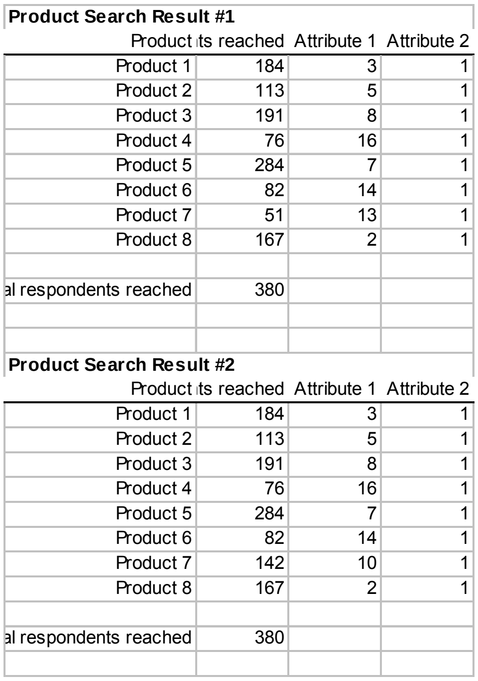
(Note that results have been altered.)

Using the customized Sawtooth Advanced Simulation module, the optimal combination of five core flavors for the menu produces increases over the current menu offer:

- Reach: increased from 82% to 91%
- Share of Requirements satisfied: increased from 17% to 21%
- Favorite Coffee on menu: increased from 21% to 31%

Optimization is an iterative process, particularly so when adding additional flavors to a product line. To address the key question of the number of flavors that should be offered on the menu, the product line is gradually increased from 5 to 10 flavors.

In this example from the optimization simulations, two alternative menu lines will "reach" an identical number of respondents (380).



Product Search Result #1

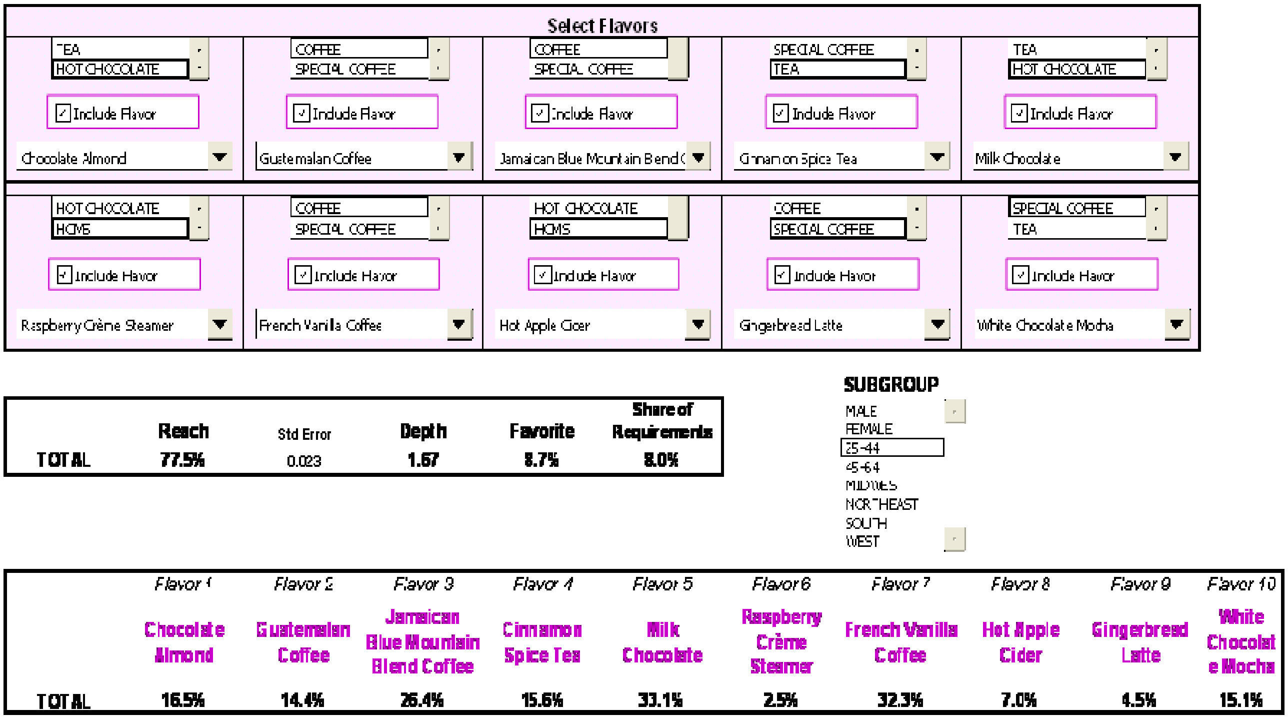
Product	its reached	Attribute 1	Attribute 2
Product 1	184	3	1
Product 2	113	5	1
Product 3	191	8	1
Product 4	76	16	1
Product 5	284	7	1
Product 6	82	14	1
Product 7	51	13	1
Product 8	167	2	1
al respondents reached	380		

Product Search Result #2

Product	its reached	Attribute 1	Attribute 2
Product 1	184	3	1
Product 2	113	5	1
Product 3	191	8	1
Product 4	76	16	1
Product 5	284	7	1
Product 6	82	14	1
Product 7	142	10	1
Product 8	167	2	1
al respondents reached	380		

In the first simulation, the flavor (13) in position 7 will reach only 51 respondents, while in the second simulation flavor 10 is substituted with a reach of 142 respondents. Thus, the number of flavors "liked" on the menu is increased.

To allow interactive use of the information, the data are further taken into an interactive Excel-based simulator, such as the one shown below:



Select Flavors

TEA HOT CHOCOLATE	COFFEE SPECIAL COFFEE	COFFEE SPECIAL COFFEE	SPECIAL COFFEE TEA	TEA HOT CHOCOLATE
☑ Include Flavor	☑ Include Flavor	☑ Include Flavor	☑ Include Flavor	☑ Include Flavor
Chocolate Almond	Guatemalan Coffee	Jamaican Blue Mountain Blend C	Cinnamon Spice Tea	Milk Chocolate
HOT CHOCOLATE HOMS	COFFEE SPECIAL COFFEE	HOT CHOCOLATE HOMS	COFFEE SPECIAL COFFEE	SPECIAL COFFEE TEA
☑ Include Flavor	☑ Include Flavor	☑ Include Flavor	☑ Include Flavor	☑ Include Flavor
Raspberry Crème Steamer	French Vanilla Coffee	Hot Apple Cider	Gingerbread Latte	White Chocolate Mocha

	Reach	Std Error	Depth	Favorite	Share of Requirements
TOTAL	77.5%	0.023	1.67	8.7%	8.0%

SUBGROUP

MALE
FEMALE
25-44
45-64
MIDWES
NORTHEAST
SOUTH
WEST

	Flavor 1	Flavor 2	Flavor 3	Flavor 4	Flavor 5	Flavor 6	Flavor 7	Flavor 8	Flavor 9	Flavor 10
	Chocolate Almond	Guatemalan Coffee	Jamaican Blue Mountain Blend Coffee	Cinnamon Spice Tea	Milk Chocolate	Raspberry Crème Steamer	French Vanilla Coffee	Hot Apple Cider	Gingerbread Latte	White Chocolat e Mocha
TOTAL	16.5%	14.4%	26.4%	15.6%	33.1%	2.5%	32.3%	7.0%	4.5%	15.1%

This allows the analyst to determine the reach, depth of menu offering, inclusion of the "favorite" flavor and share of requirements met across the different product types—in this example, coffees, teas and so forth.

SUMMARY

This paper presents the case that incorporation of the Maximum Difference Scaling approach into TURF analysis enables the researcher to broaden the approach to bring insight to a wide range of marketing problems. In its broadest terms, Maximum Difference Scaling can be viewed as a one attribute Choice model. Individual-level results obtained through HBA allow the measurement of not only "reach" of line of offerings, but also depth of the line, inclusion of "favorite" items and the degree to which the line satisfies the requirements of the individual.

It is particularly applicable to situations in which customers seek variety. This extension of TURF is applicable beyond "fast moving consumer package goods," the traditional application for the approach. It can be used in the formations of service offerings, such as rewards programs, and retail issues, such as product displays.

These case studies, as well as similar studies, have shown this approach to be quite useful in optimizing product lines. However, as the issues of product line development become more complex, we are seeking ways to incorporate additional purchase decision dynamics into the approach. TURF analysis traditionally focuses on "reach" as a primary criterion. This paper argues that, not only should criteria be broadened, but issues such as substitutability should be incorporated into the analytic framework.

COMMENT ON BUROS

BRYAN ORME
SAWTOOTH SOFTWARE

Karen has done some nice work here. She's demonstrated some practical uses for TURF-based optimization routines and shown some attractive spreadsheet-based what-if simulators. There is little to criticize, but in the spirit of improving research I'd like to point out some weaknesses of TURF analysis and also question whether the standard MaxDiff measurement technique should be preferred for this type of research. In place of standard MaxDiff results used within TURF analysis, I'd suggest Adaptive-MaxDiff.

DRAWBACKS OF TURF

In a 1999 article in *Marketing Research* magazine by Conklin and Lipovetsky (Conklin and Lipovetsky, 1999), the authors highlighted a few drawbacks of TURF analysis:

- The top many TURF solutions are often statistically indistinguishable,
- TURF "tends to add products to a line that have unique appeal to a small group of customers,"
- If the competitive landscape changes, the TURF solution may no longer be very optimal,
- Techniques which more heavily weight items with overall broader appeal will perform better in the face of uncertain competitive moves.

The first point (many near-optimal solutions) is both good news and bad news. It is good news in the sense that a client can have a variety of nearly equally-preferred options to choose from based on his or her expert opinion, knowledge of marketing/development costs, etc. The bad news is that a client may become frustrated that there aren't statistically significant differences among the top n candidates.

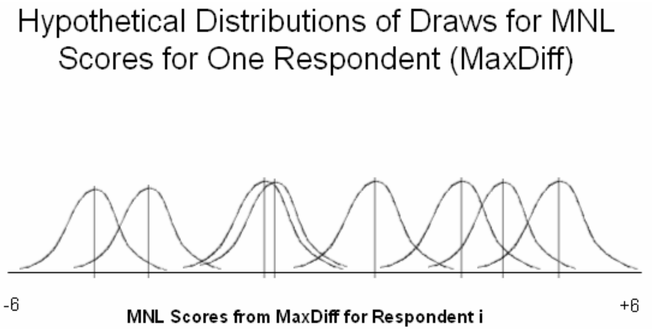
Conklin and Lipovetsky argued for a related search technique called Shapley Value to overcome these issues. It's not my intention to further elaborate on this here, and the interested reader may refer to the article.

ADAPTIVE-MAXDIFF QUESTIONING

The last point I'd like to make stems from some recent research we've conducted (but not yet published) on Adaptive-MaxDiff questioning.

Traditional MaxDiff focuses on obtaining precise estimates for both very important (or preferred) and very unimportant items. Each item appears approximately an equal number of times in the experiment, and the respondent takes time within each set to report which item is best and which is worst.

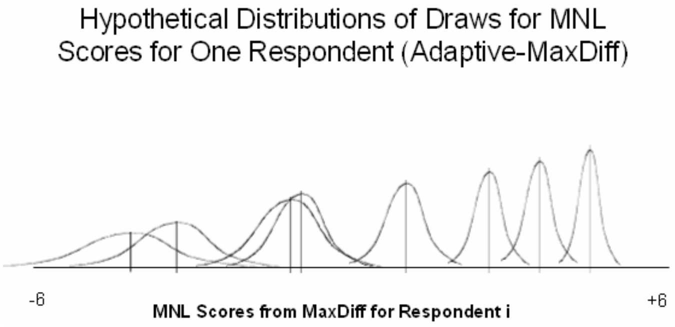
Consider a hypothetical respondent's item scores resulting from a MaxDiff questionnaire, with HB draws around each item's point estimate of importance. The widths of the distributions characterize the precision of the estimates. For standard MaxDiff analysis, the distributions of draws may look something like this:



Each item is estimated with roughly equal precision. But, if the goal of our modeling is to leverage or identify the items of most importance (such as with TURF), are both the design strategy and the "Worst" half of the MaxDiff question really compatible with this goal? After all, only the few most important items for each respondent affect TURF solutions. Why not ask each respondent to trade off the items of most importance to him/her relatively more times (to obtain higher precision for these items) and those of least importance fewer times? This suggests an Adaptive-MaxDiff Strategy.

With an adaptive MaxDiff strategy, we might begin by showing, say, 5 or 6 items per set. Then, we discard any items from further consideration that are marked "worst." The subsequent sets show fewer items per set (only the "surviving" items) in progressive stages (four items, then three items, etc.). In the final stage, the best few items are shown in pairs (MPC) until an overall winner is identified. Estimation follows as with traditional MaxDiff, for example using HB. A clever aspect of this strategy is that adaptive questionnaires lead to greater utility balance in later sets, which has been demonstrated to increase the precision of utility estimates for choice studies estimated under MNL (Huber and Zwerina, 1996). And, as the amount of utility balance increases, the complexity of the task (number of items per set) is reduced, to try to avoid increased respondent error and fatigue.

With Adaptive-MaxDiff, the distribution of draws for a hypothetical respondent may appear more like:



Precision is higher for items of most importance and lower for items of least importance. With Adaptive-MaxDiff, we're more likely to identify correctly the most important item(s) for this respondent, which is critical to TURF.

We have already conducted a split-sample methodological study comparing Adaptive-Maxdiff to standard MaxDiff. Our results were favorable for the adaptive form, and we'll present the results in the upcoming joint Sawtooth Software/SKIM conference in Munich, Germany later this year.

REFERENCES:

Conklin and Lipovetsky (1999), "A Winning Tool for CPG," *Marketing Research*, Winter 1999/Spring 2000.

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, (August), 303-317.

AGENT-BASED SIMULATION FOR IMPROVED DECISION-MAKING

DAVID G. BAKKEN
HARRIS INTERACTIVE

ABSTRACT

Agent-based modeling and simulation is applied to the problem of forecasting adoption for new consumer durables. This approach extends the typical marketing forecast based on measures of purchase intent or conjoint-based share predictions to incorporate dynamic processes of consideration set formation and manufacturer reactions to the level of demand.

INTRODUCTION

The emerging discipline of *complexity science* offers new tools for improving our understanding of and ability to predict complex phenomena. One of the more fascinating developments to come out of this field is agent-based simulation. A central tenet of agent-based models is that complex *macro level* behavior arises from the behavior of individuals who follow relatively simple rules of action. Another core idea is that small changes in initial conditions can lead to large changes in the macro level pattern.

Traditional marketing research can provide retrospective insight into the impact of marketing actions on buyer behavior, but our interview-based methods give us a rather static picture. Because our traditional methods are not well-suited for capturing or revealing the effects of interactions between buyers, marketers, and their competitors, at best we have an incomplete view of the marketplace; in some cases the view may be misleading.

Most of our survey-based methods rely on aggregate statistical analysis and modeling. The mean, median, and top two box percentages are widely used measures in quantitative marketing research. Multiple regression analysis and variants, such as logistic regression, are the workhorses of predictive modeling, but these techniques most often are limited by reliance on aggregate rather than individual level analysis. New methods of estimation, including latent class and hierarchical Bayesian models, are changing this, but these models still provide a static rather than moving picture of a market.

At the other end of the research spectrum, qualitative interviewing methods provide rich information about differences in consumer preferences, perceptions and behavior, but these methods also do not enable us to predict the behavior of a community of consumers interacting in real time with other consumers, their environment, and competing offers.

These limitations have important consequences for decision-making. While decision-making is a somewhat complex process involving several different activities, an essential element of *good* decision-making is the ability to consider the likelihood of each of the possible outcomes of each of the alternative courses of action. Even simple decisions involving a small number of variables can generate a large number of potential actions and outcomes. Agent-based simulation offers a method for estimating the likelihood of many possible outcomes to many possible decisions.

OVERVIEW OF AGENT-BASED SIMULATION

Most marketers and marketing researchers will be familiar with some form of simulation as applied to marketing. Consider the simple "What if?" scenario analysis applied to the results from a conjoint analysis, or more elaborate "war games" in which competing teams of managers (or MBA students) develop and implement strategies in an interactive fashion. Many of us were introduced relatively early in life to a simulated real estate market in the form of the Monopoly® board game.

As a concept, simulation has almost as many meanings as applications, but a central, underlying principle is the representation of one process or set of processes with another, usually simpler in some useful sense, process or set of processes. The Monopoly® board game, for example, simplifies a rather complex economic system into a limited set of transactions. Part of the fun of playing the game derives from the degree to which the outcomes (wealth accumulation and bankruptcy) resemble the outcomes from the real world process that is being simulated.

While agent-based simulations differ in significant ways from other approaches to simulation, there are some similarities; therefore it is worth briefly reviewing those approaches. *Systems dynamics* models treat the process under consideration as a series of *aggregate* flows or state changes. Large systems of differential or difference equations are used to plot trajectories of variables over time. For example, we might model factory output as a system where consumers flow into the system, demand products at some price and the factory makes the product and accumulates inventory, shipping products out to customers. At any point in time, all of these variables have some value (e.g., number of customers, price they are willing to pay, current inventory), at the next point in time the aggregate values of those variables will change. The basis for simulating change lies in the functional relationship between the variables. For example, $x_{t+1} = f(x_t; \boldsymbol{\Theta})$ indicates that the value of x at time $t+1$ is a function of the value of x and time t and some parameter $\boldsymbol{\Theta}$.

Discrete event models are one step closer to agent-based simulation. Discrete event models are often used to simulate the flow of *customer agents* through some service system. A source generates new customers who join the queue at a server. After being served, they exit the system. Such models are used to determine the impact of changing the number of servers, splitting or aggregating process steps, and so forth. An important distinction is that the *agents* in this case (which might be customers or servers) are not autonomous, and make no decisions.

Cellular models (also known as cellular automata) are very simple agent-based models. Agents occupy a two (or more) dimensional space and are connected in a regular lattice. Cellular models are used to simulate communication processes between agents. In a cellular model, each agent's *current state* is determined by its state at the previous point in time and the state of its immediate neighbors. Many cellular models employ a *von Neumann* neighborhood, which consists of the four neighboring cells to the north, east, south or west of the target cell, but it is possible to use the *Moore neighborhood* of all eight contiguous cells.

The defining characteristic of an *agent-based* simulation is the autonomous decision-making of the agents. Systems dynamics and discrete event models are governed by an overall structure and process that determines the state of individual agents or subunits in the model. In contrast, agents in an agent-based simulation are equipped with sensing mechanisms and decision rules that govern their behavior in response to changes in their environment. Agents are characterized by at least four essential attributes: *autonomy*, *asynchrony*, *interaction* and *bounded rationality*. *Adaptation*, or the ability to learn, is another characteristic that may be essential in many agent-based models for marketing.

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

An Agent-based Simulation of Adoption for a New Automobile Model

Forecasting sales for new consumer durable goods such as automobiles has proven to be one of the biggest challenges for market research. Bayus, Hong and Labe (1989) cite a few reasons for this difficulty, including greater fluctuations in consumer spending on durables, great variability in the timing of purchases, and the fact that markets for durables are not static. Most models for forecasting demand for new durable products are based on the Bass (1969) model. In the Bass model, sales are a function of both innovation (e.g., "early adopters") and imitation. The model is simple and elegant but requires a number of assumptions. These assumptions create the opportunity for applying agent-based simulation to the problem of forecasting sales for a new consumer durable[1]. For example, the Bass model assumes that market potential remains constant over time, that the diffusion of one innovation is independent of all other innovations, that marketing actions do not affect the diffusion process, and that there are no supply restrictions. The Bass model is an aggregate model. Variable factors that drive individual consumer choices (such as individual preferences, awareness, and ability to obtain the product) are ignored.

Bass developed this model before conjoint methods became widely available. We can now measure, with considerable reliability, the decision-making processes of individual consumers. Consider the introduction of a new car model. Most automakers conduct "stage-gate" research throughout the development of a new model, refining the concept and resulting product in response to feedback from consumers. We might, for example, have data from a conjoint study that will allow us to estimate the relative utility of the new model for each survey respondent. We can use this as the starting point for an agent based simulation of the adoption of a new model or *name plate* (the newly introduced Ford Fusion is an example of a nameplate).

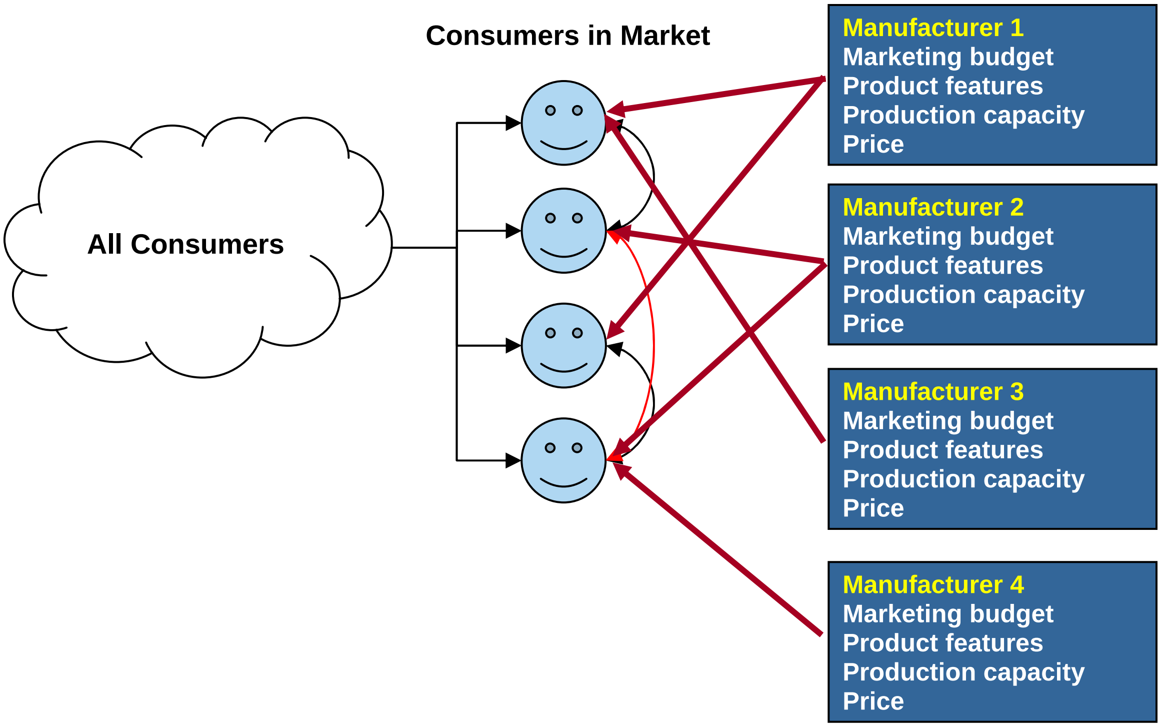
Implementation of the Agent-Based Simulation

The first step in implementing the agent-based simulation is specification of the process to be modeled. **Figure 1** presents a schematic representation of this process for the new vehicle. Consumers enter the market at some point, where they are confronted with marketing communications from various manufacturers as well as information from other consumers. There are a number of *subprocesses* that need to be captured. These include processes for (among consumers) becoming aware of marketplace alternatives, formation of consideration sets,

[1] I should note that a number of researchers have developed modifications of the original Bass model based on addressing one or more of these assumptions.

and choice of a vehicle from among those in the consideration set. For manufacturers, subprocesses include setting prices to maximize profit.

Figure 1. Overview of the Marketing Process



The model consists of two types of agents, **buyers** and **sellers**.

Buyers are assumed to have a fixed set of preferences for brands, features and price, represented by the "part-worths" or "utilities" estimated from a discrete model. The probability of choosing a particular vehicle, conditional on being in the market and conditional on the vehicle being in the buyer's consideration set, is a function of the ratio of the total utility of the specific vehicle to the utilities of other vehicles in the buyer's consideration set. At a decision point, the buyer "purchases" (via a first choice method) the vehicle in the consideration set with the highest utility, as long as that utility is greater than a "threshold" reflected in the part-worth estimate for "none of the above." This subprocess should be familiar to most readers as the mechanism by which static preference shares are predicted in conjoint market simulators.

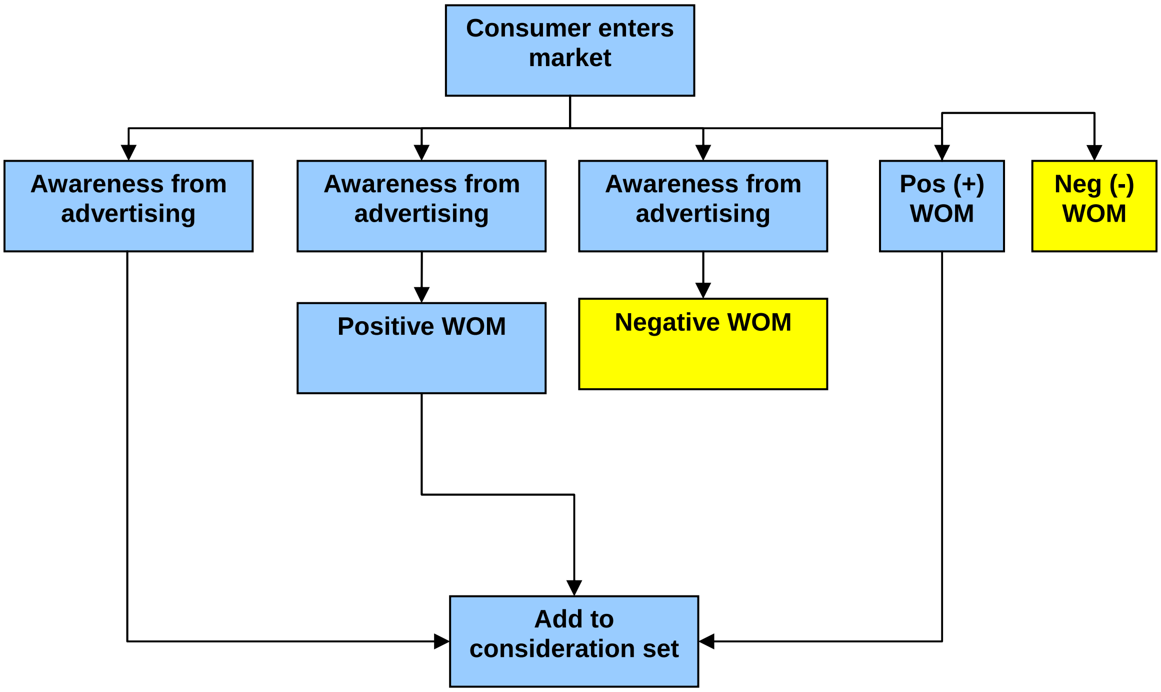
Sellers are analogous to manufacturers. The vehicles offered for sale are fixed in terms of feature configuration; the features do not change over the course of the simulation. For each vehicle, the manufacturer establishes a production rate, or the number of new vehicles that become part of dealer inventory in each time step. If inventory accumulates too rapidly, sellers can adjust the rate downward. If there is more demand than supply, sellers can increase the production rate. In addition, sellers can vary the price for a vehicle, raising price when demand is high and lowering price when demand is low.

Consideration set formation

Each buyer must form a consideration set of vehicles in order to make a purchase. A vehicle can enter the consideration set if the buyer becomes aware of the vehicle from one of two sources: advertising by the seller or word-of-mouth from another buyer who already has purchased the vehicle. These flows are diagrammed in **Figure 2**.

Agent-based simulations rely on stochastic processes to assign values to variables that, in the real world, are typically unobserved and that can be assumed to be random at the agent level. For *existing* vehicles, the probability of awareness from advertising for any one buyer is a random variable set to be equal to measured awareness from advertising. For the new vehicle, probability of awareness from advertising is a function of the advertising level (expressed as number of exposures per time period) which, in turn, is a function of the manufacturer's available resources.

Figure 2. Consideration Set Formation



The process for becoming aware through word of mouth is somewhat different, and relies on a **small world** network of connections between buyers. Small world networks combine elements of "regular" and "random" networks in a way that minimizes the number of connections required to move information from one buyer to any other buyer in the market. To implement the small world network we create a two-dimensional array in which each element (cell) of the array represents an individual buyer. **Figure 3** illustrates the difference between the regular and random connections. Any buyer can communicate with any other buyer through a chain of immediate neighbor connections, but the random connections directly connect buyers who are not immediate neighbors. **Figure 4** presents a "snapshot" of a word of mouth simulation in progress.

Figure 3. Regular and Random Networks Equal "Small World"

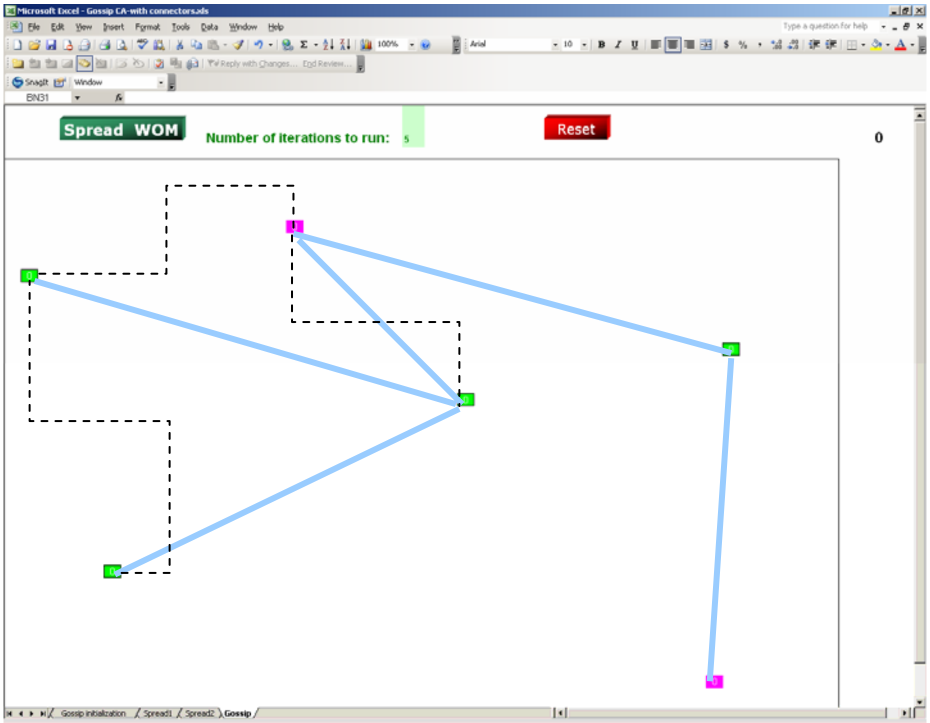
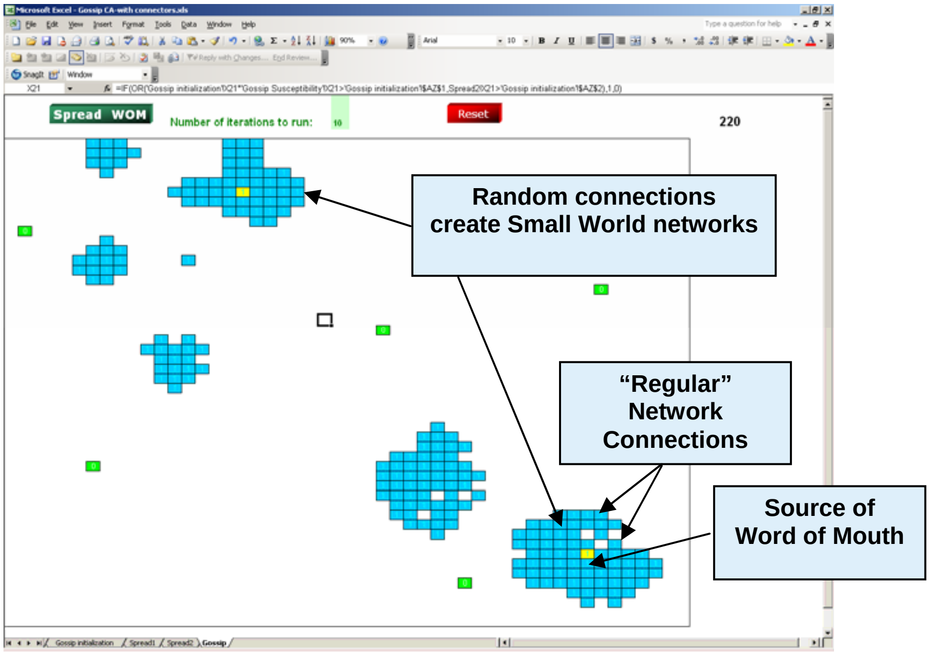

Microsoft Excel - Gossip CA-with connectors.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
Arial
10
BN31
Spread WOM
Number of iterations to run:
5
Reset
0
Gossip initialization
Spread1
Spread2
Gossip

Figure 4. Word of Mouth Simulation


Microsoft Excel - Gossip CA-with connectors.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
Arial
10
X21
Spread WOM
Number of iterations to run:
10
Reset
220
Random connections create Small World networks
"Regular" Network Connections
Source of Word of Mouth
Gossip initialization
Spread1
Spread2
Gossip

=IF(OR('Gossip initialization'!X21*'Gossip Susceptibility'!X21>'Gossip initialization'!$AZ$1,Spread2!X21>'Gossip initialization'!$AZ$2),1,0)

At initialization, all buyers are considered potential *receivers* of word of mouth. Once a buyer agent makes a purchase, it becomes a potential *sender* of word of mouth. Buyers are characterized by *susceptibility* to word of mouth as well as propensity to spread word of mouth.

These are random variables from a uniform probability distribution (min=0, max=1) and may change from one time step to the next, reflecting real world day to day variability in an individual's propensity or ability to send or receive word-of-mouth.

Senders can transmit either positive or negative word of mouth. The likelihood of negative word of mouth is a function of "things gone wrong," which is a measure of product quality in the automotive industry.[2]

A buyer may "forget" about a brand if some (variable) number of time steps pass without a new exposure to advertising or word-of-mouth. For each buyer agent, a random variable specifies the number of time steps required to forget about the vehicle.

In summary, a vehicle enters a buyer's consideration set if either advertising or word of mouth makes the buyer aware of the vehicle. However, *negative* word of mouth, generated as a result of quality problems, can prevent a vehicle from entering the buyer's consideration set.

Seller Agents

The "seller" agents (reflecting manufacturers/dealers) make offers to those buyers who include the seller's vehicle in their consideration sets. For this simulation, we assume that each vehicle in the simulation represents a unique seller (that is, there are no sellers who are trying to maximize an outcome across multiple vehicles). Whereas buyers seek to maximize their utility from a single purchase decision, sellers attempt to maximize a profit function that requires that they match supply (inventory) to demand. The key metric for this is "days of supply" or the expected number of days of inventory, given current supply, production rate, and purchase rate. If vehicles sell faster than the production rate, revenue may be lost due to the fact that some buyer's are willing to pay more than the asking price. If vehicles sell more slowly than the production rate, profit suffers as inventory accumulates and must be financed. Sellers have two mechanisms for managing demand—pricing and advertising. Advertising impacts the number of buyers who include a vehicle in their consideration set, while pricing impacts the number of such buyers who will prefer the vehicle to all other alternatives.

At each time step, the seller evaluates the current inventory against an *ideal* level, which is a user input specified at initialization. If the inventory departs from the ideal by a specified amount, the seller can change price, change advertising levels, or both.

Choice Simulator

Awareness from advertising or word of mouth determines if a particular vehicle is included in the choice simulation for a given respondent.

For each respondent, a preference share is calculated for *all* vehicles in the market scenario, whether in the consideration set or not. If a vehicle is not in the consideration set, the utility for that vehicle is multiplied by **0**. Preference share or probability of purchase is then calculated for the vehicles in the consideration set. The alternative with the highest preference share (including "none") is considered the first choice, and is purchased by the respondent. When this happens, an indicator for the vehicle make is set to 1, and this buyer drops out of the choice simulator. However, the buyer remains in the market for purposes of spreading word of mouth.

[2] Not all details of the simulation are provided, for simplification. For example each agent follows "rules"—a set of conditional relationships—that determines if net word of mouth is positive or negative.

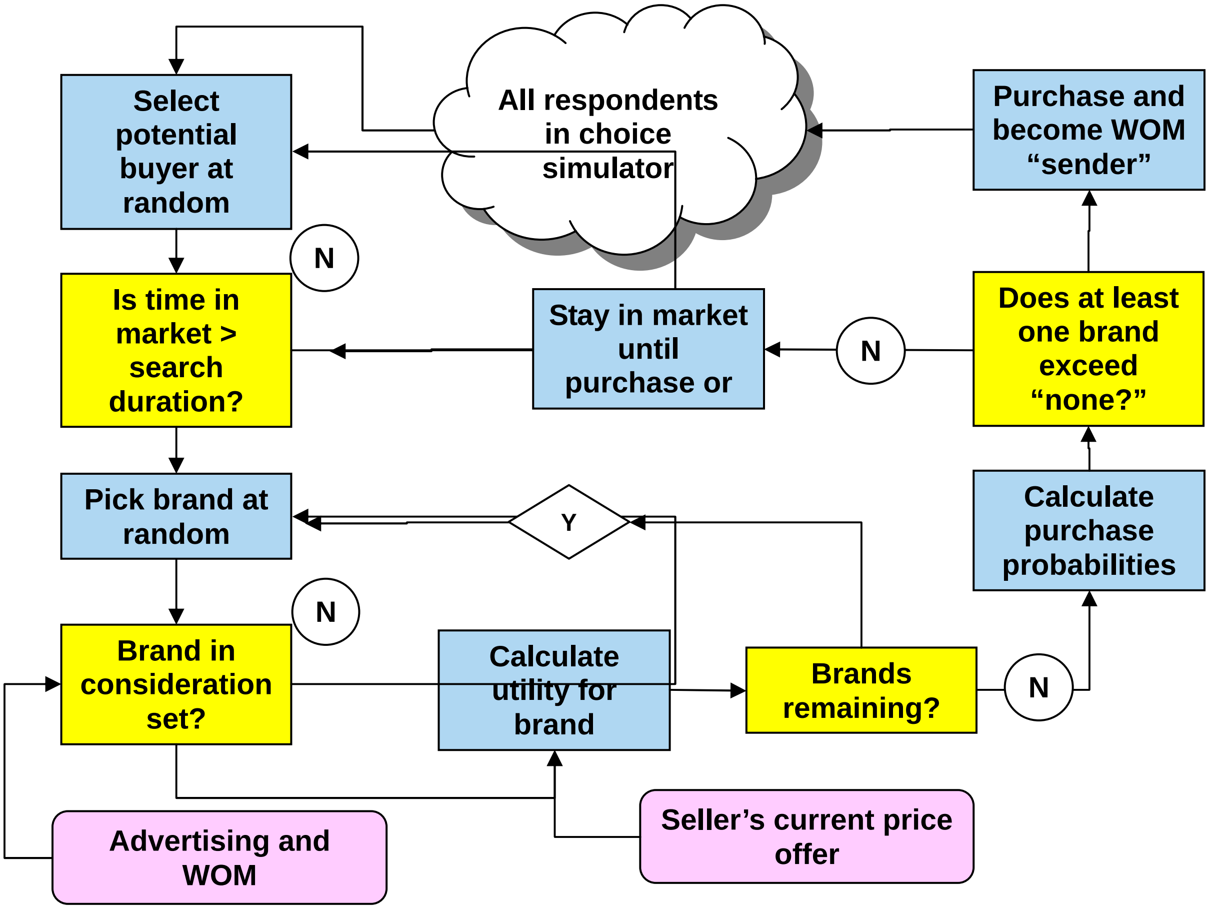
For each vehicle, a starting price point is set at initialization. The price is set for a time step, and all buyers at that time step are offered the same price (note—this is only one of many possible ways of simulating price negotiation or adjustment). Sellers can change prices between time steps.

The choice simulator incorporates three indicator variables. For each buyer, a variable will indicate whether or not that buyer is in the market. If the value is **0** (not in the market), all market share calculations for that buyer are "zeroed out." A second variable for each vehicle indicates if it is in the buyer's consideration set. Finally, there is a variable to indicate if a vehicle has been purchased by a buyer agent.

General overview of the simulation process

Figure 5 presents a schematic representation of the overall simulation. A subset of respondents in the choice simulator are selected at random to be "in the market."

Figure 5. Overview of Consumer Search and Decision Process


Select potential buyer at random
All respondents in choice simulator
Purchase and become WOM "sender"
N
Is time in market > search duration?
Stay in market until purchase or
N
Does at least one brand exceed "none?"
Pick brand at random
Y
Calculate purchase probabilities
N
Brand in consideration set?
Calculate utility for brand
Brands remaining?
N
Seller's current price offer
Advertising and WOM

Prior to running the simulation, several variables must be initialized, either by user input or random number generation. Some of these variables maintain a fixed value over the course of the simulation. For buyers these include:

Average network size

Maximum number of periods a buyer stays in market without purchasing

Number of periods without ad exposure until forgetting occurs

Initial seller fixed (constant) variables include:

Ideal inventory (days of supply)

Marketing resources

Advertising productivity (number of exposures to create awareness)

Things gone wrong[3]

Random variables are generated by one of two methods. For variables that reflect a number of events, such as number of ad exposures in a time step, a random number is drawn from a Poisson distribution with lambda equal to the expected value or average across all units of observation. So, for example, if a seller allocates 5 units of marketing resources, and each unit is expected to generate 1.5 exposures per buyer per time step (the productivity of advertising), the value for lambda would be 7.5. A few of these random variables are set at initialization and remain constant for the run, but many are updated at each time step.

For variables used to establish thresholds for action, such as propensity to spread word of mouth, a random value is drawn from a uniform probability distribution (with range between 0 and 1). The result may be used either as a weight or multiplier (number of ad exposures times susceptibility) or as a conditional value. For example, in the case of initial awareness for *existing* vehicles, each case gets a random value between 0 and 1. Separately, the user sets the expected awareness from advertising. Assume that this value is 40%. Then, every case where the random variable is less than or equal to 0.4 will get a value of 1 in the awareness indicator for that vehicle.

The general flow of the simulation within a time step follows:

1. Select a buyer at random.
2. Select a vehicle at random.
3. Set number of ad exposures for this buyer in this time step.
4. Add number of exposures in this time step to number of exposures in all previous time steps.
5. If no exposures to advertising in this time step, check to see how many time steps since last exposure.
6. If number of time steps since last exposure is equal to number of time steps to forget, set awareness from advertising to 0.

[3] Quality can, of course, improve over time. For simplicity, we assume that quality was fixed for the forecast time period.

7. If current plus cumulative number of exposures multiplied by susceptibility to advertising is > 1, set awareness from advertising to 1.
8. Scan network members (network for each buyer is generated at initialization) and sum word of mouth values for all network members. If absolute value of this sum multiplied by susceptibility to word of mouth is >= 1, set awareness from word of mouth to +1 if sum of word of mouth is > 1, or to -1 if < -1(note—if 0, set to 0).
9. If sum of awareness from advertising and word or mouth > 0, set consideration set to 1 (included).
10. Repeat for next vehicle.
11. Repeat for next buyer.

There are other subprocesses to determine whether a buyer of a given vehicle will spread positive or negative word of mouth, whether a seller will adjust price, and how much the price will be adjusted.

This simulation was programmed in Microsoft® Excel. MS Excel is an object-oriented programming language. Object-oriented languages are well suited for agent-based simulation, although other types of languages have been used. The spreadsheet format allows us to accumulate summary data for each time step for each individual agent in the simulation. **Figure 6** shows data for one respondent for one vehicle. Some of the variables represent external or environmental factors, such as the number of ad exposures in a period and cumulative exposures. Others are "instance" variables that are either stable or varying traits of the individual buyer. For example, "TSF" ("Time steps to forgetting") is the instance variable that determines the number of consecutive periods without exposure to advertising that will lead to "forgetting." This introduces an advertising decay effect. "Susc" is an instance variable that changes from one time step to the next and indicates the buyer's susceptibility to advertising in that time step. Finally, "aware" is a summary indicator variable that tells us whether the buyer is aware of the vehicle in a given time step. Other instance variables include the number of periods a buyer will stay in the market without finding an acceptable alternative and, for those who have purchased, propensity to spread word of mouth.

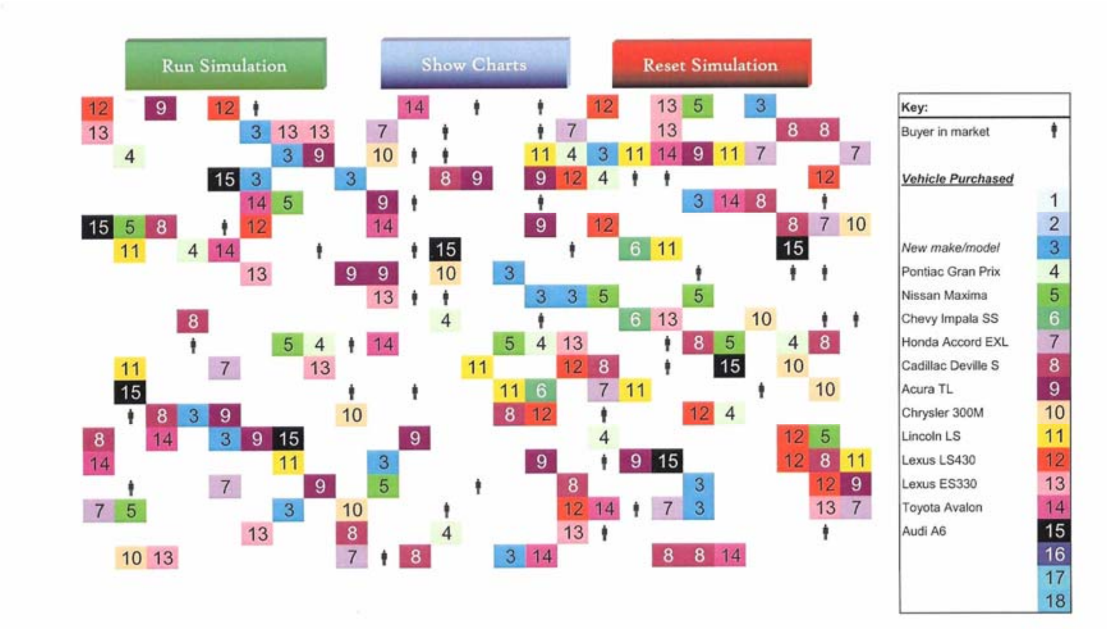
Figure 6. Data for One Respondent, One Vehicle

Time step	1	2	3	4	5	6	7	8	9
#Ad exp	0	2	3	0	0	0	2	1	0
Cum #exp	0	2	5	5	5	0	2	3	3
TSF	3	3	3	3	3	3	3	3	3
Susc.	.11	.26	.33	.78	.51	.62	.13	.89	.45
Aware	No	No	Yes	Yes	Yes	No	No	Yes	Yes

Running the Simulation

One of the benefits of simulation is the ability to visualize the changes in key variables over time. In particular, visualization makes the effects of changing initial conditions apparent. For purposes of illustration, in **Figure 7** we show a spatial representation of vehicle choice that reveals the impact of word of mouth. In a number of instances, immediate or near neighbors purchased the same brand; 65% of the buyers of the new vehicle are no more than one cell away from another buyer (and others may be connected via small world networks). This form of visualization could be used to estimate the impact of varying levels of quality ("things gone wrong") on adoption rate, as an example.

Figure 7. Spatial Display of Vehicle Adoption

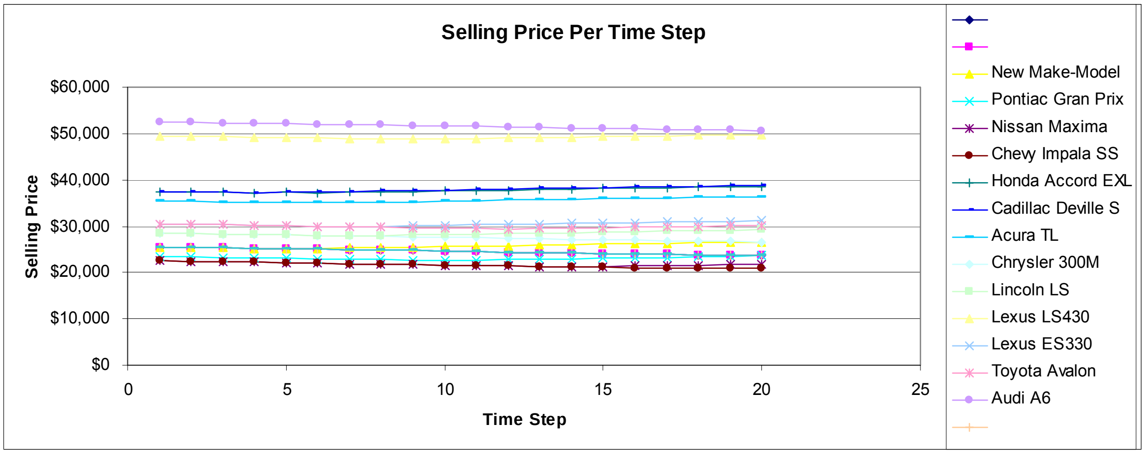


Key:	
Buyer in market	🯅
Vehicle Purchased	
	1
	2
New make/model	3
Pontiac Gran Prix	4
Nissan Maxima	5
Chevy Impala SS	6
Honda Accord EXL	7
Cadillac Deville S	8
Acura TL	9
Chrysler 300M	10
Lincoln LS	11
Lexus LS430	12
Lexus ES330	13
Toyota Avalon	14
Audi A6	15
	16
	17
	18

Plotting summary measures over time is also informative. **Figure 8** shows the results of "price negotiation" over the course of a simulation as manufacturer agents adjust price in order to meet their inventory targets. It appears that the new vehicle was priced slightly below what consumers are willing to pay, and the seller was able to increase price slightly over the course of the simulation.[4]

[4] All simulations depend on initial conditions. For simplification, all sellers were given the same inventory targets. Assigning different targets would result in different patterns of price adjustments.

Figure 8. Vehicle Price by Time Step



Programming Options for Agent-Based Simulations

Several “toolkits” for developing and running agent-based simulations are available. Most of these are based on object-oriented languages such as Java or Objective-C and contain subroutines and pre-programmed modules that simplify the task of creating an agent-based simulation. Such toolkits usually include routines for displaying summary results and for visualizing the simulation in progress.

One such toolkit is Swarm (www.swarm.org) which was developed at the Santa Fe Institute. The Swarm toolkit utilizes Objective C, so some familiarity with this language is a prerequisite for using Swarm.

NetLogo (http://ccl.northwestern.edu/netlogo/) is a Java-based toolkit that is derived from StarLogo, a toolkit based on the Sugarscape simulation described by Epstein and Axtel (1996). Agents populate a two-dimensional space, similar to the word-of-mouth component in the simulation presented in this paper. NetLogo includes a library with a wide variety of models, so that users often can find examples of code to use in building a new simulation.

Discussion

While agent-based simulation has been applied to developing an understanding of emergent behavior for a wide variety of human and non-human phenomena (predator-prey relationships, spread of viral infections, foraging behavior in ants, traffic congestion, trade, combat and so forth), there appear to be few direct applications in marketing research to date. Adoption of innovative products and services is one area where agent-based simulations may add value to marketing research. This application has two aspects that favor agent-based simulation. First, the adoption process unfolds *over time*. Second, interaction among consumers and between manufacturers (in the form of marketing actions, price changes, word of mouth, and so forth), affect the rate of adoption.

We often know a great deal from market research about the *marginal* distributions of the variables we measure in our surveys. We know far less about the *joint* distributions of many variables considered at the same time. For example, we might know how much consumers like a particular brand, how important quality is, and whether or not they tend to share their brand or product experiences with others, but we would be hard pressed to use that knowledge to predict the impact of differences in initial product quality on the spread of negative word of mouth and the subsequent impact on first year sales for a new product. Rather than try to estimate these probabilities using an aggregate flow model, we can create a population of agents that reflect the marginal distributions (in terms of the distribution of intrinsic and acquired characteristics), let them interact with each other, and observe the outcome. We can systematically vary the incidence of quality problems to determine the point at which poor quality will undermine the launch effort.

Choice simulators based on individual-level utility estimates are an excellent starting point for applying agent-based simulation for marketing research. Choice simulators already contain two key agent populations—buyers and sellers. We need only to specify the rules for interaction between the agents over time to create a dynamic agent-based choice simulator.

When a client asks a market researcher to "predict first year sales for this product," she usually expects a single value (perhaps with some confidence interval around the estimate). Agent- based simulation offers a different type of prediction—one where we identify the potential variability in outcomes as a function of different starting conditions and different rules for action. Agent-based simulations can simplify the process of identifying several different possible futures and, through repeated simulations, give us an idea of the likelihood of each of those different futures. This would seem to be a better basis for making decisions than single value predictions—which is somewhat like placing a bet on a poker hand by looking at only one of the five cards.

REFERENCES AND ADDITIONAL RESOURCES FOR AGENT-BASED SIMULATION

References

Bass, F. (1969. "A new product growth model for consumer durables," Management Science, 15 (January), pp. 215-227.

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

Epstein, J. and Axtell, R. (1996), Growing artificial societies: social sciences from the bottom up, Cambridge, MIT Press.

Resources

Axelrod, R. (1997), The complexity of cooperation: Agent-based models of competition and collaboration, Princeton: Princeton University Press.

Kennedy, J. and Eberhart, R. (2001), Swarm intelligence, New York: Academic Press.

Santa Fe Institute, www.santafe.edu

Center for the Study of Complex Systems, www.cscs.umich.edu

HOW MANY CHOICE TASKS SHOULD WE ASK?

MARCO HOOGERBRUGGE AND KEES VAN DER WAGT
SKIM ANALYTICAL

ABSTRACT

Through a meta analysis of many CBC studies, the authors seek to recommend an optimal number of choice tasks after which there is no substantial improvement in model fit or substantial changes to respondents' answers. Regardless the type of study and the complexity of the design, it appears that there is virtually no gain in asking more than 10-15 choice tasks.

INTRODUCTION

At the Advanced Research Techniques (ART) Forum of 2005, there was a lot of attention focused on Hierarchical Bayes techniques, whether stand-alone or in combination with Choice-Based Conjoint. Academics were trying to improve the performance of HB models by integrating more variables into the model, e.g. to include the time that a respondent takes to evaluate the choice task in the model. Unfortunately the model extensions that were presented were not as successful as had been hoped.

During subsequent discussions, it appeared that academics generally take a low number of choice tasks (e.g. 6) in a Choice-Based Conjoint interview as a given fact and concentrate their efforts therefore more on possibilities for model improvements. However, for practitioners, the number of choice tasks in an interview is one of the variables that they can influence. And it may be the case that setting a larger number of choice tasks in an interview can improve the performance of the CBC/HB model to a large extent.

Thus, the idea for this paper was born—what would be the effect of increasing the number of choice tasks on the "performance" of HB utilities?

COMPARISON WITH EARLIER WORK STUDYING THE NUMBER OF CHOICE TASKS

The most influential paper so far on the subject of the number of choice tasks in CBC is "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" by Richard Johnson and Bryan Orme, presented at the ART Forum of 1996. Their frame of reference for CBC analyses at that point in time was aggregate logit analysis. One of the main conclusions of their paper was that respondents seem able to complete more choice tasks than the authors previously thought was wise to do, and that doubling the number of choice tasks provided about as much increase in the precision of aggregate parameters as doubling the number of respondents.

Currently, we generally no longer use aggregate logit analyses and, instead, we rely on individual utility values generated by HB. Or maybe we should speak of 'quasi-individual' utility values since these utility values are based on the individual's choices only insofar it is possible to derive them from the individual's choice data. The remainder of the content of the individual's utility values is filled up based on aggregate data.

In this new situation, the question becomes: should we perhaps *minimize* the number of choice tasks per interview? In the case that we do not have a sufficient amount of choice data per individual, it does not make sense to rely on individual utility values in the first place.

One of the outcomes of the Johnson/Orme paper was that (some) consumers change their choice pattern in the course of the interview. The other outcome was that (some) respondents become *less* inconsistent after the first few choice tasks and that consistency rates remain constant thereafter (no sign of respondent fatigue).

We have tried to tackle these problems by choosing a benchmark (holdout task) late in the interview. That would solve the second problem and, depending on the severity of the first problem, lead to a more conservative, higher number of choice tasks required rather than a lower number.[1]

Method and sample size

This paper contains two main sections:

1. The first section describes a more sophisticated method that has been applied to (only) two real commercial studies. The application to one of the two studies will be described here in detail.
2. The second section describes a more concise method that in contrast has been applied to a huge amount of real commercial studies (35). The two studies in the first section are also included in these 35.

The first method: criterion

A critical issue when determining the minimum number of choice tasks is what criterion to use to establish this number.

First, we needed to have a benchmark. For that purpose, we held one choice task out of the utility estimations. With the utility estimation that we had in a certain stage, we tried to match the prediction based on the utility values with the actual respondent's choice. This is a well-known procedure. However, we applied this procedure partly in a less common way in two respects:

1. Since many of the commercial studies did not contain a fixed choice task, it meant we had to use a variable choice task as a holdout task. We do not regard this as a methodological problem, since the main issue is that the choice task is being held out of the utility estimation. The main difference from using fixed choice tasks was that it was more work to process the matching of prediction with real choice.
2. The more important difference is that a bad match between forecast and actual choice is less relevant for the purpose of our paper. As we are trying to make a forecast, we have a total error between forecast and reality that consists of two sources that we need to distinguish carefully:

[1] In the one example with multiple holdout tasks we have spread the holdout tasks across the interview, except we did not have a holdout task in the earliest stage of the interview.

- the estimation error of the forecast (i.e. error caused by perhaps too few choice tasks or a too high variance of the respondent);
- the random error of the forecast (i.e. error caused in the holdout task where the respondent happens to be less in line with all the other choice tasks; this may happen merely as a result of presumed probabilistic consumer behavior).

It should be emphasized that we are only interested in the estimation error of the forecast (which depends on the number of choice tasks), not in the random error (which is independent of the number of choice tasks).

Though we have not finalized the discussion about the criterion yet, we will, in between, show an example of our approach. In one commercial study we had 16 choice tasks in the interview, each with 4 concepts. Our procedure was as follows:

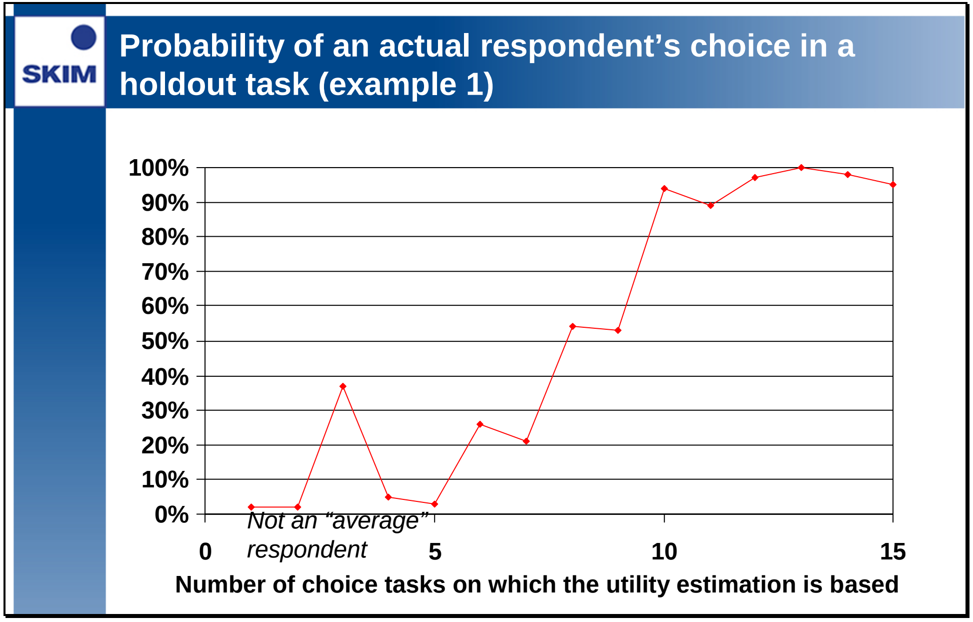
- We calculated HB utilities based on only the 1st choice task[2], and used those utilities to make a prediction for the 13th (holdout) task.
- We calculated HB utilities based on the 1st and 2nd choice task, and used those utilities to make a prediction for the 13th (holdout) task.
- We calculated HB utilities based on the 1st to 3rd choice task, and used those utilities to make a prediction for the 13th (holdout) task.…
- until: We calculated HB utilities based on choice tasks 1-12 and 14-16, and used those utilities to make a prediction for the 13th (holdout) task.

Thus, we had 15 predictions in total, based on an increasing number of choice tasks used for the utility estimation.

Please note that the prediction for the holdout task actually consisted of a probability distribution for each of the concepts. We matched the prediction with the respondent's actual choice by studying the predicted probability of the concept that has actually been chosen by the respondent. Figure 1 demonstrates this for one example respondent.

[2] Even though estimating HB models with a single task is not considered an appropriate use of HB (there is no ability to capture heterogeneity in our models using a single task), we did that here to show the increased benefit of asking tasks beyond the first task. With a single task, the individual-level parameters reflect just random normal draws from population parameters.

Figure 1



SKIM

Probability of an actual respondent's choice in a holdout task (example 1)

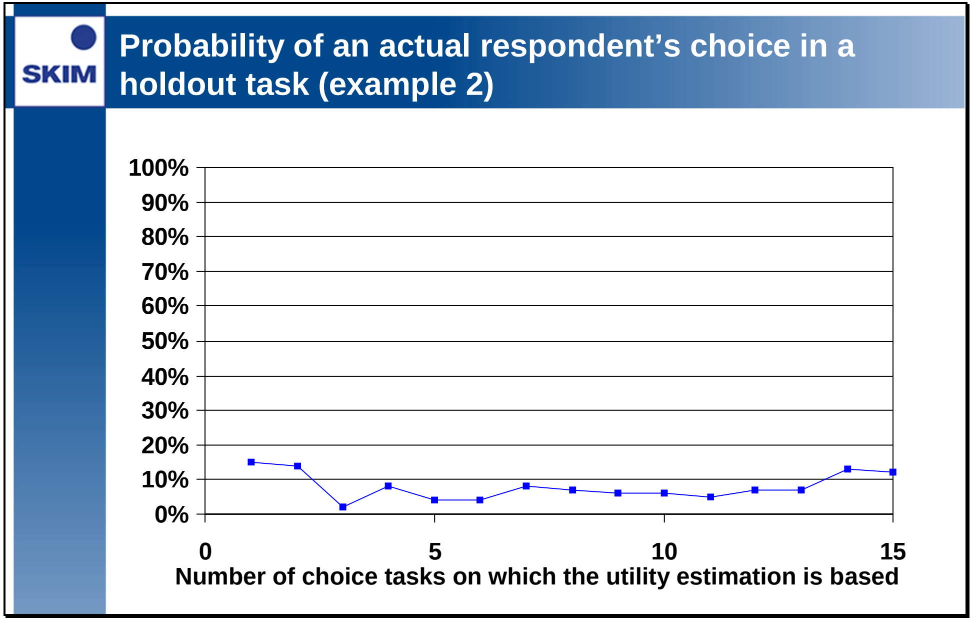
The interpretation is as follows:

- If we base the utility values on at least 10 tasks, the predicted probability of the concept that has actually been chosen is very high (more than 90%). The prediction is very good, but that is partly a matter of luck. After all, a probability of 90% means at the same time that there is 10% chance that the respondent would have done something completely different.
- If we would have asked only very few choice tasks, say, choice tasks 1-5, the predicted probability of the concept that has actually been chosen is very low (less than 5%). Since HB utilities based on so few choice tasks depend heavily on aggregate choice data, it means apparently that we are dealing with a very atypical respondent for whom—as an individual—the aggregate choice data are not very meaningful. As a result, we would need a substantial number of choice tasks from this individual in order to get the forecast right.

In short, the amount of random error is very low, and the amount of estimation error is (apparently) decreasing while adding more choice tasks to the utility estimation. The estimation error keeps decreasing until the 10th choice task when there is hardly any total error left.

Figure 2 shows the results of another respondent. Here we see that the predicted probability of the concept that has actually been chosen is very low for this second respondent, regardless of how many choice tasks are included in the utility estimation. Hence, the amount of random error is very high, and (therefore) we cannot say anything meaningful about the amount of estimation error.

Figure 2



Probability of an actual respondent's choice in a holdout task (example 2)

Finally we will return to the discussion of the decision criterion for the number of choice tasks. Here is our view:

- If the probability of the chosen concept remains stable, while increasing the number of choice tasks on which the probability estimation is based, we assume that there is no more reduction possible in estimation error.
- We consider the number of choice tasks after which the probability remains stable as "sufficient".

Applying this to the first respondent, we would argue that asking 10 tasks is sufficient for a proper utility estimation of this respondent. Applying it to the second respondent, we would argue that we could live with just one choice task from him or her. A bad result (in terms of forecast) leads in this case, surprisingly enough, to a very low requirement of number of choice tasks.

There are two main objections against this criterion:

1. We cannot *really* determine if such an amount of choice tasks is enough because the *true* number of choice tasks required may fall "out of range". For example, with the 2nd respondent, the probability of the chosen concept might only start increasing after 20 or 30 choice tasks. We cannot check that with this approach because the original study did not contain more than 16 choice tasks.
2. We used only one holdout task. It is then not sufficient to say that you need a number of choice tasks for a certain respondent. We should rather say that we need a number of choice tasks for a certain respondent in order to forecast one certain holdout task. It would probably require more choice tasks in order to forecast multiple holdout tasks. In the example for the second respondent, should we have chosen a different holdout task, it would have been likely that we would have concluded on a larger number of choice tasks than one.

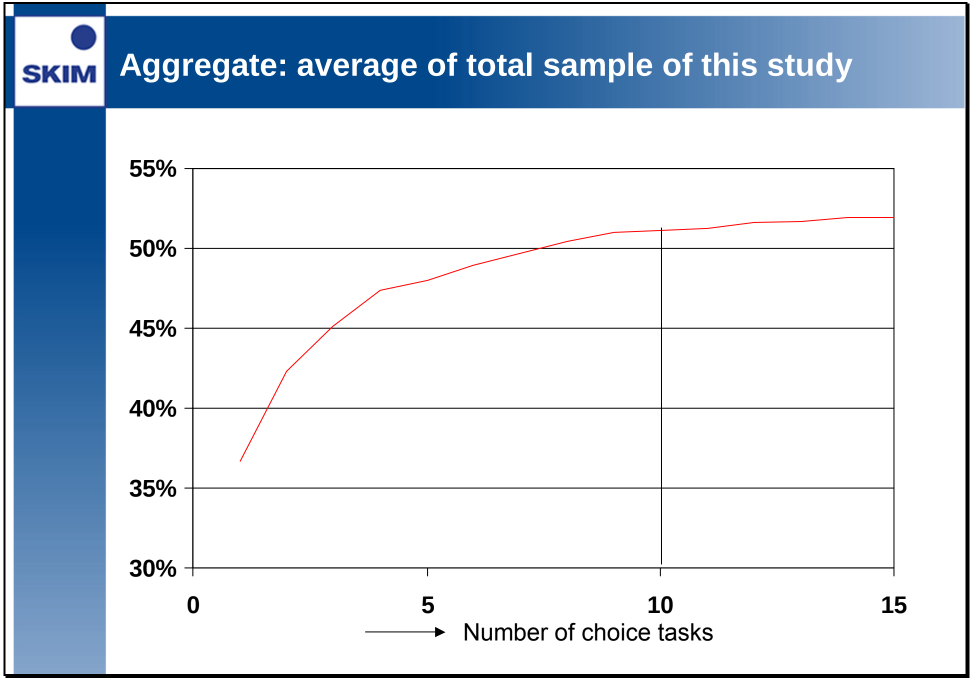
We will come back to both objections later, when discussing the results.

THE FIRST METHOD: RESULTS

In the previous section we have shown graphs for two individual respondents with the forecasted probability versus the number of the choice tasks in the utility estimation.

Figure 3 shows the average probability of all respondents of this study together.

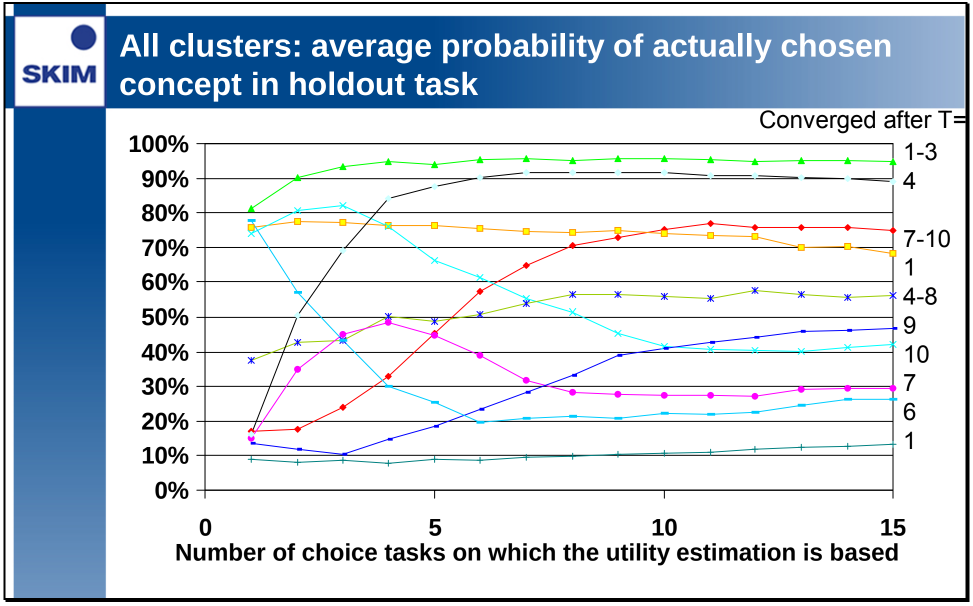
Figure 3



It is clear from the picture that, on the aggregate level, there is hardly anything to gain when increasing the number of choice tasks beyond 10.

We can also show results on a slightly more disaggregate level. To this end, we have conducted a cluster analysis on the series of forecasted probabilities while increasing the number of choice tasks in the utility estimation. Figure 4 shows the cluster means of a 10-cluster solution.

Figure 4



For example, the top line in the graph depicts a cluster of respondents (13% of the sample) for whom the prediction of the actual choice was very good from the very beginning. Increasing the number of choice tasks from 1 to 3 leads to a slight increase in probability to near 100%, so thereafter there is clearly nothing to gain in increasing the number of choice tasks.

The bottom line in the graph shows a cluster of respondents (23% of the sample) for whom the actual choice could never be predicted, regardless of how many choice tasks were included. The probability is around 10%, which is even worse than random (because we had 4 concepts, a random choice would have a probability of 25%). These respondents apparently answered something in the holdout task that was in contradiction with all the other choice tasks. They were not necessarily bad respondents, but probably we just had bad luck with their reply on the holdout task.

For both the top and the bottom clusters, increasing the number of choice tasks hardly has any effect. There are several clusters in the middle of the graph, however, for whom an increase in number of choice tasks is worthwhile. Roughly we see two types of clusters, depicted by increasing and decreasing curves respectively:

- Clusters of respondents that clearly benefit from increasing the number of choice tasks, in terms of an increasing probability. All these clusters converge after or within 10 choice tasks though, which means there is nothing to gain for them if we would increase the number of choice tasks from 10 to 15.
- Clusters of respondents for whom the probability is initially high, but—when more choice tasks are being added—it appears that the initial high probability was an overestimation, perhaps based on a coincidental similar "mistake" in an early choice task and the holdout task. In the later choice tasks the picture is being revised and convergence takes place after or within 10 choice tasks.

The conclusion of this cluster analysis is, while on an aggregate level 10 choice tasks would be necessary, for a substantial number of respondents we could have sufficed with a far lower number of choice tasks. This cluster analysis is apparently a good add-on to the aggregate results as in Figure 3. In the aggregate results the results of increasing and decreasing curves might have been cancelled out against each other and we might have drawn a too optimistic conclusion on the aggregate level. Fortunately (but perhaps accidentally) this did not really happen, since both at the aggregate and the cluster level we see that an increase beyond 10 choice tasks is not worthwhile.

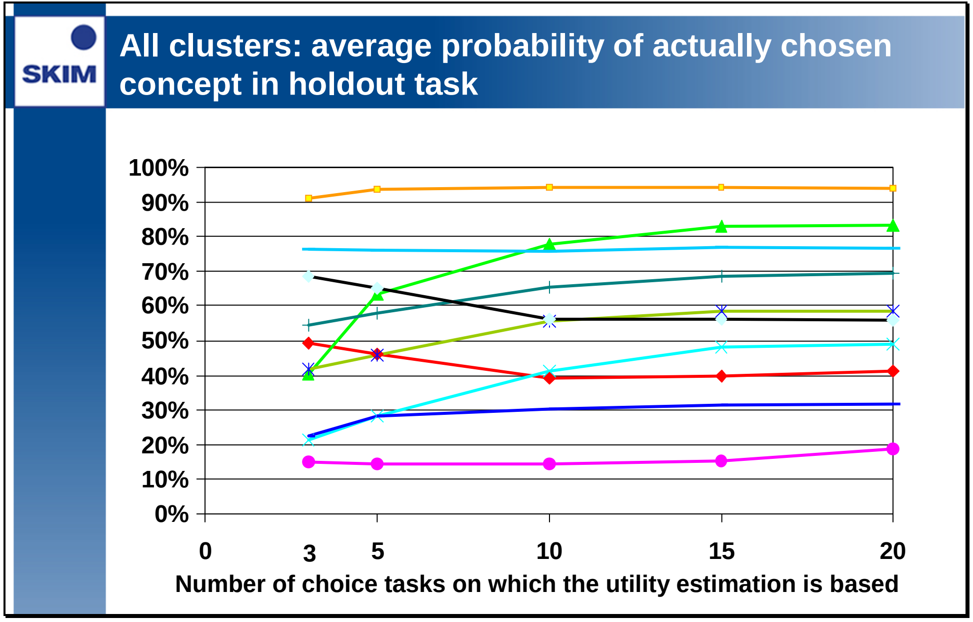
We come back to the earlier mentioned objectives against using a criterion like we did:

1. We cannot *really* determine if such an amount of choice tasks is enough because the *true* number of choice tasks required may fall "out of range". However, on a cluster level we have seen gradual increase or decrease for some clusters in the range up to 10 choice tasks, and a steady line for *all* clusters between 10 and 15 choice tasks. Though theoretically possible, it is practically speaking very unlikely that one or more clusters would suddenly change pattern again after 15 choice tasks.
2. We used only one holdout task. This is something that we will address when discussing the next study briefly.

In the second study that we will show, we did not use one holdout task, but instead we used four holdout tasks. We could easily sacrifice so many choice tasks to become holdout tasks since we had 29 choice tasks in total. The number of concepts per choice task was 15 in this case. Here we have only estimated HB utilities based on 3, 5, 10, 15 and 20 choice tasks.

We have calculated now, per respondent, the mean probability of the chosen concept across the four holdout tasks. Figure 5 shows the cluster results, again based on a 10-cluster solution, for this study:

Figure 5



Like in the previous study, we have clusters here that have converged almost immediately, even though, interestingly enough, this is based on four holdout tasks. The worst-performing cluster, consisting of only 4% of the respondents, shows an average probability of 15% which is at least clearly above random choice (with 15 concepts, random choice would result in an average probability of 7%). Most of the clusters that represent 58% of the sample have converged after or within 10 choice tasks. The other clusters converge after or within 15 choice tasks, strangely enough with the exception of the small cluster with very low probabilities. In any case, convergence takes place after or within 15 choice tasks for 96% of the sample.

THE SECOND METHOD: CRITERION

It would obviously be a drawback to draw conclusions based on only two commercial studies. However, though the previously discussed method was very promising and rather sophisticated, it was also too elaborate to apply on a larger number of studies.

Therefore, we decided on the following shortcut:

- we would include many other commercial studies in our analysis
- we would only compare actual choice of holdout tasks to a *first choice prediction*, rather than to a probability
- we would only take one holdout task per study
- we would only look at aggregate results
- we would take advantage of having many studies in our analysis by running a cross-study regression analysis that would show us a "cleaned" pace of convergence after removal of random (study specific) error

In our new approach, we had 35 studies at our disposal, conducted for two industries:

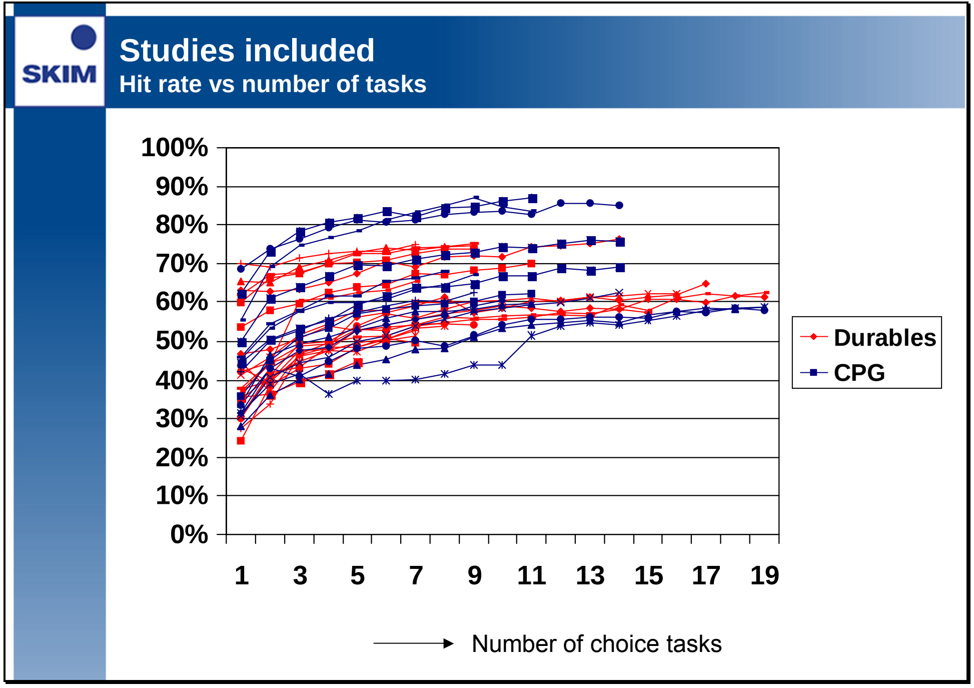
1. studies with CPG products, mostly with few attributes (generally: brand, package and price), mostly with many levels (many brands, many price levels), mostly with many concepts (each depicting a brand-package combination on the market) and always with full-profile tasks;
2. studies with "durables" products, mostly with many attributes, but a limited number of levels per attribute, often partial profile and always with a limited number of concepts per task.

Per study we used one choice task as a holdout task (dependent on the original total number of choice tasks) and—again—we examined to what extent adding choice tasks to the utility estimation would improve the hit rate for the holdout task.

THE SECOND METHOD: RESULTS

Figure 6 shows, in just a single graph, all raw data of mean hit rates per study.

Figure 6



Studies included

Hit rate vs number of tasks

It is interesting to see that hit rates fluctuate a lot from one study to another. Based on 10 choice tasks, the hit rate varies roughly from 50% to 80%. This is another indicator that one should never draw conclusions based on just one or two studies, as studies are too different from one another.

In addition, hit rates of some CPG studies are a lot higher than of durables studies, but from other CPG studies they are a lot lower. But since the number of concepts is generally a lot higher in CPG studies, it would be fair to state that CPG studies perform slightly better than durables studies.

For the purpose of this paper, the most important thing is that, for all studies we can see an increase of hit rate when the number of choice tasks is being increased, but the pace of increase is getting slower and slower as choice tasks are added.

As said, we have fit these data in a regression analysis in which we tried to explain the hit rate by the following variables:

- the number of choice tasks (obviously, since that is the purpose of our study)
- the parameters of the study (such as CPG versus durables, number of concepts etc.)

In addition, we have tried optimising the fit by trying to incorporate non-linear and interaction effects of these variables.

Because the dummy variable type of industry (CPG versus durables) correlates too heavily with other study parameters, we have decided to develop separate equations for the two types of industry.

Both for CPG and for durables we ended up in a multiplicative model with the following explanatory variables:

1. Number of choice tasks
2. Number of concepts in a task
3. Number of attributes and/or levels in the study (differently defined for CPG studies and durables studies)
4. Number of "active" attributes in a choice task (for durables studies only, applicable in case of partial profile)
5. An interaction effect between the number of choice tasks and the number of attributes

With respect to the interaction effect, it should be noted that the hit rate does not converge as quickly if the study is more difficult (i.e. has more attributes), which is logical. However, there were no interaction effects between the number of choice tasks and any other variables that were indicators of the complexity of the study, which is quite surprising. It means, rather, that the complexity of the study is not such an important argument for increasing the number of choice tasks.

The regression equations are as follows:

Durables:

Hit rate = constant * #choicetasks$^{0.21}$ * #attributes$^{-0.15}$ * #activeattributes$^{0.04}$ * #concepts$^{-0.34}$ * (#choicetasks / #attributes)$^{-0.10}$

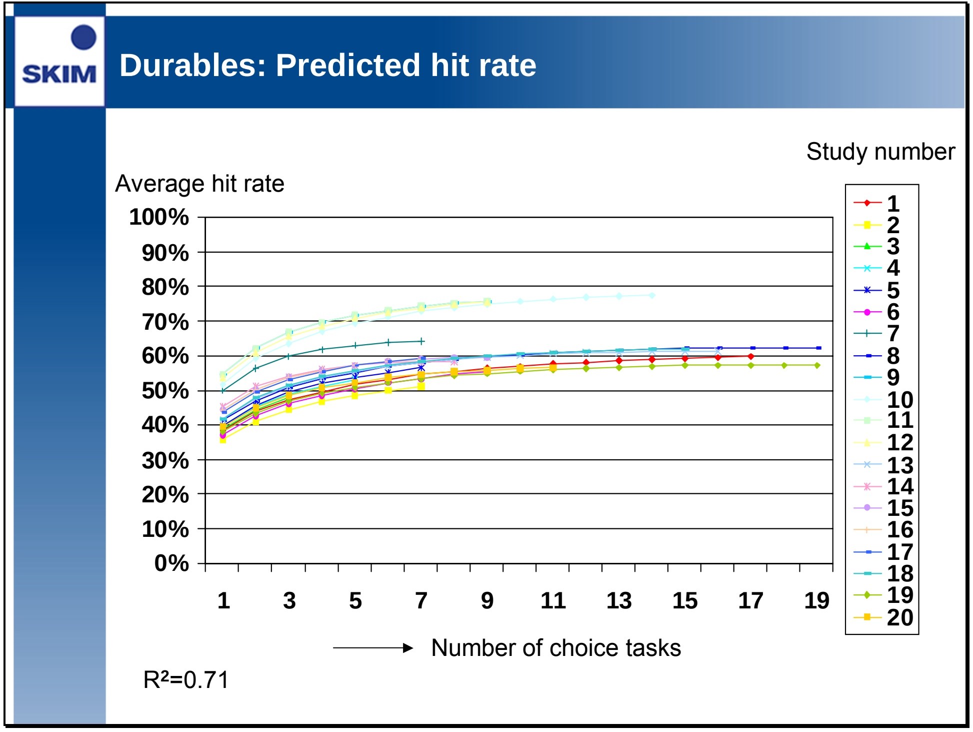
CPG:

Hit rate = constant * constant * #choicetasks$^{0.21}$ * (#attributes+0.03*#levels)$^{-0.50}$ * #concepts$^{-0.21}$ * (#choicetasks / #attributes)$^{-0.06}$

Both models had an R^2 of about 0.70 so that is a very good fit. Interestingly, both models have the same coefficient for the number of choice tasks.

If we now, finally, make a graph of the "cleaned" hit rates (the predicted hit rates by the regression equations), we get the following two pictures for durables and CPG respectively:

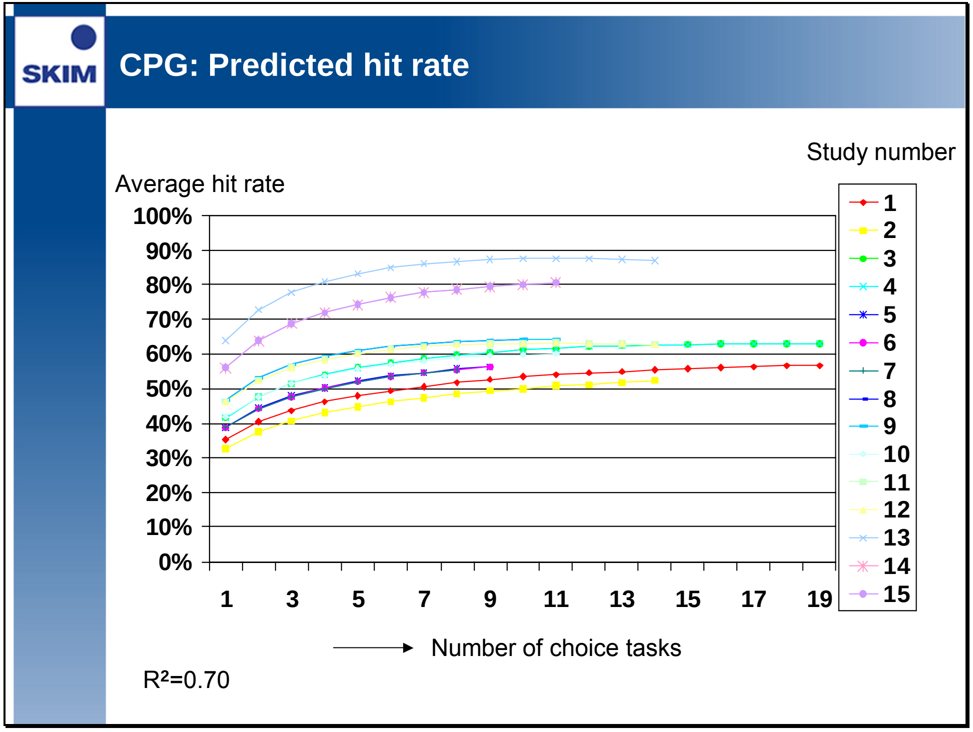
Figure 7



Durables: Predicted hit rate

R²=0.71

Figure 8



CPG: Predicted hit rate

R²=0.70

From these graphs, it is clear that generally it is sufficient to have only 10 choice tasks for the majority of studies. For some studies there is still an increase in hit rate after 10 choice tasks, but in these cases the difference in (predicted) hit rate gain between 10 and 15 choice tasks is so small that it would virtually not harm to stick to 10 choice tasks.

The complexity of the study (number of attributes, levels, concepts) has an important influence on the absolute magnitude of the hit rates but, as noted before, the influence on the pace of convergence is rather limited.

Discussion

We used two different methods and applied these on two same studies. (The second method was applied to 33 other studies as well.) The results are similar. Using the more sophisticated method leads to a minimum requirement of 10 choice tasks for one study and 15 choice tasks for another, while using the more shortcut method leads to an overall required minimum of 10 (or maybe 15) choice tasks.

We would like to make one additional observation about this. Obviously using a first choice prediction (as in the second method) is a less precise instrument than using a probabilistic prediction (as in the first method). But then, one could link this phenomenon with the type of model that one is using in the simulator. When using the first choice model in the simulator, one could argue that 10 choice tasks are enough for a good prediction (as from the second section of this paper). While, when one is using share of preference, it makes it worthwhile to increase the number of choice tasks from 10 to 15 (as in one of the two studies in the first section of the paper).

SAMPLE PLANNING FOR CBC MODELS: OUR EXPERIENCE

JANE TANG, WARREN VANDALE, AND JAY WEINER
IPSOS INSIGHT

ABSTRACT

How big a sample do I need? How many choice boards do I need to ask each respondent? Did I mention this is a sample of cardiologists? If I ask the respondent to allocate his spending (constant sum instead of single choice), do I still need to ask as many choice boards? What happens if I add 4 more factors to this design? These are the questions frequently asked by the project team. Most researchers today develop heuristic answers based on experience, rules-of-thumb and budget constraints.

This paper presents a systematic approach to sample planning for CBC models based on our experiences from over 250 models. We examine the impact of population characteristics, choice model input, model complexity on the effect of sample size and number of choice boards required. An Excel™-based tool is used to provide quick calculation. Examples from real life studies are used to demonstrate the use of the tools, along with hit rates resulting from the data.

BACKGROUND

Although most of the principles that influence the planning of a choice model are based on statistics, successful researchers develop heuristics for quickly determining sample sizes based on experience, rules-of-thumb and budget constraints (Orme, 1998). Communicating these heuristics throughout the organization is a difficult, if not impossible task.

The two main drivers of cost in any research project are sample size and interview length. Client service teams need to know the number of respondents needed to provide accurate cost details to the client. Frequently in the early stages of discussion the number of attributes and attribute levels in choice models is a moving target. Often the conversations begin with everything but the kitchen-sink and evolve into something more manageable. It is important to be able to come up with a reasonable estimate of the sample size and to be able to communicate the impact on sample requirements or interview length if attributes are added or removed.

There is no easy way to calculate sample size required for a given model. The formal approach of sample planning involves generating synthetic datasets of various sample sizes based on *a priori* utilities and design efficiency, and checking for resulting standard errors (Chrzan and Orme, 2000). This approach is not always easy to use in field applications and often not practical during early project planning discussions. This paper presents a simple Excel™-based planning tool that can quickly provide an estimate of sample size or the number of profiles that would need to be shown to estimate any specified model.

RULES OF THUMB

Most choice modelers began their careers working with conjoint. Because conjoint required complete records to estimate OLS regression, all the researcher needed to do was count the number of degrees of freedom required to fit the model, consult the Addleman (1962) tables for an orthogonal design and you had your experimental design (allowing for enough cards to assess consistency). Since conjoint was modeled at the individual level, sample size really didn't matter as long as you were comfortable making decisions based on the final sample.

Early choice models were fit using aggregate models. Typical rules of thumb based on regression models often drove the sample size. Typically, at least five observations per parameter is desirable. Again, counting degrees of freedom and multiplying by 5 seemed to work. One key advantage of choice based models was the fact that we could often test more factors and levels by showing respondents a subset of the total design. Blocked designs lead to the heuristic of 75 respondents per block. Again, we could work from the total degrees of freedom required for the model, decide how many tasks we could show respondents and compute the number of blocks required.

The application of these heuristics or rules of thumb is usually based on experience and the guidelines are often not clearly stated. These rules generally have little allowance for different sample and design specifications. For example, the more homogeneous the target sample, the fewer respondents you will likely need to fit a model. This paper attempts to put some order in these rules of thumb and present an Excel™-based systematic approach to sample planning based on our experience. The tool is designed to serve two primary functions. First, it is to help our less experienced marketing scientists deal with the planning of choice experiments under a variety of conditions. Second, it is to give confidence to our client service team dealing with minor changes in the design specification without having to constantly rely on the marketing science team. For example, if we need to add one more factor, how many more choice tasks does each respondent have to do?

SOME LIMITATIONS OF THE TOOL

This is ***not*** a formal power analysis approach. We do not attempt to establish the link between sample sizes and estimate standard errors. We don't know how the efficiency of any experimental design would affect the results. Our approach is validated by whether the results agree with the heuristics/rules of thumb commonly used. After the study is fielded and model estimated, we check Hit-rate statistics on the holdout task—it needs to be significantly better than chance. If a block design is used, we also check to see if MAEs on all random holdout tasks are well within the margin of error assuming random choice.

OUR EXPERIENCE

Since its inception in 1998, Marketing Sciences at Ipsos Insight has conducted conjoint research in a broad array of industries including packaged goods, pharmaceuticals & healthcare, telecommunications, technology, financial services, lotteries and gaming, airlines, automotive, agricultural, mass media, building products, energy utilities, and tourism. The department has conducted more than 275 choice models. More than 65% of those models are choice based conjoint or discrete choice models.

The category of choice models includes many different forms. The planning tool can be used for: Paired Comparison, MaxDiff Comparisons, Full Profile Choice Based Conjoint and discrete choice models. The planning tool should not be used for: Ranking/Rating Based Conjoint, Self-explicated Conjoint, Adaptive Conjoint Analysis, or Partial Profile Conjoint.

SOURCES OF ERROR IN CBC PROJECTS

Errors are deviations from "truth." There are several sources of error in any research project. Sampling error occurs when a sample of respondents deviate from the population.

With a random sample, increasing sample size can reduce sampling error, but not bias. With convenience sampling, there is no way to objectively evaluate the amount of sampling error. Measurement error is the error introduced because we cannot precisely measure respondents' preference. Measurement errors can be reduced by getting better/richer information from each respondent, for example, by using maxdiff selection instead of single choice selection. Measurement error can be reduced by asking respondents to do more choice tasks.

THE BASIC RULE FROM JOHNSON & ORME (ORME, 1998)

For aggregate-level CBC, treat the choices like proportions. Having respondents complete more tasks is approximately as good as having more respondents. Johnson recommends,

$$nta / c >= 500$$

where:

n = number of respondents
t = number of tasks
a = number of alternatives per task
c = number of "analysis cells"

When considering main-effects, c is equal to the largest number of levels for any one attribute. If you are also considering all two-way interactions, c is equal to the largest product of levels of any two attributes.

Johnson's rule is for determining minimum sample sizes for aggregate level modeling. Since an HB model is an improvement over an aggregate model, we reason that if a sample meets the criteria for an aggregate model, it should do fine for disaggregate HB modeling.

We interpret "c" as an indication of the model complexity. Johnson used "c" as a useful heuristic when Sawtooth Software's CBC program was limited to a small number of factors and levels (6 factors). With larger numbers of factors and levels, this doesn't seem to be appropriate. A model with 10 factors (each with 3 levels) is more complicated than another model with only 3 factors (each with 3 levels), and should require more observations to estimate. Most of our models are disaggregate HB models with main effects estimates for individual respondents. We feel the model degrees of freedom, e.g. the number of parameters to be estimated from the non-constant/full-profile alternatives, is a better indicator of the model complexity.

Disaggregate models with interactions are just more terms for the model, so the same degrees of freedom rule should still apply. Our experience with modeling interactions at the individual level is limited. Most disaggregate modelers feel that the heterogeneity in the individual estimates accurately capturers the information that might be gained in fitting interaction models.

We interpret the constant "500" as being an indicator of the inherent variability in the model from a general population sample. Samples from specialized populations may have smaller variability. Specialized populations are likely to be more homogeneous. Certain populations may also be better at decision making, i.e. more logically consistent choices. For example, when compared to the general population, the physician population is both more homogeneous with respondents having similar educational and socio-economic backgrounds, and is better at making logically consistent choices.

The Homogeneity/Logical Consistency Index (HLI) in our model is a measure used to indicate the amount of variance expected within a certain sample: the more homogeneous the population, the lower the index; the more logical the respondents' answers, the lower the index. For general population studies, we suggest that HLI retains the original value Johnson proposed - 500. For specialized populations, e.g. General Practice MD, a smaller HLI value of 100 has worked well for us.

We interpret "t" as the quantity of information given by each respondent. When a respondent chooses a constant alternative (including NONE), he gives little information regarding his preference for factor levels in the non-constant alternatives.

Only those choice tasks where a non-constant/full profile alternative is chosen count towards modeling. We compensate for the exclusion of constant choices by also excluding the constant alternative parameters from model complexity.

We interpret "a" as the richness of information given by each choice task. We reason that having more non-constant alternatives in the choice task gives us more information.

Realistically most CBC experiments on the web are done with no more than 5 non-constant alternatives. The nature of choice task also contributes to the richness of information being collected. Constant Sum Allocations provide more information than MaxDiff designs. Dual Response tasks asking preference followed with the likelihood to purchase provides more information than single choice questions.

We suggest using the following grid to determine "a" - the richness of information available for a choice task. The value of "a" generally retains the original value Johnson proposed, but with a less steep increase for the larger number of non-constant alternatives for single choice, and larger values for the more informative type of choice task.

These values have worked well for us. The resulting sample size and number of tasks match what we would expect from experience. The model based on the subsequent actual data has been found to have acceptable hit rates and MAE from the holdout task.

	Number of Non-constant Alternatives				
Nature of Choice Task	1	2	3	4	5+
Single Choice	1.0	2.0	2.5	3.0	4.0
Dual Response	N/A	2.2	2.7	3.2	4.2
MaxDiff	N/A	N/A	3.0	3.5	4.5
Allocation	N/A	2.5	3.0	4.0	5.0

Our revisions to Johnson's formula:

$$\frac{nt(1-c\%)a}{d.f.} > HLI$$

Where:

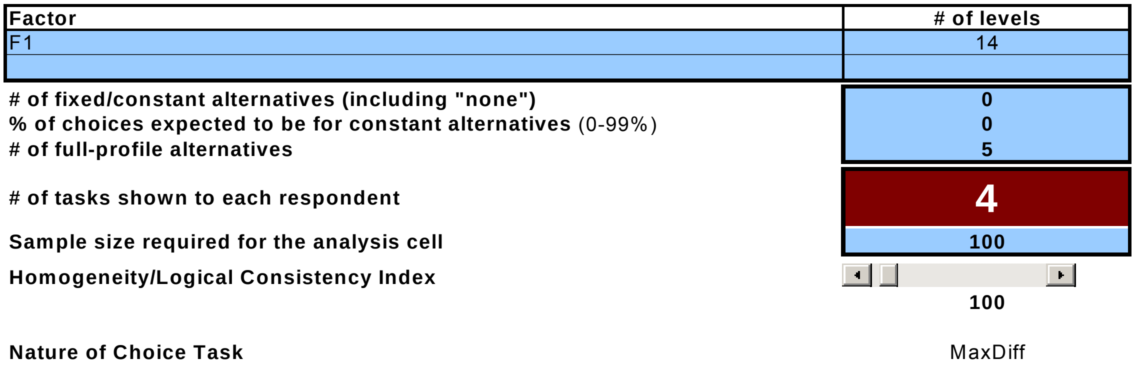
n = sample size
t = number of choice tasks
a = richness of information from each choice task
c% = percent of time the constant alternatives are chosen
d.f. = degrees of freedom of the model (excluding constant alternatives)
HLI = Homogeneity/Logical Consistency Index

Excel™-based Calculator

To aid in the planning of choice models, we developed an Excel-based tool. The spreadsheet allows the user to solve for either the number of tasks (t), or the sample size (n). The calculator allows for one random holdout task. The number of tasks required should allow each level to be shown at least once. The calculator solution is best used as the minimum requirement for sample planning as Johnson originally intended.

Example 1

The client has budget to sample 100 General Practice MDs. We are interested in the importance of 14 attributes associated with the treatment of a particular condition. If we want to use a MaxDiff design, how many choice tasks do we need to present to each respondent? We set the planning tool as follows:



Factor	# of levels
F1	14

# of fixed/constant alternatives (including "none")	0
% of choices expected to be for constant alternatives (0-99%)	0
# of full-profile alternatives	5
# of tasks shown to each respondent	4
Sample size required for the analysis cell	100
Homogeneity/Logical Consistency Index	100
Nature of Choice Task	MaxDiff

How well did this work?

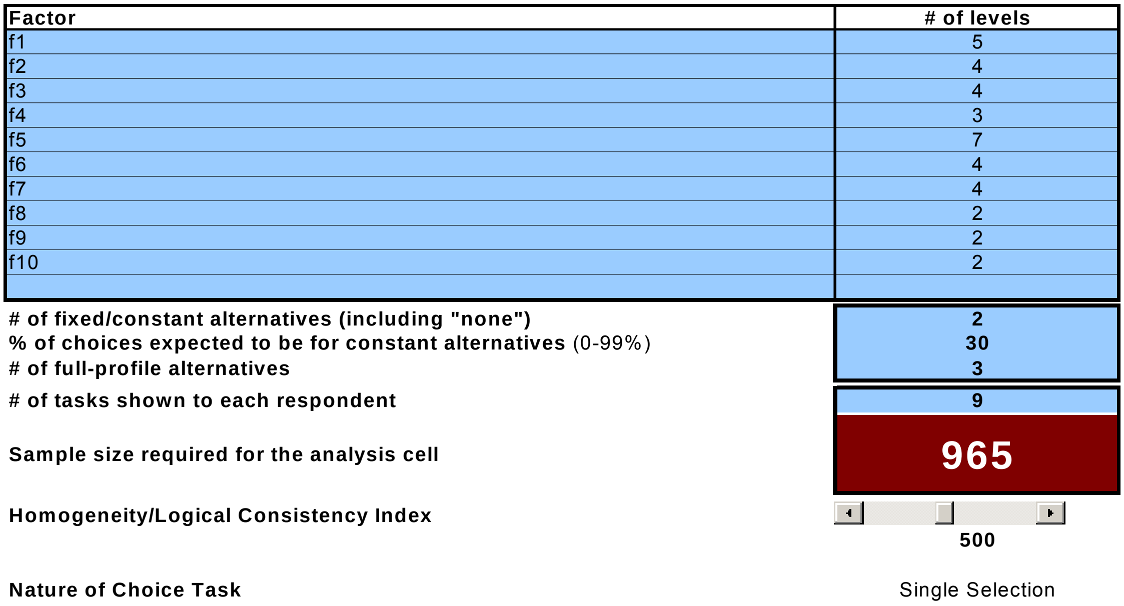
The final sample obtained was n=97. Using a 5-block design, each respondent completed 5 choice tasks. Using 3 tasks per respondent in an HB model, hit rate on random holdout tasks was 55%. MAEs from all holdout tasks were well within the margin of error

Example 2

The client is interested in testing the preference of performance attributes for its product.

The choice tasks are set up as a single choice among 3 full-profile alternatives and the 2 constant alternatives. The client wants to show no more than 9 choice tasks to each respondent.

The product is to be marketed to the general population. It is estimated that approximately 70% of the population may be interested in the product. We set the planning tool as follows:



Factor	# of levels
f1	5
f2	4
f3	4
f4	3
f5	7
f6	4
f7	4
f8	2
f9	2
f10	2

# of fixed/constant alternatives (including "none")	2
% of choices expected to be for constant alternatives (0-99%)	30
# of full-profile alternatives	3
# of tasks shown to each respondent	9
Sample size required for the analysis cell	965
Homogeneity/Logical Consistency Index	500
Nature of Choice Task	Single Selection

How well did this work?

The final sample was n=1,049 with a 10-block design. Using 8 choice tasks per respondent in an HB model, hit rate on random holdout tasks was 84%. MAEs from all the boards were well within the margin of error.

Summary

We present a revised rule of thumb for sample planning for CBC models. The Excel^{TM}-based tool allows researchers to easily implement design/sample specifications. The various values used in the formula are based on our experience. The tool is validated by other commonly used heuristics/rules for sample planning and by the hit rates/MAEs in the holdout tasks once actual data have been collected. There is a disconnection between the formal approach to sample planning (based on design efficiency and synthetic data) and field practice. By highlighting the issues faced in the daily practice of choice modeling, we hope to challenge fellow researchers to bridge that disconnection, and develop an easy to use link between sample & design specifications to the standard error of the estimates, hit rates and MAEs.

References

Addelman, S (1962): "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments" in Technometrics, Vol. 4(1), pp. 21-46.

Chrzan, K. & Orme, B. (2000): "An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis," in Sawtooth Software Technical Papers, www.sawtoothsoftware.com/techpap.shtml.

Orme, B. (1998): "Sample Size Issues for Conjoint Analysis Studies," in Sawtooth Software Technical Papers, www.sawtoothsoftware.com/techpap.shtml.

BRAND IN CONTEXT: BRAND DEFINITION IN VOLATILE MARKETS

ANDREW ELDER
ILLUMINAS

ABSTRACT

What happens when an established brand in one marketplace seeks to leverage its brand equity into a new product category? In the case of a brand extension, previous research would suggest the company should seek to selectively leverage its qualities into a new category that customers perceive as a reasonable fit. Arm & Hammer can leverage its brand into other cleaning products, while Harley-Davidson would be advised to avoid cake decorating kits.

As usual, however, the acceleration of technology has created unusual variations on a familiar theme. Such is the case with the "Triple Play," wherein telecom, cable, and internet providers seek to leverage their brand equity to provide services in a convergent market. The digitization of media allows each provider to lay claim to voice, video, and data services, and brands are rushing to capture each other's customers through bundling.

When modeling adoption of the merged Triple Play offering, traditional brand attributes carried similar weight as in a traditional "single-category" model, demonstrating that core branding components are not strictly bound to their established category. Yet this study also found that certain attributes related to the category itself had a significant impact on Triple Play adoption, suggesting clear contextual effects that affected the ability for certain brands to leverage their reputation into the convergent category. Some of these results are consistent with brand extensions, while others appear unique to the Triple Play market, and may be applicable to technology convergence more generally.

INTRODUCTION

In a simpler time, the world was a simpler place. You bought your meat from a butcher. You bought books from a bookstore. Your music player came from a consumer electronics manufacturer, and your PC came from a computer OEM. And you bought voice, video, and data services from tidy, semi-regulated providers.

Deregulation, technological innovation, and competition have blurred the lines between established product categories. The computer has evolved from a specialized calculation tool to an appliance that combines information and media with connectivity. Voice, video, and data providers all serve as conduits into and out of this gateway, leading to competition for new and existing services.

Convergence is happening everywhere. In the retail world, mega-stores offer beef and best sellers under the same roof. In the virtual world, voice, video and data are all coming through the same pipe. In both cases, there are new and evolving battles as to who best provides those services.

In a simpler time, branding was also a simpler task. A company could compare themselves against competitors who provided similar services, and differentiate themselves accordingly.

The brand positioning might not directly reference the end product – "the quiet company" or "think different" have no inherent product or service connotation – but the goal was implicitly to build a brand identity that would sell more life insurance or personal computers.

But what happens when markets converge? Consumers face considerable hurdles to deciphering evolving technology markets, and even more confusion when familiar services are offered in new formats from unfamiliar faces. And brands have the daunting challenge of communicating these evolving value propositions to the end consumers while protecting their historical brand advantages.

For years, networking companies have been touting the advantages of IP infrastructure as an alternative to public switched telephone network (PSTN) for voice delivery. Digital networks are built for transmitting data, and voice traffic can effectively be manipulated as simply another form of data.

By transforming voice data into digitized packets, businesses and consumers can gain cost advantages over standard analog toll pricing. VOIP also opens up the possibility that any high-speed network can carry voice traffic, so that telephone service can be delivered by:

- Telecom providers
- Cable providers
- Internet service providers

The challenge for all three providers is to educate consumers about VOIP, and merge this evolving service into their existing brand propositions. In early 2005, most providers were looking for ways to package VOIP with their existing services. The integration of voice, video and data was generally referred to as the "Triple Play."

The evolving brand structure necessitated by the Triple Play service offering is emblematic of the overall trend towards convergence in technology marketplaces (e.g. Apple + iPod, Dell + TVs). This not only creates substantial challenges for brands, but also for brand measurement.

There is no shortage of research available on brand extensions, which is the practice of applying an existing brand name to a new product category. Product extensions are predominantly created according to one of two scenarios:

1. An established brand invents a new category that builds on their existing strengths. The brand, for a time, has this new category to themselves, but also carries the burden of creating and promoting an entirely new concept. Example: Iams creates insurance for pets.
2. An established brand moves into another category that is already established, but different from the brand's core offerings. The brand is likely entering a crowded space, but hopes their brand strengths will differentiate them from the competition. Examples: Starbucks markets a coffee liquor, Emeril Lagasse markets cookware, etc.

The number of brand extensions that are simply plying a new name in an existing category (scenario 2) far outnumber the creation of new products (scenario 1). And neither accurately encompasses the emergence of Triple Play, which is characterized by both the diffusion of a new service (digital phone) and the simultaneous convergence of competition among brands that were previously established in separate categories.

Our goal was to create a brand research methodology that could address the uncertainty in the Triple Play marketplace, and provide a structure for evaluating cross-category brand convergence more generally.

THE RESEARCH OPPORTUNITY

To understand the Triple Play purchase process and branding opportunities, we explored the following areas:

Triple Play At the core of the research is whether respondents know and understand the Triple Play concept, how much they value the opportunity, and at what levels they expect to adopt the technology.

Branding Even without using VOIP, respondents have extensive experience and perceptions around traditional service providers that will influence their preference for Triple Play adoption.

Personality Individuals have unique characteristics and attitudes that determine their interest in new products, innovation, and early adoption beyond Triple Play.

The overall structure for this research is complicated by the fact that there is no clear and consistent category to form a common basis for comparison among brands. Further complicating the assessment is that cable and telecom service areas are fragmented, with even the largest players having limited service area and scope. This makes it impractical to compare Time Warner against Comcast when the two competitors have limited regional overlap, and thus fragmented customer experience (perception).

Thus we constructed a survey instrument that attempted to mimic consumer decision-making, with brands and categories defined primarily by current usage. Triple Play was introduced as a concept, and then evaluated from both a category (cable vs. telecom) and brand (Time Warner vs. SBC) perspective.

This approach is a departure from traditional brand measurement, which tends to be narrow (defined strictly by category) or broad (completely unrestricted comparisons).

Our study was fielded during March of 2005 using a web-based survey consisting of approximately 30 distinct questions. The survey typically lasted between 17 and 20 minutes, using online panelists from the US only. Respondents were accepted as qualified participants if they currently used at least two categories of the Triple Play service offering (voice, video, and/or Internet), and had at least minimal familiarity (name recognition) with at least two voice providers and two video or data providers.

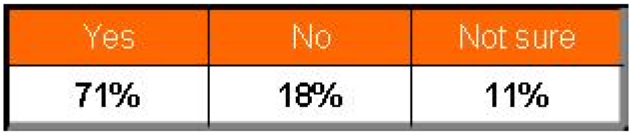
Our screening criteria emphasized a higher level of usage and awareness than the market average. Our assumption was that those with minimal services currently would be relatively uninterested in Triple Play, and those with minimal knowledge would lend little insight into brand and category considerations.

The VOIP market was and continues to evolve rapidly. In the time frame of this study, stand-alone VOIP was less relevant than Triple Play, although the same study today would likely incorporate Skype, Vonage, and similar players.

TOP-LINE FINDINGS

The initial results produced some insights into the Triple Play opportunity. Most (71%) respondents had some knowledge of the concept of bundling video, voice, and data, but only four percent of our sample used the same provider for all three services at the time of fielding. Despite low adoption levels, our sample expressed relatively strong interest in purchasing a Triple Play bundle, with 64% rating themselves as "interested" (top 3 box on a 7-point scale). These elevated levels of awareness and interest, combined with almost non-existent adoption, suggests a clear opportunity for integrating digital voice offerings with video and high-speed Internet service packages.

Figure 1. Have you ever heard of a company providing all of these services together [cable/satellite TV, home telephone and high-speed Internet access]?



Yes	No	Not sure
71%	18%	11%

Figure 2. Which company do you currently subscribe to for your [cable/satellite TV, home telephone and high-speed Internet] service?

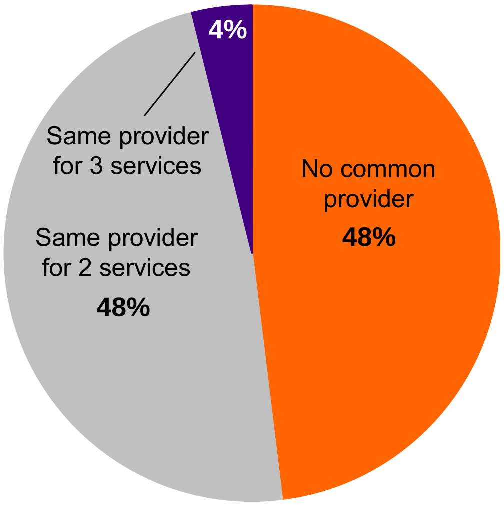
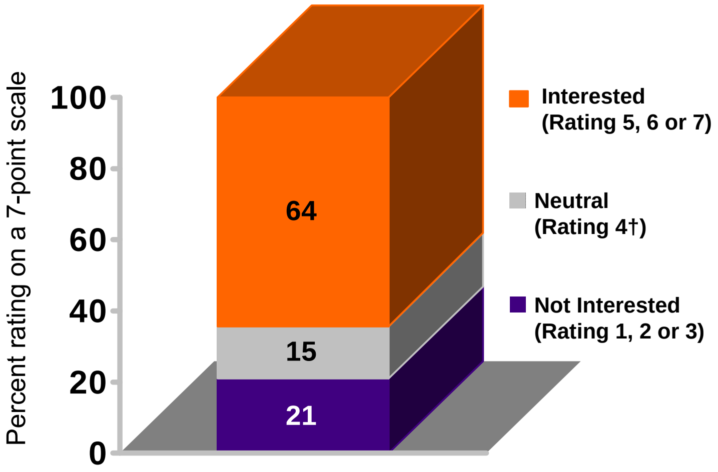
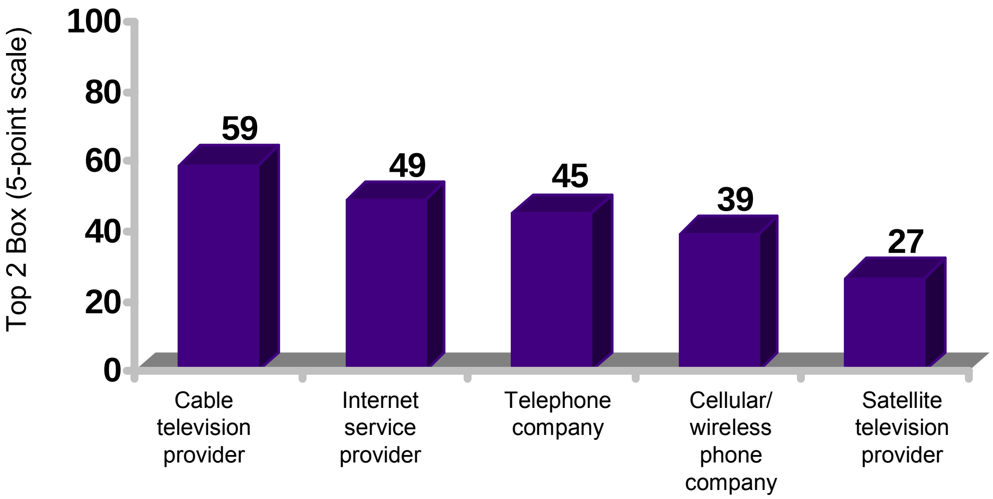


Figure 3. How interested would you be in purchasing video, voice and high-speed Internet services from a single provider?



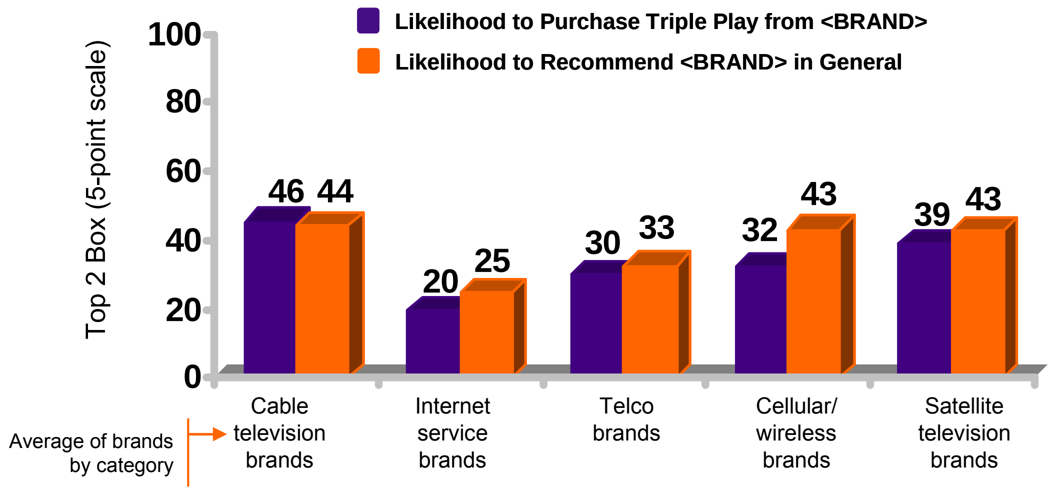
Finding consumer interest in Triple Play leads to the question of "yes, but from whom?" Interestingly enough, the answer changes subtly depending upon how you ask the question. First, we asked respondents what type of company they'd consider for their Triple Play purchase. Without imposing specific brands on the decision, consumers preferred cable companies. More surprisingly, their next preference was for Internet Service Providers over telephone companies, wireless providers, or Satellite TV providers.

Figure 4. How likely would you be to consider purchasing Triple Play from each of the following types of companies?



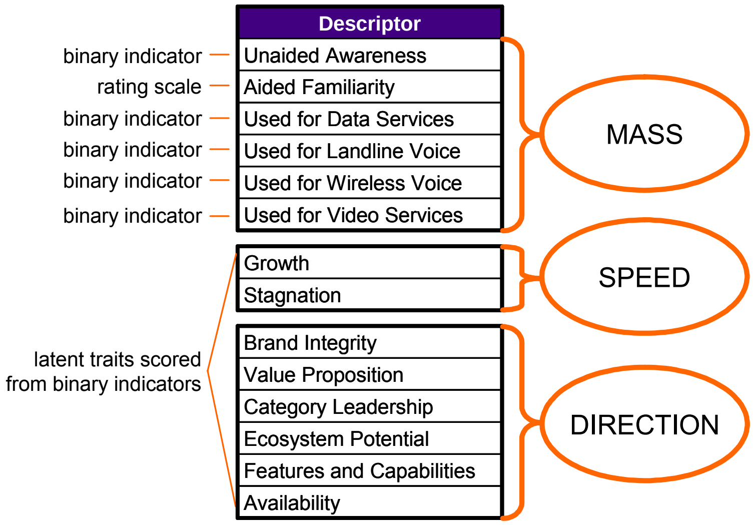
This unexpectedly strong performance for ISP's becomes all the more curious when Triple Play interest is evaluated in terms of specific brands. If asked to evaluate specific providers, dedicated ISP's fare worse than every other category, swapping positions with Satellite television brands for second place behind cable brands. This raises the question of why stand-alone ISP brands are much less desirable than a generic Internet supplier, while satellite TV brands have the opposite effects.

Figure 5. How likely would you be to purchase Triple Play from each of the following brands? / recommend each brand to others?



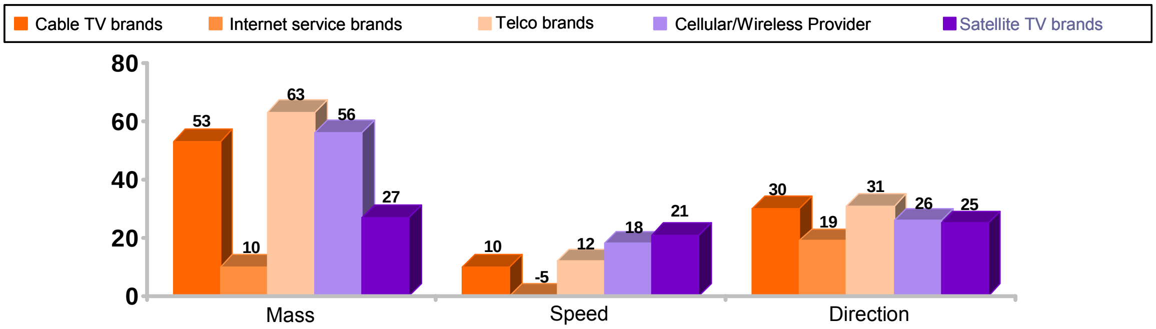
One element of this explanation is apparent from the component attributes that are evaluated across brands. Illuminas uses a brand evaluation model that is based on the physical concept of momentum that measures brands according to their Mass, Speed, and Direction. Mass attributes represent the brand experience or "footprint," as measured by knowledge and usage. Speed and Direction concepts are latent dimensions aggregated from binary attributes that measure each brand's growth or decline (Speed) and the characteristics that constitute the brand's overall perceptual composition (Direction). When combined, these components provide a succinct description of each brand's "momentum" in the marketplace.

Figure 6. Components of the Illuminas Brand Reputation Model



The component attributes show a clear relationship with the preference outcomes shown in Figure 5, especially with reference to the brand weakness among ISP brands. From this table we can surmise that ISP brands fail to garner much consideration or recommendation because they generally lack Mass, Speed, and Direction relative to competitors in other categories.

Figure 7. Brand Attribute Evaluations

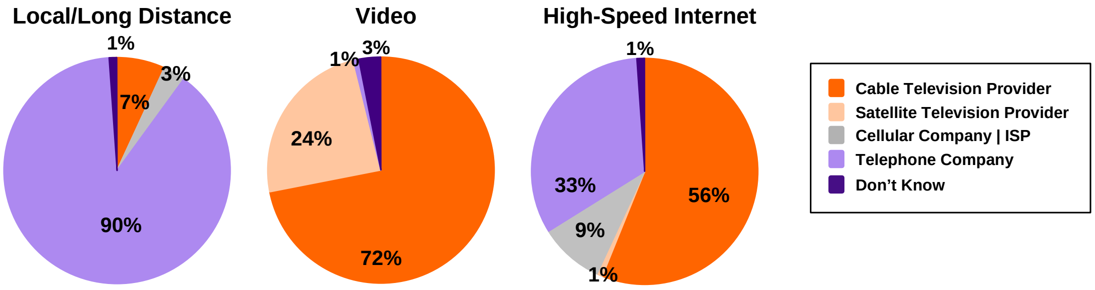


And yet the component dimension scores do not completely explain our brand-level outcomes, nor the discrepancy between brand and category evaluations. Most notably, the component scores would predict that Telco brands, based on their strong Mass and Direction scores, should be more preferred than their cable counterparts. Yet in both the brand and category evaluations, cable is clearly stronger than Telco. The component scores are similarly vague as to why satellite TV brands should perform so much better than their generic category evaluation would suggest.

Introducing Context

One explanatory factor may have to do with the context that surrounds each of the service providers. Cable companies may benefit in Triple Play consideration because their current penetration and consumer exposure reaches across multiple service categories – high speed Internet and video services. While Telephone companies "own" voice services, they do not have the dominant levels of cross-over that Cable currently enjoys.

Figure 8. Which company do you currently subscribe to for each service?



As brands converge together in new markets, their value proposition is at least partly determined by the categories in which they compete. In other words, their context changes as they cross category lines. Respondent expectations from a brand are at least partly defined by the services that they fulfill.

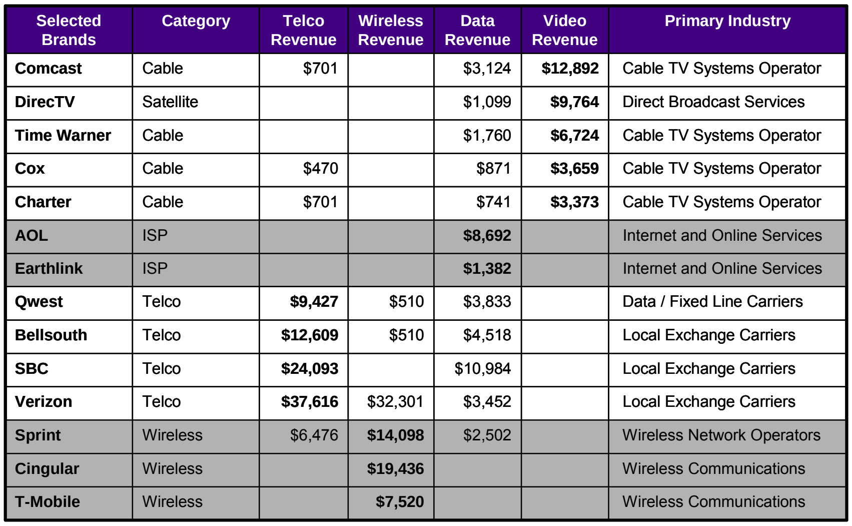
Why do respondents demonstrate strong consideration in Triple Play from “Internet Service Providers,” but have such low interest in purchasing it from actual ISP brands? The short answer is likely that respondents identify VOIP as requiring high-speed internet, and thus make the connection that Triple Play is desirable coming from any company that can also supply broadband. But relatively few individuals get their broadband from a pure ISP. Instead they are more likely to look to the Telco or cable companies who fulfill their broadband to also supply Triple Play.

Recall that our goal was to create a brand research methodology that could address the uncertainty in the Triple Play marketplace, and provide a structure for evaluating cross-category brand competition more generally. As a result of observing the inter-relationship between brand- and category-level results, we hypothesized that brands are partly defined by the categories in which they compete. The primary category characteristic that a brand must either leverage or overcome is the relevance of their core category to the end user.

One of our first hurdles in testing this hypothesis is finding a consistent method for meshing category-level observations with brand-level observations. In some cases this linkage is simple, such as with DirecTV; all perceptions of satellite TV as a category can be linked to DirecTV as a brand, since the brand and category are virtually synonymous.

The situation becomes much more complicated, however, if we consider brands such as Verizon and Comcast. Both brands have substantial overlap in the Triple Play offerings, which confounds our ability to identify category-level linkages with the brand. In the case of Comcast, the brand derives the bulk of its revenue from video, so most consumers are likely to consider Comcast as a cable brand with relatively little dilution from data and voice. Verizon is perhaps the most complex brand in this emerging category, since it derives almost equal revenue from both wireline and wireless voice services.

Figure 9. Revenue Sources by Brand



Selected Brands	Category	Telco Revenue	Wireless Revenue	Data Revenue	Video Revenue	Primary Industry
Comcast	Cable	$701		$3,124	**$12,892**	Cable TV Systems Operator
DirecTV	Satellite			$1,099	**$9,764**	Direct Broadcast Services
Time Warner	Cable			$1,760	**$6,724**	Cable TV Systems Operator
Cox	Cable	$470		$871	**$3,659**	Cable TV Systems Operator
Charter	Cable	$701		$741	**$3,373**	Cable TV Systems Operator
AOL	ISP			**$8,692**		Internet and Online Services
Earthlink	ISP			**$1,382**		Internet and Online Services
Qwest	Telco	**$9,427**	$510	$3,833		Data / Fixed Line Carriers
Bellsouth	Telco	**$12,609**	$510	$4,518		Local Exchange Carriers
SBC	Telco	**$24,093**		$10,984		Local Exchange Carriers
Verizon	Telco	**$37,616**	$32,301	$3,452		Local Exchange Carriers
Sprint	Wireless	$6,476	**$14,098**	$2,502		Wireless Network Operators
Cingular	Wireless		**$19,436**			Wireless Communications
T-Mobile	Wireless		**$7,520**			Wireless Communications

All revenue estimates shown in millions of USD.
Estimates pulled from 2004 annual reports where possible.

One impetus was to affix category-level data to a brand based on respondent usage. Thus, if one respondent uses Verizon only for wireline services and another uses Verizon only for wireless services, then the two respondents would have different category assessments assigned to the same brand. In theory, this would account for the differential experience of both respondents, which is meaningful to the extent that companies often have separate organizations delivering various services, creating the opportunity for diverse customer experiences and perceptions.

The primary problem with this approach is two-fold. The first problem is how to juggle overlapping delivery. Many respondents use Verizon for both wireline and wireless services, but does this inherently mean that their perceptions of Verizon are evenly divided between the two? While both experiences undoubtedly color brand perception, such a cleaving seems unlikely. The second hurdle is the gap left by brands that are not directly experienced. If a respondent uses SBC for wireline services and Cingular for their wireless, how do we assign their category affiliation for Verizon?

Both problems could perhaps have been overcome had we asked respondents to assign affiliations between brands and categories, perhaps even assigning weights or constant sums across multiple categories. Given the length and complexity of the questionnaire (already running 20 minutes in length, on average), this exercise was abandoned. In light of the uncertain linkage between brand and category, perhaps such an exercise should be explored in future efforts.

Without self-explicated associations, we relied on *a priori* assignment to link brands with categories. Using the revenue distributions shown in Figure 9, brands were assumed to have an affiliation with whichever category provided the largest revenue source for the brand. This rule generally defined each brand according to its historical business and brand associations. Even this simplistic approach has complications due to the increasingly-interconnected corporate ownership of distinct brands, such as Time Warner's ownership of AOL. Although AOL is part of Time Warner, its revenues are reported separately and clearly should not influence Time Warner, which is distinctly branded as a cable provider.

Figure 10. Correlations Between Category Consideration and Brand Purchase

			Categories		
Brands	Satellite	Telephone	Cable	Cellular	ISP
DirecTV	**0.667**	-0.058	-0.099	0.079	0.046
Bellsouth	0.206	**0.526**	0.059	0.350	0.281
SBC Communications	0.139	**0.446**	-0.006	0.023	0.104
Verizon	0.173	**0.314**	0.116	0.278	0.215
Comcast	-0.029	0.130	**0.476**	0.152	0.220
Time Warner Cable	-0.134	-0.066	**0.461**	0.113	0.073
Cingular	0.310	0.229	0.021	**0.389**	0.296
T-Mobile	0.255	0.121	0.055	**0.373**	0.245
Earthlink	0.181	0.221	0.041	0.373	**0.315**
AOL	0.100	0.118	0.050	0.227	**0.099**

All attributes rated on a 5-point scale. Pearson correlation coefficients shown.
Bolded scores represent expected intersection between brands and categories.

This *a priori* assignment proved to be more effective than attempting to shuffle or merge ratings based on usage. We asked respondents how likely they were to consider Triple Play from each category, and (separately) how likely they were to consider Triple Play from each brand. When correlating these two scores, as is shown in Figure 10, there is an expected synchronicity between brand and category, with the highest scores occurring at the expected intersection. As previously noted, DirecTV has the strongest overlap with its category. Notably, the ISPs have the weakest overlap, perhaps due to the cannibalization of high-speed Internet service from cable and Telco providers. It is interesting to note that Verizon does suffer greater dilution than other Telco brands, while Comcast retains a stronger focus on its core cable base.

MODELING THE TRIPLE PLAY

Most branding studies from the technology sector examine brand within the confines of a particular product category. This has less to do with industry or corporate considerations, and more to do with the purchase scenario we are trying to measure.

If I want to evaluate the server market, I identify respondents who are responsible for server purchases. This inherently places any brand observations within the context of a server purchase and an appropriate competitive set, which implies that brand is valuable because it influences or informs a sale. But companies like IBM, Sun, and Dell have a brand presence ("halo") that extends far beyond servers. Can we truly expect brand metrics to pertain directly to a category purchase?

At the other extreme are branding studies that compare players across industries, regardless of the differences in their products and/or services. Such an approach looks at branding as a general corporate asset, which implies that brand is a valuable component of a company's valuation.

This strategy helps quantify the largest and most successful brands, but isn't necessarily as informative for comparing companies that aren't Coke or Google. And how relevant is it to compare a soda against an Internet portal? From an investment perspective, it is highly relevant. From a product and purchase perspective, not so much.

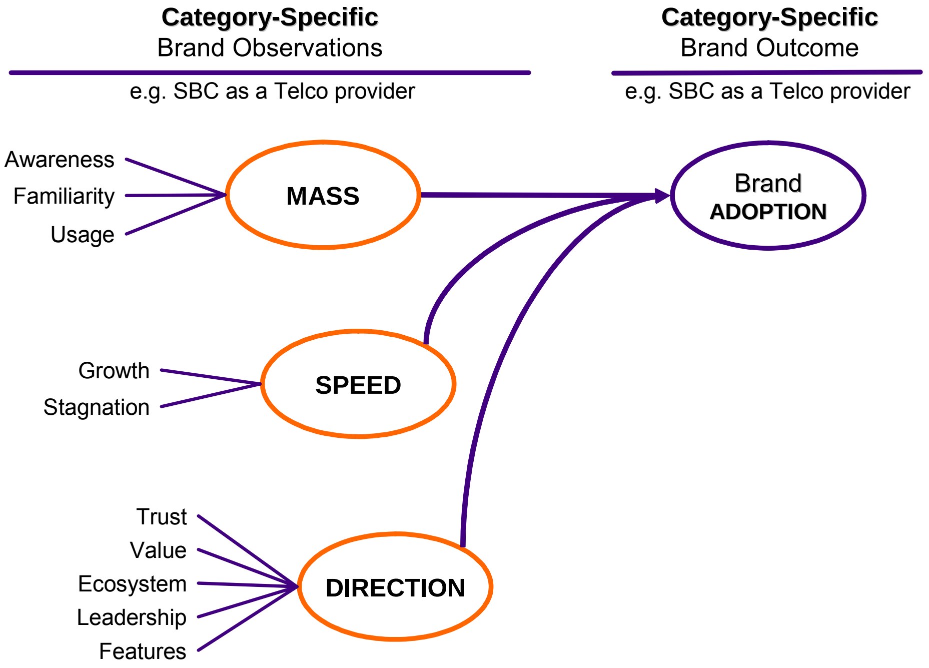
Since this study is intended to measure purchase intent, our inherent hypothesis is that brands are built to drive customers towards a specific purchase. This purchase is based on different criteria across different product categories, thus we expect to find brand attributes with varying influence depending upon the desired outcome.

To this end, we employ a brand model that is grounded in a structure that is common to Geoffrey Moore's notion of "crossing the chasm," and similar approaches by David Aaker and others. The underlying assumption is that a brand must be built and positioned, because a relatively small portion of any technology market (the early adopters) will make a purchase based solely on the tangible characteristics of the product. In order to achieve greater diffusion, technology products must convey attributes beyond "speeds and feeds" that will resonate with less technically-inclined decision-makers.

At the foundation of these attributes is brand awareness, built by previous experience with products and knowledge obtained through various formal and informal channels. From that knowledge, each individual forms an opinion about the brand, which we measure based on association with a series of attribute statements. Brands are evaluated within a competitive context of other "familiar" brands in order to engender comparisons.

After building a brand definition based on these attributes, we then use these associations to predict some element of overall brand affinity, such as willingness to recommend the brand to others or intent to purchase the brand for a specific product. In either scenario, the inputs are traditionally aligned with the output in terms of context; if we are attempting to predict the purchase of a file server, then we evaluate both the attributes and outcomes within the context of server products.

Figure 11. A Traditional Brand Model

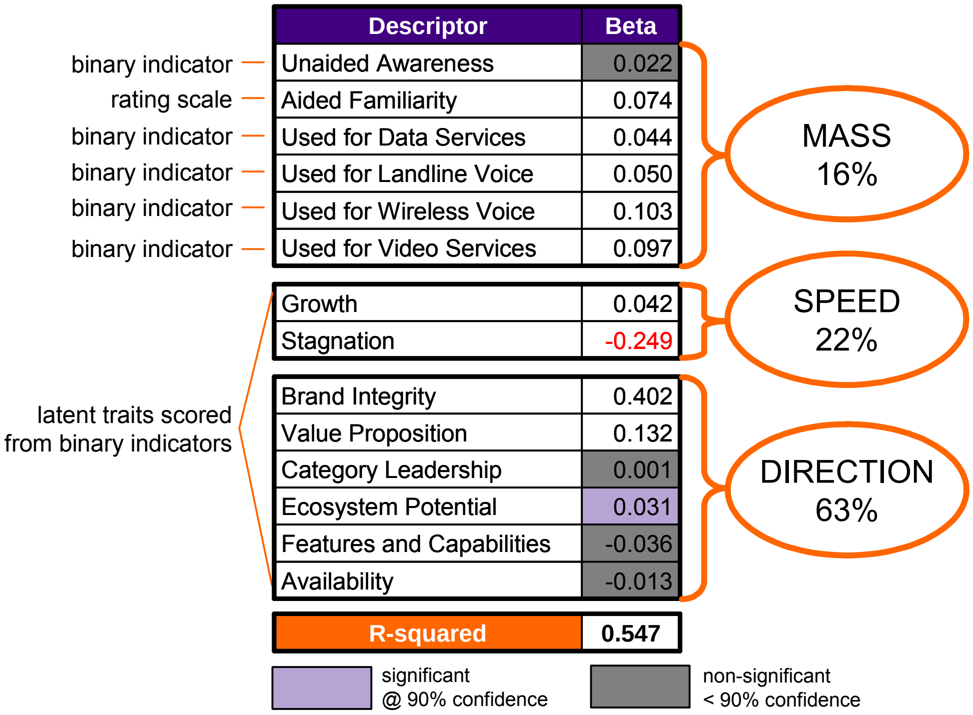


Our study contained components for a traditional brand model, because the brand observations (inputs) and brand recommendation (outcome) were asked generically across all brands, without positioning them as providers of Triple Play. Since all other brand assessments to this point in the questionnaire had been in the context of traditional roles as providers of voice, video, or data services, a model using these components is essentially asking respondents to relate brand characteristics in their familiar context.

Using this approach as a baseline, we found our traditional structure worked well, even incorporating brands from various categories together in a single model. The components of Mass, Speed and Direction explained more than half the variation in brand recommendation, with the "Brand Equity" dimension (part of Direction) having the largest standardized impact. The other primary impact came from brands that are successful at avoiding images of stagnation. These two characteristics alone are more influential than familiarity and usage, or aspects of brand leadership and performance. When service providers are competing in a familiar category, it is most important to be trusted and current, but not necessarily leading the way to growth and innovation.

These results differ somewhat when the regression is run separately for each of the five product categories. A Chow test is significant, signifying differential levels of explanation across categories, but the overall model is generally applicable to each industry with only slight interpretive variations, so the individual equations are not explored further here.

Figure 12. Results of a Traditional Brand Model Predicting Likelihood to Recommend Each Brand of Service Provider (Generic Context)



	Descriptor	Beta
binary indicator	Unaided Awareness	0.022
rating scale	Aided Familiarity	0.074
binary indicator	Used for Data Services	0.044
binary indicator	Used for Landline Voice	0.050
binary indicator	Used for Wireless Voice	0.103
binary indicator	Used for Video Services	0.097
	Growth	0.042
	Stagnation	-0.249
	Brand Integrity	0.402
	Value Proposition	0.132
	Category Leadership	0.001
	Ecosystem Potential	0.031
	Features and Capabilities	-0.036
	Availability	-0.013
	R-squared	0.547

latent traits scored from binary indicators

MASS 16%

SPEED 22%

DIRECTION 63%

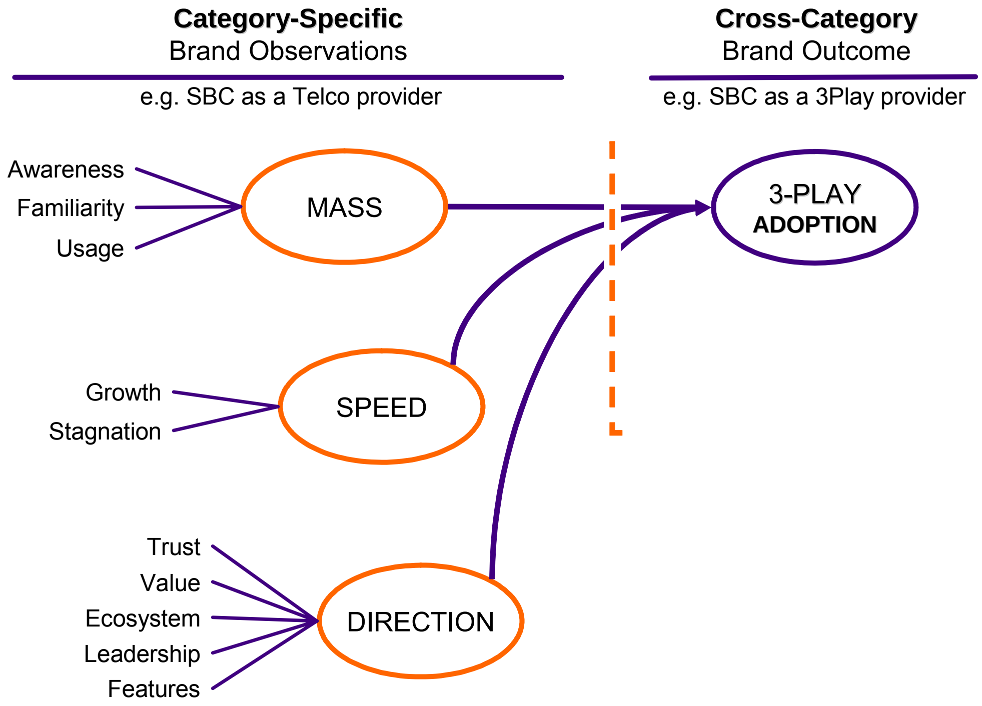
▨ significant @ 90% confidence

■ non-significant < 90% confidence

After evaluating the traditional brand model, our next step was to use the same format to predict Triple Play adoption. Asking consumers to evaluate brands in the context of an emerging category (especially one with only 4% penetration) can be problematic since respondents are unlikely to have a firm grasp on what characteristics are relevant, let alone how individual brands fare in that unknown context. Think about asking a consumer to rate Microsoft as a provider of web browsers circa 1993, and that is similar to the situation we faced in early 2005 assessing Triple Play.

Unlike a brand extension, the brand convergence embodied in Triple Play forces respondents to cross yet another chasm; not just to simply accept the brand in a new role, but also to engage with an evolving product category (particularly involving digital voice) and choose which established competitor best suits this new offering. This raises several questions as to how well traditional brand models can evaluate this scenario, and if so how the results might differ from our traditional approach. If context does affect the brand's role in the purchase process, then what might this say about how brands should approach convergent markets?

Figure 13. The Traditional Brand Model as Applied to Convergent Markets



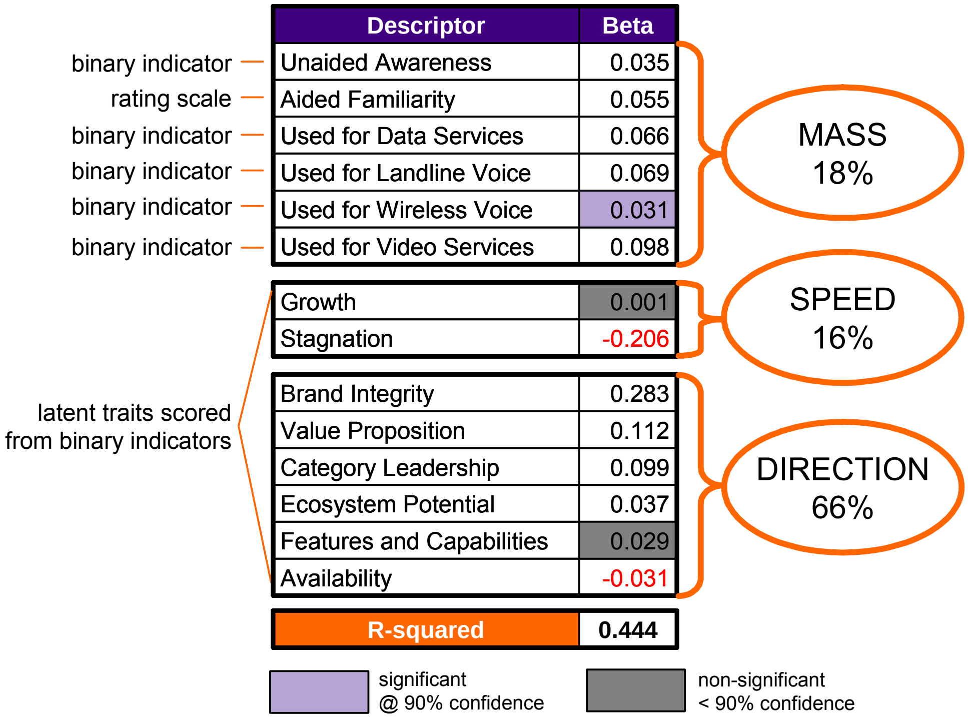
Despite asking respondents to "cross the chasm" by applying their ingrained brand perceptions to an emerging category, the traditional brand model holds up well. Brand metrics are slightly less predictive of Triple Play than they were of brand recommendation, as indicated by an r-squared that is somewhat lower (.444 vs. .547). But while this suggests the presence of other factors influencing interest in Triple Play, there are also more significant relationships observed. In a larger sense, the distribution of explanatory power across Mass, Speed and Direction is roughly comparable to the traditional model, but there is a marginally greater role for innovation in predicting Triple Play adoption.

Our initial findings, simply from changing the outcome from a generalized context to a convergent one, are:

- the impact of attributes shifts slightly, such that trust and (avoiding) stagnation are less influential relative to leadership and direct brand experience; and
- the traditional brand model is robust and applicable to Triple Play, but suggests a role for other factors to predict adoption in this convergent market.

Fortunately, our study has a number of other factors that could be incorporated into the model. As mentioned earlier, one of our hypotheses entering into this research is that the relevance of the brand's core category would be a significant contributor to predicting adoption of Triple Play; an example being that if a particular consumer doesn't find cable relevant (perhaps they avoid watching TV or live in an area not serviced by cable) then they are unlikely to be swayed by a cable company's Triple Play offer, regardless of their favorability towards the individual brand. These category attributes were applied uniformly by brand based on the *a priori* relationships discussed previously.

Figure 14. Results of a Traditional Brand Model Predicting Likelihood to Purchase Triple Play (Convergent Context)



	Descriptor	Beta	
binary indicator	Unaided Awareness	0.035	MASS 18%
rating scale	Aided Familiarity	0.055	
binary indicator	Used for Data Services	0.066	
binary indicator	Used for Landline Voice	0.069	
binary indicator	Used for Wireless Voice	🟪 0.031	
binary indicator	Used for Video Services	0.098	
latent traits scored from binary indicators	Growth	⬛ 0.001	SPEED 16%
	Stagnation	-0.206	
	Brand Integrity	0.283	DIRECTION 66%
	Value Proposition	0.112	
	Category Leadership	0.099	
	Ecosystem Potential	0.037	
	Features and Capabilities	⬛ 0.029	
	Availability	-0.031	
	R-squared	0.444	

🟪 significant @ 90% confidence

⬛ non-significant < 90% confidence

In addition to category considerations, our enhanced model also incorporated product and individual characteristics. These account for overall knowledge and interest in the Triple Play concept, as well as each individual's likelihood to adopt innovative technology or exhibit brand loyalty. This enhanced model is shown in Figure 15.

The impact of our basic brand attributes did not change substantially after adding the brand-agnostic components to the model, with integrity and stagnation remaining the two most influential individual attributes. The overall model fit improved to .524 on the basis of several significant new contributors, some of which were rather surprising.

Most notably, the impact of category consideration is nearly as important as integrity and stagnation to predicting Triple Play adoption. If you recall the incongruity discussed around Figures 4 and 5, the position of category consideration serves as a powerful equalizer in this brand model because it helps moderate some of the disconnects between brand attributes and brand-based Triple Play purchase intent. Why do Telco brands score lower than cable brands on Triple Play adoption, even though the Telco brands have higher attribute scores? Because the Telco brands are being held back by their own category. Consumers may know a lot about SBC, think they provide a good value, and exhibit technology leadership, but still be resistant to the idea of getting all their Triple Play services through the phone line.

Figure 15. The Traditional Brand Model as Applied to Convergent Markets with Brand-Agnostic Components Added

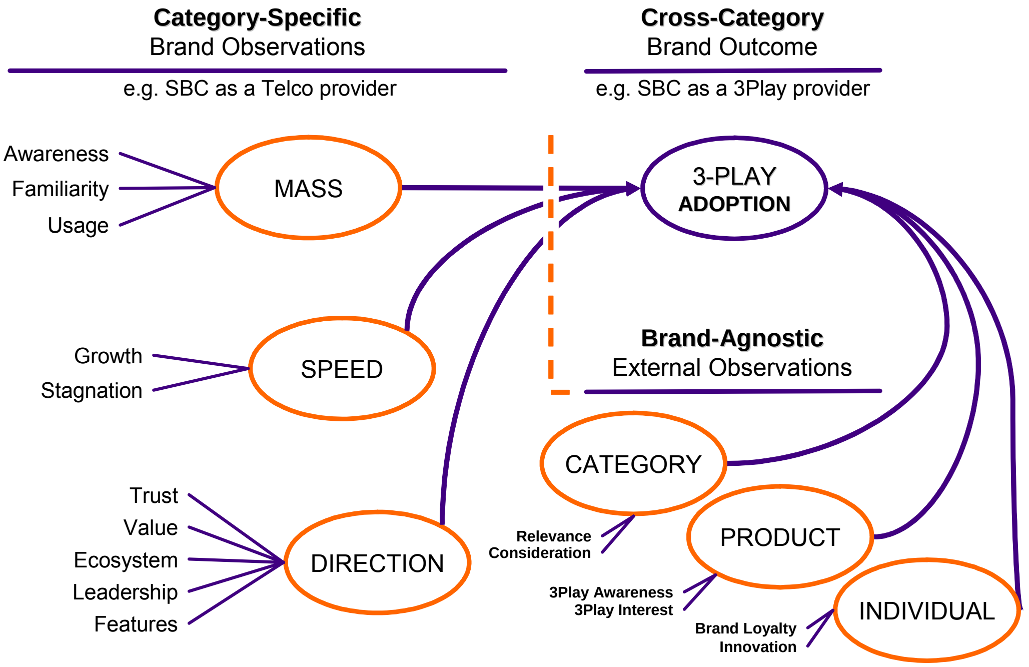


Figure 16. Results of Adding Brand-Agnostic Components to a Traditional Brand Model Predicting Likelihood to Purchase Triple Play (Convergent Context)

Mass:	**13%**	**Category:**	**18%**
Speed:	**15%**	**Product:**	**3%**
Direction:	**48%**	**Individual:**	**3%**

	Brand Descriptor	Beta
binary indicator —	Unaided Awareness	⬛ .023
rating scale —	Aided Familiarity	.033
binary indicator —	Used for Data Services	.059
binary indicator —	Used for Landline Voice	.063
binary indicator —	Used for Wireless Voice	.075
binary indicator —	Used for Video Services	.063
latent traits scored from binary indicators	Growth	⬛ .013
	Stagnation	-.205
	Brand Integrity	.285
	Value Proposition	.101
	Category Leadership	.085
	Ecosystem Potential	⬛ .001
	Features and Capabilities	⬛ *
	Availability	⬛ -.014
	R-squared	**.524**

	Category Descriptor	Beta
rating scale —	Category Relevance	-.061
binary indicator —	Viable TELCO provider	🟪 .024
binary indicator —	Viable VIDEO provider	.103
binary indicator —	Viable BROADBAND prov.	⬛ -.003
rating scale —	Category Consideration	.197

	Product Descriptor	Beta
binary indicator —	Triple Play Awareness	-.058
rating scale —	Triple Play Interest	.087

	Individual Descriptor	Beta
rating scale —	Brand Loyalty	.036
rating scale —	Innovation	.091
rating scale —	Careful Research	🟪 .027
rating scale —	Value and Quality	⬛ .002
rating scale —	Get What You Pay For	-.072

🟪 significant @ 90% confidence

⬛ non-significant < 90% confidence

* Attribute removed due to collinearity concerns

The presence of category consideration does beg the question as to how brands can influence this important attribute. Unfortunately, our questionnaire did not allow for additional time to explore the category consideration process beyond the other category attributes shown in the model. The other category descriptors (relevance, viability) had very little co-linearity with consideration, suggesting this is an area ripe for further research.

One additional equalizer is the viability of a brand to providing various aspects of the Triple Play offering. In the brand extension literature, there is a recurring theme that brands must demonstrate a "fit" with their target category. With a convergent category, things cannot be categorized quite so neatly because the candidate brands already provide some element of the offering. The question is to what degree respondents will allow each brand to take over the services currently provided elsewhere. To that end, we asked which brands would be acceptable providers of voice, video and data services, even among brands that did not currently offer such services directly. Of those three components, it was most important that a brand be perceived as a viable provider of video services. Video has been the hardest medium to deliver given current bandwidth constraints, although this barrier will diminish to the degree that IP video and further compression technology advances take hold.

In the meantime, these results indicate that brands entering a convergent market would be wise to do whatever they can to have their existing category help drive consideration for the emerging product category. One can hypothesize success by the cable industry in this regard by their ability to bundle a number of services and create a multi-media entertainment hub. Effective bundling, previous brand extensions with DVR's, responsive customer service; these are all potential areas for driving cable consideration for Triple Play. When combined with the cable category's inherent advantage as a current provider of video services, it is clear from the contextual category data why cable brands have the upper hand providing Triple Play to consumers.

One strategy that does not seem to work is to rely upon category relevance, as defined by usage frequency and perceptions of being "essential," as drivers of product adoption. The impact of our category relevance rating was statistically significant, but negative. The counter-intuitive result largely reflects that respondents use wireless and ISP services frequently and consider them to be "essential," but are unwilling to purchase Triple Play services from them. This situation is somewhat exacerbated by our inexact linkage between brands and categories, but even searching for individual brands with this relationship proved unsuccessful. Only Satellite TV brands demonstrated any substantial relationship between category relevance and Triple Play adoption, providing an indication of the strength of their brand affinity and the severity of the difference between Satellite users and non-users. One of our primary hypotheses is clearly left unsupported by the data.

DISCUSSION

There are several complicating factors that may have muddled the relationship between category and brand. Some of these areas could be addressed within the study, while others suggest the need for further research.

- Brands seldom fit neatly into discrete categories. Perceptions can change dramatically based on experience and perspective. Verizon is generally defined as a Telco, but has more growth and exposure in wireless. We controlled for this to some extent based on usage, but could have further explored respondent perceptions related to category involvement.

- Brands are in a constant state of flux. Several of the brands studied last year are no longer in existence due to M&A activity, or have shifted strategic focus into or out of our targeted product categories. This creates challenges for tracking brand identity and category affiliation.
- The VOIP product offering continues to evolve. Our client's goal was to study the Triple Play, but stand-alone services (Vonage, Skype) are working their way to customers through different channels, which has different branding implications from the bundled scenario.

There were several options for exploring the relationship between brand and category, but ultimately *ad hoc* associations proved to be the most consistent. Brand metrics correlated intuitively with category metrics: satellite brands had the strongest linkage with their category, while select Telco brands had muddied relationships.

In an attempt to capture this variation in brand-category definition, we created a ratio between each brand's strongest association and its average across all categories. When entered into the full regression model, this ratio carried a positive and significant relationship with Triple Play interest, indicating that brands with stronger category associations tended to generate higher Triple play interest.

While offering some insight into the scores for Satellite brands, this result also seems counter-intuitive to the need to establish video credibility while also offering data and voice services. It was held out of the final model due to its arbitrary calculation and uncertain implications, but the concept of "focus" seems a worthwhile one to include in future research regarding convergent branding.

Several questions and issues arose during this analysis about the best way to model the data. I chose to process and present the results based on linear regression because it effectively and efficiently captured the baseline relationships, but there are certainly other approaches available.

- Linearity (or lack thereof) within the component attributes. Rating scales have questionable interval properties which can skew results. Since this analysis relies on several scaled metrics, this issue warrants further examination.
- Respondent-level heterogeneity. Some of the issues surrounding brand definition and category impact could be resolved by allowing individual-level relationships. Comparing aggregate results against HB/REG utilities might improve prediction and reveal underlying relationships obscured at the aggregate level.
- Potential relationships between category and brand. Given the hierarchical nature of category and brand, there is the opportunity to explore more complex structure involving the given inputs and outputs of the Triple Play model.

A Categorical Regression of the various Triple Play models revealed that there were examples of non-linearity present within the data. In some attributes, the CATREG attributes led to a re-coding (typically a simplification) that restored linearity. Even allowing for these issues, however, the CATREG produced nearly identical results to the linear regression, both in terms of model fit and coefficient strength.

An individual-level model with HB/REG produced results similar to the aggregate regression, but with somewhat stronger brand effects relative to category effects. However, the individual coefficients have yet to shed light on additional structures for exploration. Given the relative lack of individual-level observations, HB/REG seems less than ideal in this application.

The hierarchical alternatives (SEM, HLM) seem to offer the best hope for additional insights, and are the subject of ongoing analysis. However, any hierarchical model will be beset with the same issues discussed elsewhere – that is, the complex (and indeterminate) relationship between category and brand.

CONCLUSION

This approach demonstrates that brands are influenced by the context of their purchase scenario. But the nature of this influence – in terms of content and magnitude – is quite different than was hypothesized. While broader category and individual characteristics have a significant impact on adoption, brand-level information is clearly the dominant influence on Triple Play purchase intent.

If we consider brand influence to be a continuum between purchase driver and universal corporate asset, this study suggests it falls somewhere in the middle. Two brand attributes (Integrity / Stagnation) remain the predominant contributors to both within-category and cross-category outcomes. Even though Leadership becomes more important for the new category, there is a consistent relationship between branding characteristics and (stated) interest even when the reference category changes.

The primary caveat here is that, while evolving rapidly, the Triple Play category is still not that far removed from its component pieces (voice, video and data). In our scenario, brands seem to have a relatively easy time transferring equity from a core category to a partially-related category. However, brands are unlikely to benefit from this same effect if the category were farther afield, or if outsider brands were competing amongst entrenched service providers.

In this respect, Triple Play has some of the same characteristics of a brand extension. The same concepts of brand quality and category fit that are dominant in the brand extension literature appear in this study as well in the guise of brand integrity and the category viability attributes. But it is the convergent aspect of Triple Play, bringing established brands together in an evolving market, that differentiates this phenomenon from brand extensions. I believe two key components – the impact of innovation, both as a brand attribute as well as a personal descriptor, and category context – have distinctive influence in the Triple Play purchase scenario, and perhaps other technologically convergent markets as well.

Surprisingly, category context has little to do with our Category Relevance measure (a combination of "usage" and "essential" ratings), which had no meaningful impact on Triple Play adoption. In fact, it has a negative coefficient due to the strong relevance and low Triple Play interest among wireless and service provider brands.

Instead, relevance is dictated by the perceived ability of the category to consolidate services ("category consideration" ratings). Of the Triple Play components, video delivery is by far the most important, both in terms of current usage and anticipated viability. In contrast, voice and data delivery are significant contributors if currently obtained from a brand, but less relevant at the category level. Cable companies have a clear advantage (reflected in the ratings) due to their core positioning around video and relative lack of structural competitors. There is more direct competition in voice and data services, although data delivery is a key component to VOIP (and thus Triple Play) adoption.

Any brand which hopes to succeed in a convergent market should examine the role of its core category, particularly with reference to a consumer's willingness to consider any brand from that category as a provider of the convergent service. If that is not the case, then the traditional brand extension drivers such as quality and fit will be sabotaged by the brand's broader affiliation with a category that has failed to gain access to the consumer. The rapid pace of technological change, while rapidly becoming a cliché in product development, dictates that convergent technologies will continue to evolve and bring established brands into competition in evolving markets. In these scenarios, brands are influenced by perceptions and usage of the category with which they are associated, and these category characteristics should be considered an integral part of brand modeling and brand communication when entering into any convergent market.

REFERENCES

Aaker, David A. and Kevin Lane Keller (1990) "Consumer Evaluations of Brand Extensions," *Journal of Marketing* Vol. 54 (January): 27-41.

Moore, Geoffrey A. (1991) *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*, New York: HarperCollins.

Sattler, Henrik, Franziska Volckner, and Grit Zatloukal (2002), "Factors Affecting Consumer Evaluations of Brand Extensions," *Research Papers on Marketing and Retailing No. 010*, Hamburg: University of Hamburg.

BRAND POSITIONING CONJOINT: A REVISED APPROACH

CURTIS L. FRAZIER, URSZULA JONES AND KATIE BURDETT
MILLWARD BROWN

At the 2003 Sawtooth Conference in San Antonio, Millward Brown presented our initial approach to integrating conjoint with brand positioning. This three-stage model usually worked well, but allowed for a significant amount of error to enter the models. This paper will present a revised, single-stage approach illustrating a concise method to improve model fit.

BACKGROUND

Traditionally, optimization studies have taken one of two distinct tracks. The research has focused on either optimizing the product/price or optimizing the positioning of the brand. For product and pricing research, various trade-off approaches predominate. Brand positioning work is most often done through a variety of regression-based approaches, whether these are simpler linear models or more complex structural equations or time series approaches.

The problem is that the two approaches largely stand in isolation and are not integrated in a way that allows for a concrete assessment of the relative ROI of focusing on brand or product/pricing.

It should be noted at this point that while the approach was initially created to deal with branding, the approach can (and has) been extended to other concepts. For example, we have successfully used the approach to predict the conjoint utilities of concepts from HDTV technology preference to birth control methods.

THE ORIGINAL MODEL

The model presented at the 2003 Sawtooth Conference used a 3-stage approach. These three stages were:

1. Estimate a conjoint model (could be discrete choice or ratings-based conjoint) using HB to get individual-level brand utilities.
2. Using the brand utilities as a dependent variable in a regression model, with brand image attributes as exogenous predictors of the brand utilities. This regression model was initially tested using OLS, but later extended to latent class regression and HB-Regression models. These techniques resulted in better fitting models, but also sometimes resulted in over-fitting of the data.
3. Integration of the two primary models through a market simulator. Essentially, simulated changes to brand positioning would add or delete utility from the brand utilities, which would impact the share of preference estimates.

PROBLEMS WITH THE ORIGINAL MODEL

The 3-stage process was found to be quite effective—most of the time. The second stage regression models were oftentimes found to be quite predictive of the brand utilities.

Unfortunately, there were a number of cases where the models ranged from merely satisfactory to poor. We believe that the models would fail for three reasons:

1. When controlling for a significant number of product attributes in the conjoint, the brand utilities may have little meaning left. In one particular example for an online bookseller, we feel that once we controlled for attributes such as book price, shipping speed, shipping cost, book selection etc., the brand positioning stage failed because there would be little to differentiate the brands.
2. Choosing the wrong attributes. The regression model in the brand positioning construct is no different from standard regression models in that it can suffer from things such as poor input choice, multicollinearity, etc.
3. By modeling brand strength using estimated brand utilities as our dependent variable, we are allowing error to enter our model at two stages—the estimation of the utilities and then the estimation of the regression model.

The new approach is primarily aimed at resolving issue #3, though we believe it may also address issue #1.

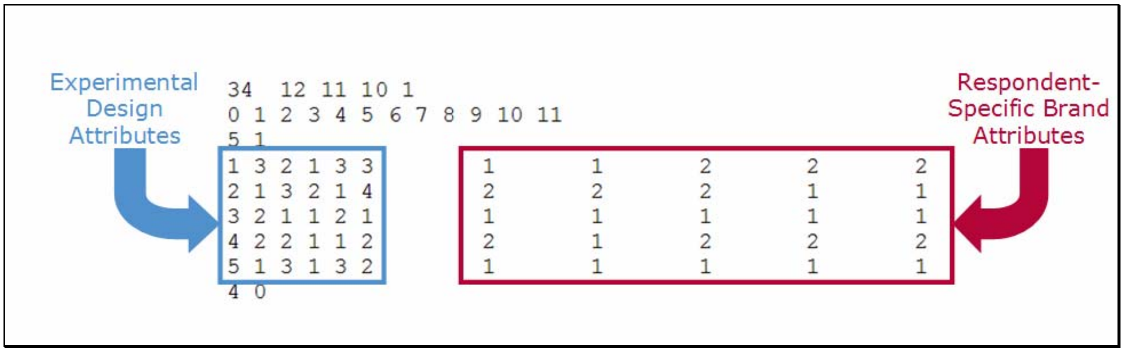
THE NEW MODEL

Based on our experience (and comments by Rich Johnson at the 2003 conference), we decided that a revised, perhaps simplified, approach was needed. Our revised approach is based on a simple assumption:

- We are assuming that when a respondent in a conjoint sees a particular brand, he or she is also *seeing* a host of other, "hidden" attributes that describe that brand. For example, a respondent who sees a choice set with Sony as one of the brands allow *sees* that the product is, perhaps, "a good value" and "a brand I trust."
- Another example might be that when the respondent sees Heineken, the respondent is also seeing "imported."

If we accept this assumption, then we can go forward with the approach. The approach specifies that the brand image data are, in fact, part of the experimental design. The primary difference is that the levels seen in each choice task are determined by the respondent. That is, respondent #1 believes that Centrum is "available where I shop" so that level becomes part of his design, while respondent #2 says the opposite.

By accepting this assumption, we can create a CHO file that includes both the true experimentally designed attributes and the user-defined part of the design.


Experimental Design Attributes
Respondent-Specific Brand Attributes

```
34   12  11  10  1
0 1 2 3 4 5 6 7 8 9 10 11
5 1
1 3 2 1 3 3     1       1       2       2       2
2 1 3 2 1 4     2       2       2       1       1
3 2 1 1 2 1     1       1       1       1       1
4 2 2 1 1 2     2       1       2       2       2
5 1 3 1 3 2     1       1       1       1       1
4 0
```

In the figure above, we can see that the respondent-specific brand attributes are coded as dichotomous variables. While there is some flexibility here, we have been setting the brand attributes as 2-levels, with the "yes" value coded as a 2 and the "no" coded as a 1.

Impact of Including Brand Image Attributes in CBC

The primary impact of including the brand image attributes in the CBC model is that the brand attributes account for a significant amount of variance that otherwise would have been assigned to the experimentally designed brand attribute. Therefore, brand utilities and the attribute importance of brand in a brand positioning conjoint will be artificially reduced.

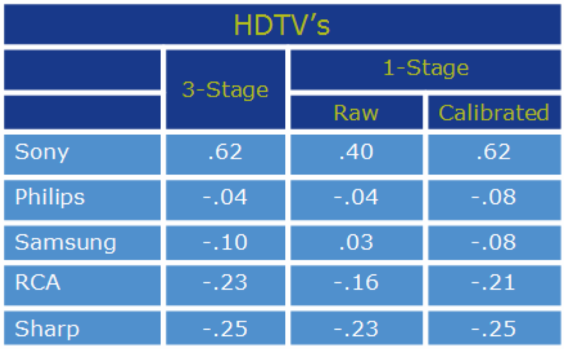
However, the utilities (and hence, the importances) of brand can be recreated to match the results from a conjoint that does not include the brand positioning attributes in the conjoint estimation. The process is straightforward:

Calibrated Brand Utility = Raw Brand Utility + $B_{Image1}X_{Image1} + B_{Image2}X_{Image2}\ldots$

For example, if Respondent #1 has a raw brand utility for Sony of 0.5, but believes that Sony is "reliable" and "an expensive brand," we would add in the positive utility for *reliability* and the (likely) negative utility of *expensive* to that respondent's Sony utility to get the calibrated value.

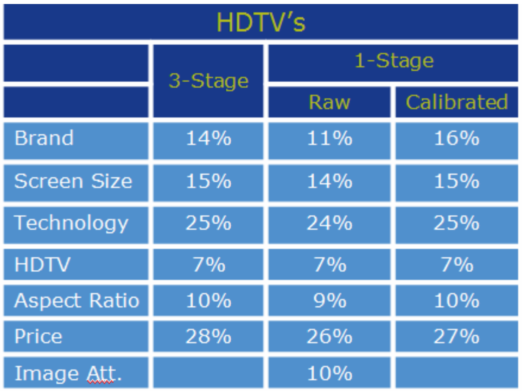
Essentially, what this equation does is add the variance that is being accounted for by the brand image attributes back into the brand utility.

Tests have shown that this simple calibration process results in brand utilities (and attribute importances) that are nearly identical to the estimates that would come out of a model that did not include the brand positioning attributes (see HDTV example below).



HDTV's

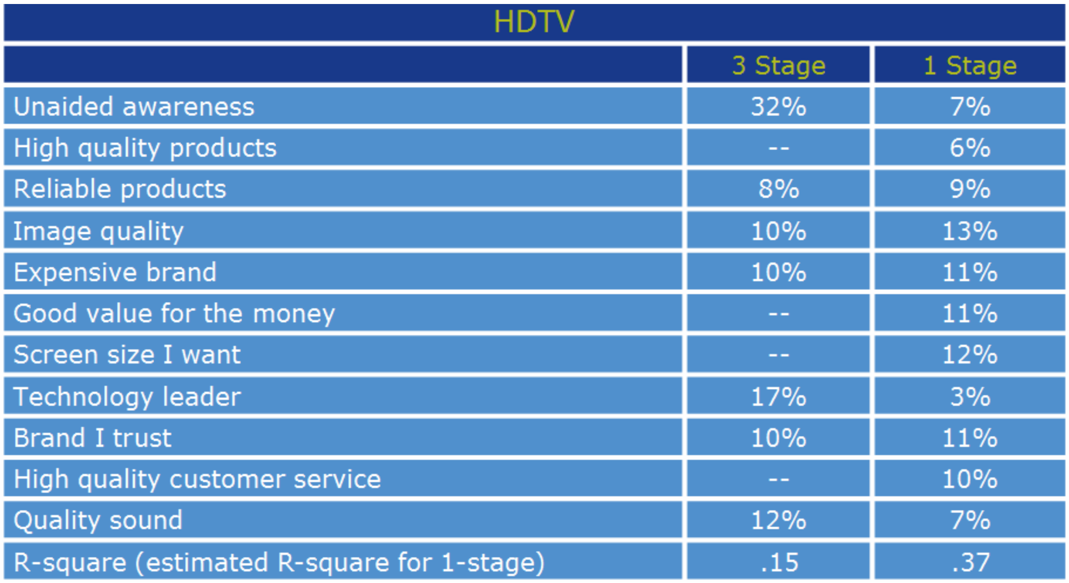
	3-Stage	1-Stage	
		Raw	Calibrated
Sony	.62	.40	.62
Philips	-.04	-.04	-.08
Samsung	-.10	.03	-.08
RCA	-.23	-.16	-.21
Sharp	-.25	-.23	-.25



HDTV's

	3-Stage	1-Stage	
		Raw	Calibrated
Brand	14%	11%	16%
Screen Size	15%	14%	15%
Technology	25%	24%	25%
HDTV	7%	7%	7%
Aspect Ratio	10%	9%	10%
Price	28%	26%	27%
Image Att.		10%	

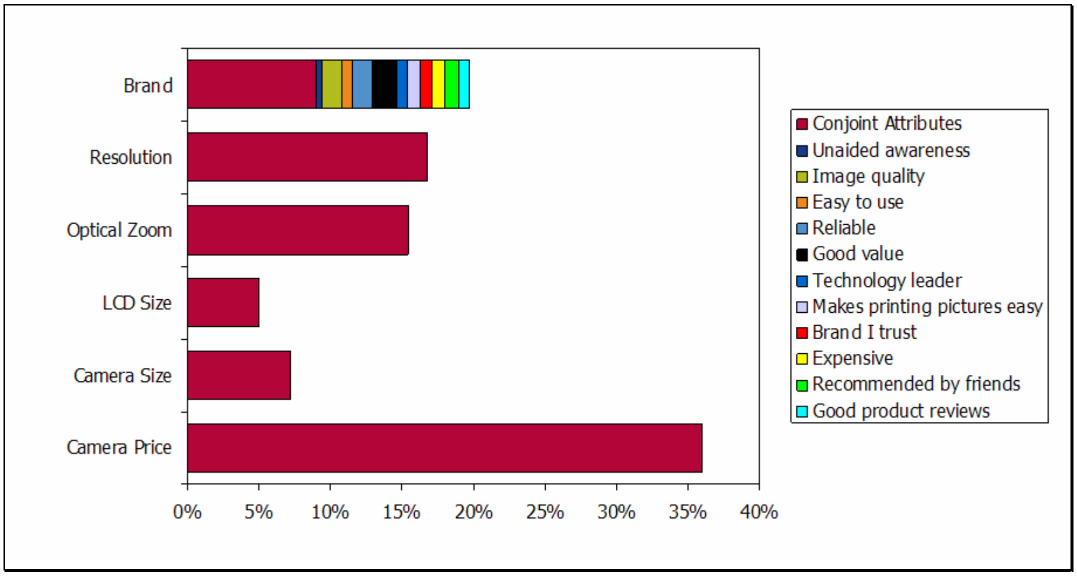
The other important impact of the new approach is the tighter linkage between the image attributes and choice. In the original approach, only a handful of attributes would be found to be statistically significant drivers of brand utility (though the number is greater with regression methods accounting for respondent heterogeneity). With the revised approach, many more attributes were found to be drivers.



HDTV

	3 Stage	1 Stage
Unaided awareness	32%	7%
High quality products	--	6%
Reliable products	8%	9%
Image quality	10%	13%
Expensive brand	10%	11%
Good value for the money	--	11%
Screen size I want	--	12%
Technology leader	17%	3%
Brand I trust	10%	11%
High quality customer service	--	10%
Quality sound	12%	7%
R-square (estimated R-square for 1-stage)	.15	.37

Using the calibration approach, it is possible to construct a measure that is analogous to an R^2. The "estimated R^2" is calculated as the difference between the calibrated and raw utilities. Using this estimate, we can parse out the impact of the image attributes, by proportionalizing the parameters as a percentage of the R^2.



CONCLUSIONS

Brand positioning conjoint is effective at allowing marketers to understand optimizing their offerings in terms of both product/pricing and brand positioning. This allows product managers to estimate ROI of a larger set of potential business decisions.

The revised approach resolves the primary problem associated with the initial approach (the multiplication of error by estimating the regression model using error-laden brand utilities as a dependent variable).

With some relatively simple back-end manipulations, the revised approach comes very close to replicating the utilities and importances from a standard choice model.

RETHINKING (AND REMODELING) CUSTOMER SATISFACTION

LAWRENCE KATZ
IFOP (PARIS)

Surveys of customer satisfaction have long constituted a major part of many research companies' commercial activity. If a study were done, however, on these same companies' own customers, the results might not be very encouraging. Too frequently, what begins as a projected long-term collaboration between researcher and client aimed at understanding and tracking customer satisfaction ends in a divorce by mutual consent, the client finding the results too static over time, too far removed from everyday managerial problems and too expensive, while the research company has little in the way of alternatives to counter these criticisms.

In what follows, we analyze various reasons for these problems and attempt to propose a way out. Basically, our argument will hinge on two premises. The first is that there is often a basic misunderstanding concerning the goals of satisfaction research; that understanding satisfaction's structural determinants is, for various methodological and theoretical reasons, incompatible with tracking its short term fluctuations. Secondly, the models and methods currently available to researchers create a propensity towards structural analyses which, though interesting to clients at first, are too stable over time to be useful as a tool for tracking. Several successive waves of a satisfaction barometer are thus often sufficient for convincing clients that their research resources can be better spent elsewhere.

Our focus here is on customer satisfaction with services (banks, insurance, telephone, transport, etc.), a domain in which, despite lively competition among providers in certain sectors, is characterized by an important degree of inertia in the customer base, especially when compared to fast-moving consumer goods. "Churners," while of utmost commercial importance, are a relative rarity, the majority of customers usually preferring to stay put and to avoid the hassle of changing once that they have chosen a provider. High levels of satisfaction and low levels of customer implication are thus common and, as we will discuss below, make both measuring and modeling satisfaction especially difficult. The problems that we discuss, however, are not unique to this domain but occur, in varying degrees, whenever measuring and modeling satisfaction are an issue.

DEEP OR SURFACE SATISFACTION?

Consider the results in Table 1 taken from a typical satisfaction study done for a service provider, in this case, say, a Conference Center[1]. The table contains the correlations between two overall satisfaction ratings—one measured at the beginning of the questionnaire and the other at the end—and 12 other "intermediate" ratings concerning the satisfaction with specific aspects and features of the service under study. The items' relative positions in the questionnaire are indicated by their item numbers. The time interval between the two overall satisfaction ratings was approximately 20 minutes.

[1] The descriptions and the item labels in the examples have been modified to preserve confidentiality.

TABLE 1
Correlations between Intermediate and Overall Satisfaction Ratings Placed at the Beginning and End of the Questionnaire

	Overall Rating	
ITEM	Beginning	End
Q13: Overall Beginning	1.000	0.384
Q15: Cleanliness	0.436	0.312
Q17: Interior Space	0.449	0.299
Q21: Parking	0.455	0.367
Q21: Security	0.298	0.213
Q25: Surroundings	0.232	0.134
Q29: Payment	0.204	0.205
Q34: Personnel	0.315	0.545
Q43: Decoration	0.200	0.428
Q46: Overall End	0.384	1.000

There are several things to notice about these findings, things that are characteristic of satisfaction studies structured in this way:

- the correlations between the two overall satisfaction ratings and the intermediate items show a marked tendency to diminish as a function of the interval (in time and in number of items) that separates them in the questionnaire
  - for example, the correlations between the initial overall rating and the first 3 items that follow it in the questionnaire are 0.436, 0.449 and 0.455 whereas its correlations with the final three intermediate items are 0.204, 0.315 and 0.200,
  - similarly, the final overall rating is more highly correlated with the items towards the end of the questionnaire than it is with those at the beginning.
- The correlation of 0.384 between the two overall satisfaction ratings—ratings that are supposed to be measures of the same thing—is far from 1.00 and only of moderate strength when compared with the other correlations in the table.

While these results raise a number of methodological issues, they also carry certain implications concerning the theoretical underpinnings of satisfaction research. For if the observed drift in the correlations across contiguous items is symptomatic of research designs in which (either by negligence or by impracticality) there is no rotation in the order in which items are administered, it also constitutes grounds for questioning the very nature and usefulness of the construct "satisfaction" that we are supposed to be measuring. If satisfaction was something that was psychologically well-grounded, we would not expect to see this sort of sequential drift. And if they are so unstable and fluctuating, of what use can our satisfaction measures be to our clients as a basis for managing their activity?

More interesting, however, is the weak correlation between the two *overall* satisfaction ratings. Even if one allows for the fact that the intervening questionnaire items can influence subjects' judgments (cf. Strack & Martin, 1987; Tourangeau & Rasinski, 1988), the correspondence between the two measures is low enough to lead us either to question their internal validity or to simply conclude that the two scales are measuring different things.

It is the second of these alternatives that seems to us to be the most compelling. Rather than conceive of the two overall ratings as being two imperfect measures that converge towards a common underlying construct of “overall satisfaction,” we will argue in what follows that the two scales measure two things— “surface” and “deep” satisfaction—that are fundamentally different both with respect to the psychological processes upon which they depend and their relevance and utility for marketing.

MONITORING OR MODELING?

Table 2 contains the complete analysis of the study presented above, framed in terms of the usual additive compensatory model: overall satisfaction with the conference center being analyzed in terms of the satisfaction elicited by 34 detailed features that compose it. The grouping of features was determined by a pilot phase and held constant across each wave of what was an annual barometer. The links between the groups of detailed items and the intermediate items, and between the intermediate items with the overall rating, were calculated by series of regressions, the weights transformed so as to sum to 1.00 within each group. The resulting schema forms a sort of “causal tree,” the weights allowing us to compare the relative importance of each intermediate “branch” in the determination of the overall rating, and within each branch, the importance of each of the more detailed characteristics. The relative importance of each detailed characteristic with respect to the overall measure can be calculated by multiplying the weights on the path that separates them (cf. Bachelet & Lion, 1988).

TABLE 2
Conference Center: Overall Satisfaction as a Function of Satisfaction with Component Attributes

	Level 1 Weights			Level 2 Weights		
	Raw	Normed		Raw	Normed	
Q5.1	0.337	**0.288**	**Q15** **Cleanliness**	0.195	**0.319**	**Q13** **Overall Satisfaction**
Q5.2	0.263	**0.224**				
Q5 .3	0.232	**0.198**				
Q5.4	0.200	**0.171**				
Q5.5	0.140	**0.119**				
Q16.1	0.140	**0.144**	**Q17** **Interior Space**	0.101	**0.165**	
Q16.2	0.220	**0.227**				
Q16.3	0.130	**0.134**				
Q16.4	0.180	**0.186**				
Q16.5	0.300	**0.309**				
Q18.1	0.290	**0.305**	**Q19** **Parking**	0.105	**0.171**	
Q18.2	0.220	**0.232**				
Q18.3	0.260	**0.274**				
Q18.4	0.180	**0.189**				
Q20.1	0.282	**0.221**	**Q21** **Security**	0.155	**0.254**	
Q20.2	0.162	**0.127**				
Q20.3	0.361	**0.283**				
Q20.4	0.470	**0.369**				
Q22.1	0.500	**0.515**	**Q22** **Surroundings**	0.000	**0.000**	
Q22.2	0.470	**0.485**				
Q28.1	0.124	**0.111**	**Q29** **Ticketing, Access**	0.015	**0.025**	
Q28.2	0.000	**0.000**				
Q28.3	0.312	**0.278**				
Q28.4	0.166	**0.148**				
Q28.5	0.349	**0.311**				
Q28.6	0.170	**0.152**				
Q30.1	0.226	**0.246**	**Q30** **Personnel**	0.016	**0.026**	
Q30.2	0.161	**0.176**				
Q30.3	0.244	**0.266**				
Q30.4	0.149	**0.162**				
Q38	0.137	**0.149**				
Q42.1	0.197	**0.195**	**Q43** **Interior Decor**	0.024	**0.039**	
Q42.2	0.372	**0.368**				
Q42.3	0.441	**0.437**				

Though the absence of a rotation left us no choice in this case but to base the model on the first overall rating, it would have been theoretically more coherent to have used the measure placed at the end of the questionnaire, once all the pertinent detailed and intermediate aspects of the service had been made salient to the subjects' minds. But while using what is certainly a more reasoned, cognitively-grounded rating would be more in line with the model, the “deep” satisfaction analyzed in this way can, as we have seen above, often be quite different from what people feel and, if questioned, express spontaneously in the field. The best choice of dependent measure thus apparently depends on whether we want to monitor or model customers' satisfaction.

This does not mean that monitoring and modeling are irreconcilable, but it does suggest that when tracking current, spontaneously felt satisfaction is one's primary aim, the usual additive model and the methods it engenders are less than optimal. As we have just said, the former is psychologically implausible, for even if we assume—as the model implies—that overall judgments repose on the assessment of the component parts, we know that this assessment can at best be partial, limited to the few points the most salient to the subjects' minds. Not only will salience be confounded with importance in the weights that we calculate for the component services, but the method based on item ratings gives us no means of measuring salience itself and thus deprives us of additional information that could be of great interest to managers.

There are also practical reasons for changing one's approach. In order to provide for a sufficient degree of reactivity on the part of management, satisfaction monitors should be frequent and fast, the aim being to come as close as possible to give a continuous record of the level of satisfaction currently felt by the client base. Lengthy questioning during which the entire set of component services are passed in review is just too cumbersome and costly to be repeated often enough to be of sufficient use to managers. While such procedures provide valuable information for establishing medium to long term priorities, an alternative procedure is needed for tracking satisfaction's dips and rises and rapidly detecting problems before their consequences become too serious.

ASYMMETRIC COMPONENT EFFECTS

Before going any further, there is one other aspect of the usual additive model that needs to be considered. As we have seen above, at the core of the model is the proposition that the overall satisfaction with a service or product is a cumulative function of the satisfaction felt towards its components. Each component's contribution is thought to be linear and symmetric in form and weighted by the component's importance. Thus, the shortest path to maximizing satisfaction is to guarantee that people are highly satisfied on the components that are most important to them, and any eventual shortcoming on one component can be compensated for by improvement on any one or several others.

Several authors have challenged these points (Herzberg *et al.*, 1959; Amstutz, 1970; Kano *et al.*, 1984; Mittal *et al.*, 1998), and a recent family of approaches called Penalty/Reward Analysis retains the basic additive model while relaxing the constraints on symmetry. Llosa (1997), for example, has developed an interesting approach based on correspondence analysis which separately evaluates each component's role as both a source of added satisfaction and added *dis*satisfaction. Since nothing forces these two roles to be equivalent, the form of each component's overall contribution is free to vary, thus enabling one to differentiate between the components according to whether they act asymmetrically or in the "traditional" monotonic manner.

As an example, consider the findings in Table 3 drawn from a study of client satisfaction done for a bank. Responses to a 5-point semantic satisfaction scale were first recoded to form a binary "Satisfied/Not Satisfied" variable for each component service which were then fed into a multiple correspondence analysis. As the resulting coordinates readily reveal, the first dimension that emerged from the analysis separated the modalities "Satisfied" and "Dissatisfied" for each component and can thus be considered as a dimension of General Satisfaction.

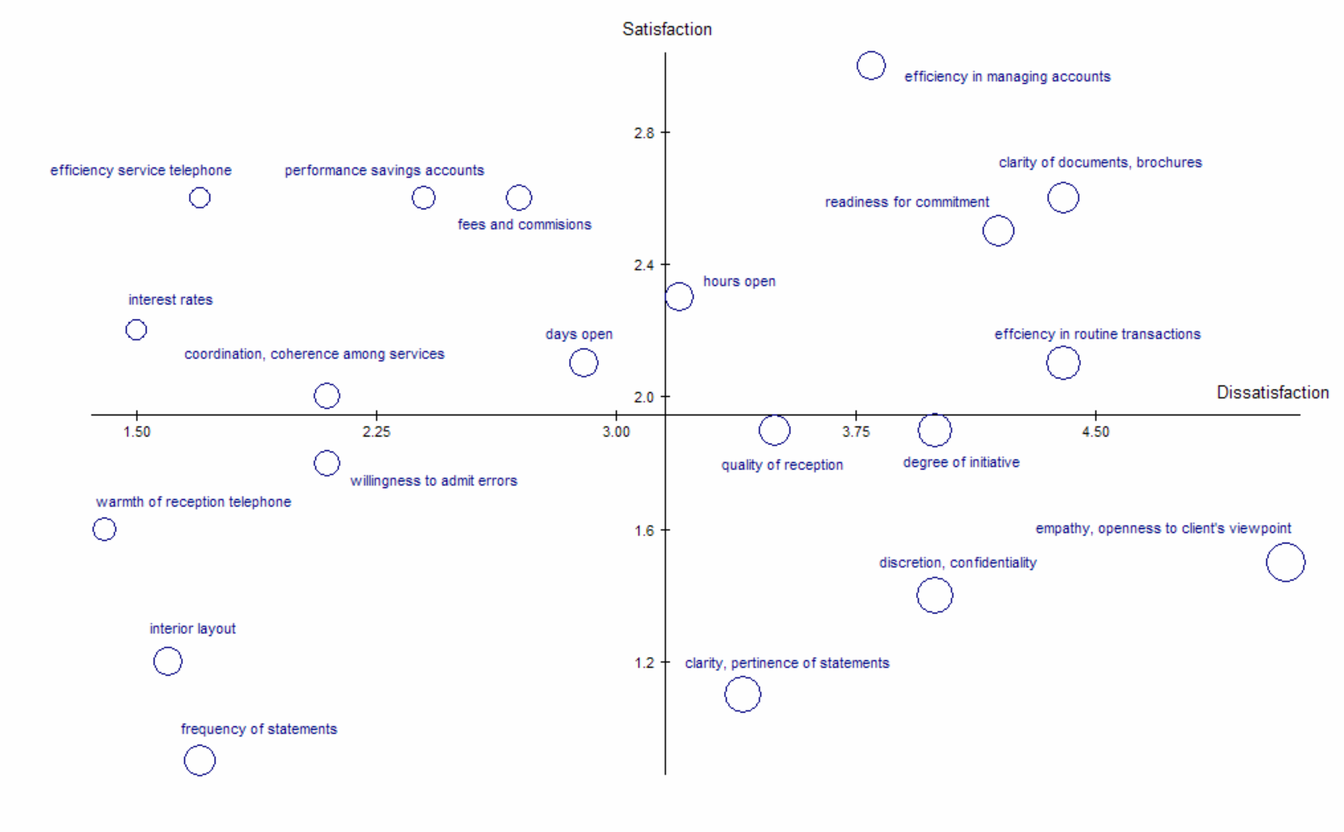
TABLE 3
Bank Satisfaction: Example of Llosa's(1997) Method)

Item	% Satisfied	Coordinates		Contributions	
		Dissatisfaction	Satisfaction	Dissatisfaction	Satisfaction
Interior layout	57	-0.48	0.37	1.6	1.2
Employee helpfulness	65	-0.79	0.42	3.5	1.9
Efficiency routine transactions	67	-0.91	044	4.4	2.1
Discretion, Confidentiality	74	-0.97	0.35	4.0	1.4
Clarity	63	-0.87	0.50	4.4	2.6
Effective account management	56	-0.73	0.57	3.8	3.0
Coordination, coherence among services	51	-0.51	0.49	2.1	2.0
Friendliness reception telephone	48	-0.41	0.45	1.4	1.6
Efficiency service telephone	40	-0.43	0.64	1.7	2.6
Days open	58	-0.65	0.48	2.9	2.1
Hours open	57	-0.68	0.51	3.2	2.3
Empathy, openness to client's viewpoint	77	-1.18	0.34	5.1	1.5
Readiness for commitment	62	-0.84	0.50	4.2	2.5
Readiness for initiative	68	-0.88	0.41	4.0	1.9
Willingness to admit errors	54	-0.53	0.46	2.1	1.8
Interest rates loans	40	-0.40	0.59	1.5	2.2
Performance savings accounts	49	-0.55	0.57	2.4	2.6
Fees, commissions for routine services	51	-0.59	0.56	2.7	2.6
Frequency of statements	64	-0.54	0.30	1.7	0.9
Clarity, pertinence of statements	76	-0.95	0.30	3.4	1.1

In the present case, however, it is the *contributions* rather than the coordinates that interest us the most, since they provide a measure (comparable to regression coefficients) of the degree to which the satisfaction and dissatisfaction felt with respect to each component participates in the constitution of this first dimension. Comparing the contributions should thus enable us to determine not only the degree but also the form (linear or asymmetric) of each component's impact on the overall satisfaction.

The first thing to notice is that most of the components impinge asymmetrically upon the overall satisfaction, and that dissatisfaction generally weighs more heavily than satisfaction. This is something that we have found to be generally the case for studies of satisfaction in the services (a domain in which samples tend to consist of mostly satisfied clients of at least moderately successful providers), and which confirms earlier findings by Mittal and his colleagues (e.g. Mittal *et al.*, 1998).

FIGURE 1
Bank Survey: Determinants of Satisfaction and Dissatisfaction



NB: The size of the points is proportional to the degree of satisfaction expressed with respect to the corresponding item

Using the contributions as coordinates for positioning the components in a two-dimensional space conveniently divides the components into four "fuzzy" classes according to the way that they contribute to the general level of satisfaction:

- Drivers: components that impinge linearly on general satisfaction, making relatively substantial contributions both in the case of satisfaction and dissatisfaction: e.g. efficiency in managing accounts, clarity of documents and brochures,
- Basic Expectancies: components that operate primarily as sources of dissatisfaction and for which increasing the level of satisfaction will have relatively little impact: e.g. willingness to listen, discretion
- Bonuses: components that are primarily sources of added satisfaction and that have relatively little negative impact when judged as being not satisfactory: e.g. efficiency service telephone, savings account performance
- Secondaries: Components that impinge less than the others on the overall satisfaction both in the positive and negative sense: e.g. interior layout, frequency of statements

Note that only the Drivers and Secondaries behave in the way prescribed by the usual, regression-based approaches based on the compensatory model and that ignoring the differences between the ways that components affect satisfaction may deprive managers of nuances that could prove valuable when fixing their priorities. Beyond a minimum threshold, devoting effort to improving satisfaction on basic expectancies may bear diminishing returns when compared to the same effort devoted to a Driver or a Bonus. In the same way, raising low scores on Bonuses should be less a priority than correcting deficits in Drivers or Basic Expectancies.

Surface Satisfaction as a Cognitive Heuristic

While relaxing the constraints on symmetry render the additive model more in line with how people may actually form judgments of satisfaction, Penalty/reward approaches such as Llosa's still suffer the same drawbacks mentioned above: the necessity of a long interview, absence of information concerning salience, and the possible confounding of salience and importance. But in order to develop a method more suited to our purposes, we first need an alternative theoretical model.

Useful in this respect is the notion of "cognitive heuristic" proposed by Kahneman and Tversky. According to these authors, many judgments (e.g. one's degree of satisfaction) made routinely in every day life are based on a set of simple and efficient rules that they call "heuristics" which can be thought of as strategies or shortcuts that permit people to carry out what might otherwise require complex thought processes in simpler ways that economize cognitive effort.

Among the different heuristics described by Kahneman and Tversky, one that they call "anchorage and adjustment" seems particularly useful for understanding spontaneous judgments of satisfaction. An "off the cuff" response to the question "How satisfied are you with X ?" would, according this analysis, involve two-steps: a first "anchorage" phase in which the respondent formulates a preliminary response on the basis of the context in which he/she finds him/herself with respect to X ("I am presently a client and have not thought about changing, so I guess I'm satisfied"), and a second phase in which the person "adjusts" this first response as a function of any experiences with X that may be salient to his mind at that moment.

The differences with respect to the approaches discussed above are fundamental. Most important is that the heuristic-based approach takes as a given the superficiality of spontaneous judgments and does not make exaggerated claims concerning the cognitive basis upon which they repose. The anchoring of the preliminary response is assumed to be done with a strict minimum of introspection, the respondent *inferring* his degree of satisfaction from his current situation much as he would for a third person.[2] The response at this stage may, as a consequence, reflect the person's disposition and comparison levels more than it does the characteristics of the service in question.

In the following "adjustment" phase, the things that modulate the preliminary judgment are not themselves abstract, more detailed judgments but rather recollections of concrete experiences that the person has had (or has heard of from others) with the service. And rather than an exhaustive review, the things that come to mind may mostly concern only a small number of the service's various components, and this with a high degree of thematic redundancy between the things cited.

[2] A similar argument was advanced by Daryl Bem (Bem, 1967) with respect to the cognitive foundations (or lack of them) of attitudes in general.

Furthermore, one can also suppose that much of what comes to mind when "adjusting" the preliminary response will depend on the respondent's expectancies concerning the service, with disappointed basic expectancies and unexpected bonuses being particularly present among the things recalled. Non-monotonic relations between overall and detailed, component-level satisfaction may be the rule rather than the exception.

Modeling Surface Satisfaction

The two stage anchorage and adjustment heuristic opens the way, with a few very simple changes in the usual satisfaction questionnaire, to doing more than simply monitor satisfaction, and this without doing too much violence to what we know about cognitive function or our methodological good sense.

As an example, consider a study done to evaluate private and business customers' (N=300 and 108, respectively) reactions to the automation of their local bank. The part of the interview that concerns us here consisted of just three questions: besides an item measuring overall satisfaction with the bank, the interview consisted simply of two open-ended questions, one asking people to cite all the things about the newly designed offices with which they were particularly satisfied, and the other the things with which they were dissatisfied. The order in which these two, open-ended questions were asked was of course rotated.

TABLE 4
Automated Banking: Bi-Polar Coding Scheme

SATISFACTION		DISSATISFACTION
Deposit money without waiting No wait; Fewer lines Fewer people	**Rapidity**	Cash deposit machine too slow
Machines are now inside rather than outside Machines are more reassuring	**Security**	Machines lack security measures Not reassuring to deposit/withdraw from a machine You're not isolated when making cash deposits
Instructions are clearly posted	**Clarity**	Lack of information Don't know where to go for special things
There is always someone to greet and receive you Personnel are more friendly, warmer, welcoming	**Reception**	Dehumanizing, there are more machines than people Impersonal, less opportunity to discuss with banker
More lively, full of light Clean The colors are bright, modern	**Interior Decor**	Cold / Impersonal
Easier access to reception desk Plenty of seats More space	**Spatial Layout**	Space too small, less room for clients The doors open all the time, create drafts The agency is too open to the outside
Always someone to inform you Information provided at the reception is sufficient	**Advice, Verification**	Check and cash deposits are not verified Receptionist more a hostess than an advisor
Machines for depositing cash and checks The design of the machines is impressive	**Equipment**	Machines for deposit often out of order Machines limited in what they can do You have to make neat packets before depositing
It's more practical Hours more flexible	**Practicality**	You have to do everthing yourself Need a debit card to withdraw money Difficult for the elderly
Advisors' offices are separate and enclosed Greater confidentiality	**Discretion**	Too visible, too open to the outside Lack of confidentiality/you hear everything

The responses to the open-ended questions were coded using the bi-polar schema presented in Table 4 in which the different reasons for satisfaction and dissatisfaction were coded into a common set of 10 categories or themes. Each theme was then formalized by a pair of binary variables which recorded for each respondent whether the theme was cited as a source of satisfaction (e.g. the décor is lively, full of light) *and/or* dissatisfaction (e.g. the décor is cold, impersonal) or not at all. The themes' salience, and hence their likely presence among the things brought to mind as respondents fine-tuned their initial anchored response, was measured by simply tabulating the frequency with which each was cited. The theme's importance, on the other hand, was measured by regressing the overall satisfaction on the pairs of binary variables. Salience and importance measures were thus distinct, and each theme's impact on overall satisfaction was free to adopt either a linear monotonic or asymmetric form according to the relative magnitude of the weights calculated for the corresponding pair of binary variables.

The results are presented in Table 5. The salience scores show that reactions were largely favorable, with Décor, Layout, Equipment and Rapidity being the sources of satisfaction most frequently brought to mind. Among the negative aspects, the lack of Discretion was the most salient, and there was a considerable degree of ambivalence with respect to the quality of Reception. The impact scores follow the same general pattern as the salience data but, as can most easily be seen with Reception, the negative citations tend to weigh more heavily than the positive.

TABLE 5
Automated Banking: Combined Sample

	Salience (Citation Rate)		Impact on Surface Satisfaction [a]		"Contribution"
	Satisfaction	Dissatisfaction	Satisfaction	Dissatisfaction	(Salience X Impact)
Rapidity	0,271	0,000	0,122		3,31
Security	0,093	0,054		-0,095	-0,51
Clarity	0,003	0,000			0,00
Reception	0,364	0,330	0,089	-0,176	-2,57
Interior Decor	0,376	0,000	0,115		4,32
Spatial Layout	0,302	0,059	0,137	-0,051	3,84
Advice, Verification	0,063	0,027	0,056		0,35
Machines	0,324	0,033	0,135	-0,103	4,03
Practicity	0,223	0,096	0,077		1,72
Discretion, Confidentiality	0,091	0,434		-0,163	-7,07

[a] **Values are standardized regression coefficients, significant at p<0.10**

Table 5 presents the separate results for private and business users. Though the two groups did not differ in their overall levels of satisfaction (3.16 and 3.13, respectively, on a 4 point scale), the cognitive basis of their satisfaction scores differs, and this not so much in the kinds of things that were salient as in the impact that these things had. Within each group, the most salient things are not necessarily the most impacting; the relatively infrequent concerns voiced by professionals with respect to the security and the practicality of the automated agency weigh heavily on the overall satisfaction level, while among private users, machine malfunction, though apparently rare, has a strong negative impact. Note also, the predominance of asymmetric effects—in this case positive—for the professionals, something we would expect from a group whose needs and expectations with regards to the bank were highly uniform.

TABLE 6
Automated Banking: Analysis by Group

	Salience (Citation Rate)		Impact on Surface Satisfaction[a]		"Contribution"
	Satisfaction	Dissatisfaction	Satisfaction	Dissatisfaction	Salience X Impact
A. Professional					
Rapidity	0,327	0,007	0,418		13,67
Security	0,044	0,102	0,427	-0,346	-1,65
Clarity	0,006	0,007			0,00
Reception	0,184	0,334	0,370		6,81
Interior Decor	0,496	0,000			0,00
Spatial Layout	0,455	0,058	0,627		28,53
Advice, Verification	0,037	0,090			0,00
Equipment	0,228	0,052	0,290		6,61
Practicality	0,341	0,112		-0,650	-7,28
Discretion	0,097	0,543	0,358		3,47
B. Private					
Rapidity	0,271	0,000	0,237		6,42
Security	0,093	0,054		-0,347	-1,87
Clarity	0,003	0,000			0,00
Reception	0,364	0,330	0,154	-0,333	-5,38
Interior Decor	0,376	0,000	0,247		9,29
Spatial Layout	0,302	0,059	0,212	-0,209	5,17
Advice, Verification	0,063	0,027	0,211	-0,390	0,28
Equipment	0,324	0,033	0,233	-0,567	5,68
Practicality	0,223	0,096	0,142		3,17
Discretion	0,091	0,434		-0,317	-13,76

[a] Values are standardized regression coefficients, significant at p<0.10

CONCLUSION

The approach that we present here is intended to fill certain gaps and correct certain biases that characterize much current satisfaction research. We intend it, however, as a complement rather than as an alternative to the usual methods. For even if, as we argue, the spontaneous judgments of surface satisfaction made by "naïve" subjects at the start of an interview are fundamentally different from the "deeper" more reasoned judgments made by the same person 20 minutes (and many questions) later, both are essential for understanding and managing customer satisfaction.

As we have said above, the regression-based methods based on additive structural models are effective for piloting satisfaction over the medium to long term, but they are too onerous to be deployed with great frequency and lack the reactivity necessary to be of use to managers in detecting problems as they arise on the field. Yet more than the ease with which it can be implemented, it's the fact that our proposed approach focuses on *what people actually say* rather than *what they could eventually respond if asked* that constitutes its main point of interest, since by so doing it can offer managers a much clearer and immediate picture of customers' problems as they arise. An ideal program for managing customer satisfaction would thus combine the two approaches, with relatively long periods between studies of deep satisfaction interspersed with regular (or even continual) surface satisfaction measures.

On a more theoretical level, the process model based on the anchorage and adjustment heuristic opens interesting new angles for thinking about customer satisfaction and tying it to other domains. Attribution theory (e.g., Kelley, 1973), for example, suggests that satisfaction ratings should become more informative as a function of their divergence from respondents' anchorage points. Estimating these anchorage points—for example, by having respondents judge a series of services other than the one under study, and entering them in a typology along with the final, adjusted ratings—might thus provide an interesting way for dissecting the highly skewed rating distributions characteristic of satisfaction studies, winnowing out low-involvement, "reflexive" responses and centering one's attention on those customers whose satisfaction is most contingent upon the successes and failings of the service in question.

Also, it is a small step to assume that the things that customers think of in adjusting their opinion are the same that they mention when talking with their entourage. The link between satisfaction and "word of mouth" or "buzz" is thus direct, and the problem of managing customer satisfaction becomes one of not only keeping one's customers but also one of keeping tabs on what one's customers may be saying to others, customers and prospects alike.

REFERENCES

Amstutz, A. (1970). *The computer simulation of competitive market response*. Cambridge Mass., The M.I.T. Press.

Bachelet, D. & Lion, J. Une Méthode d'Evaluation de l'Importance des Attributs Perçus Appliquée au Développement et au Positionnement des Nouveaux Produits, Revue Française du Marketing, no. 119, **4,** 1-11.

Bem, D.J. (1967) Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, **74,** 183-200**.**

Herzberg F., Mausner B. and Snyderman B. (1959). *The motivation to work*. New York, John Wiley and Sons.

Kahneman, D. & Tversky, A. (1973) On the psychology of prediction. *Psychological Review*, **80**, 237-251.

Kano N., Seraku N., Takahashi F. & Shinishi T. (1984). Attractive quality and must-be quality. *The Best on Quality* , 7, 165-186, John Hromi, ASCQ Quality Press, Milwaukee, Wisconsin.

Kelley, H.H. (1973). The processes of causal attribution. *American Psychologist* , **28** ,107-128.

Llosa, S. (1997) "L'analyse de la contribution des éléments du service à la satisfaction: Un modèle tetra-classe". Décision Marketing, **10** , 81-88.

Mittal V., Ross W.T. & Baldasare P.M. (1998) "The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions". Journal of Marketing, **62**, 33-47.

Strack, F. & Martin, L. (1987). Thinking, judging and communicating : A process account of context effects in attitude surveys. In H.J. Hippler, N. Schwartz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123-148). New York : Springer-Verlag.

Tourangeau, R. & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, **103**, 299-314.

DUAL RESPONSE "NONE" APPROACHES: THEORY AND PRACTICE

CHRIS DIENER
KING BROWN PARTNERS, INC.
BRYAN ORME
SAWTOOTH SOFTWARE, INC.
DAN YARDLEY
KING BROWN PARTNERS, INC

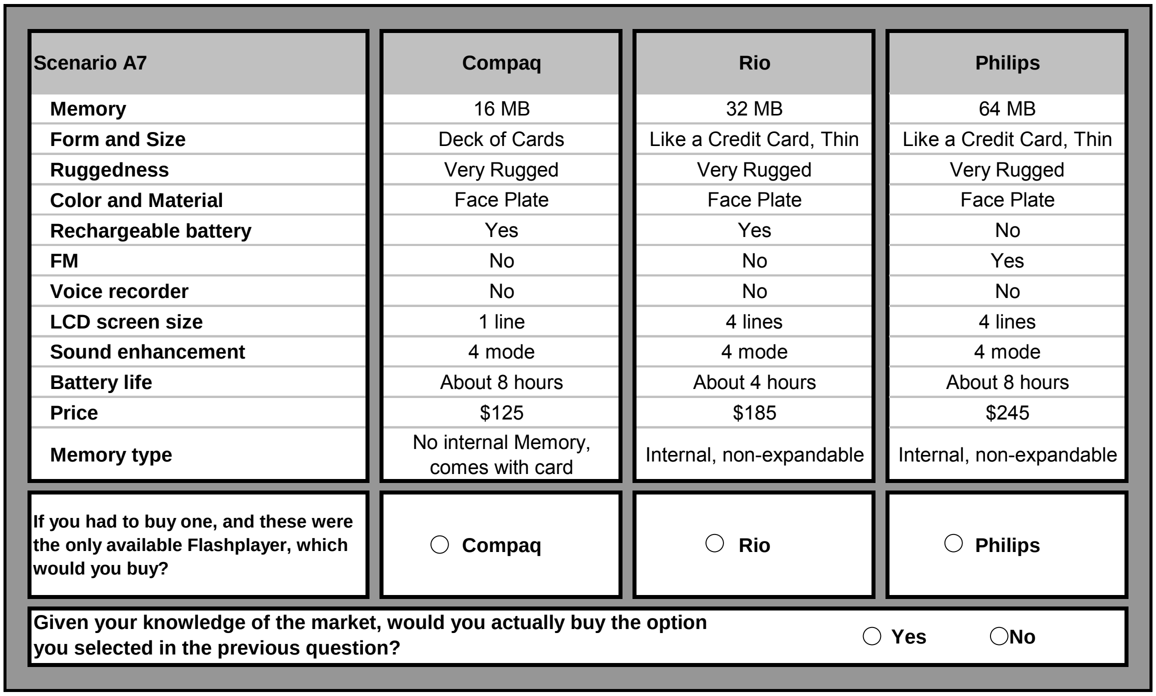
INTRODUCTION

Over the past several years a variant of choice-based conjoint referred to as *Dual Response "None"* has been the subject of increasing interest and use. With Dual Response "None," two responses are elicited for each task: a choice among available product alternatives, and then a choice between the "None" concept and the previous alternative(s). Even though this approach has been growing in use, its documentation has been sparse to date (Uldry, *et al.* 2002; Brazell, *et al.* 2006, Sawtooth Software 2005). This paper will introduce and explain Dual Response "None" and provide guidance on how to put it into practice.

DESCRIBING DUAL RESPONSE "NONE" (DR)

As the name implies, Dual Response "None" or DR, differs from standard CBC in that with each choice task respondents are asked two choice questions ("dual response") instead of only one choice question ("single response"). The first choice question (a forced choice task) elicits a choice among the described alternatives, not including the "None/Other" option. The second choice question has respondents choose whether, given what they know about the current marketplace, they would really buy the product they just selected. Essentially, it is a choice between the most preferred alternative and "None/Other." If respondents would actually buy the product, then they would not prefer "None/Other" but if they would not buy the product, then they would prefer "None/Other" over the given product.

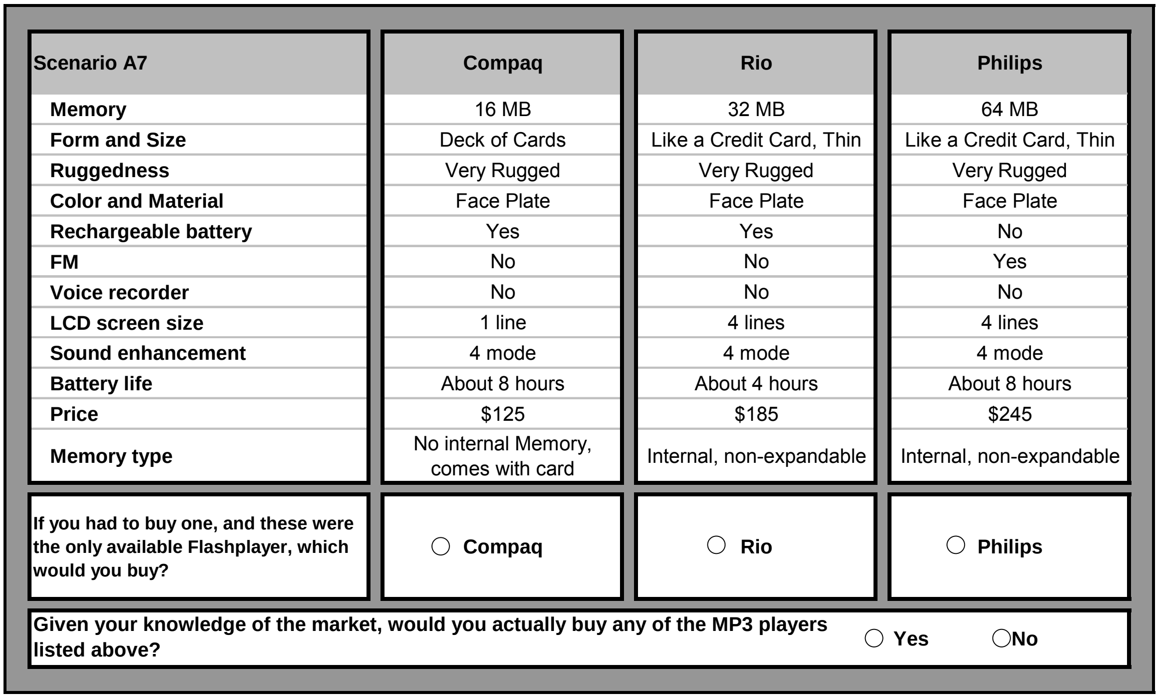
There are actually several variations of DR tasks. One is as just described, where a respondent is asked in the second choice to choose between the most preferred alternative and "None/Other." We will call this DR-2Max, because it is a choice between two alternatives: the preferred and "None/Other." Alternatively, the second task could be more global in nature, asking whether the respondent would buy any of the given alternatives. We will call this the DR-AnyMax task because it is a choice between none and any of the given alternatives. Example 1 shows a DR-2Max task and Example 2 shows a sample DR-AnyMax task.



Scenario A7	Compaq	Rio	Philips
Memory	16 MB	32 MB	64 MB
Form and Size	Deck of Cards	Like a Credit Card, Thin	Like a Credit Card, Thin
Ruggedness	Very Rugged	Very Rugged	Very Rugged
Color and Material	Face Plate	Face Plate	Face Plate
Rechargeable battery	Yes	Yes	No
FM	No	No	Yes
Voice recorder	No	No	No
LCD screen size	1 line	4 lines	4 lines
Sound enhancement	4 mode	4 mode	4 mode
Battery life	About 8 hours	About 4 hours	About 8 hours
Price	$125	$185	$245
Memory type	No internal Memory, comes with card	Internal, non-expandable	Internal, non-expandable
If you had to buy one, and these were the only available Flashplayer, which would you buy?	○ **Compaq**	○ **Rio**	○ **Philips**

Given your knowledge of the market, would you actually buy the option you selected in the previous question? ○ **Yes** ○**No**

(Example 1)



Scenario A7	Compaq	Rio	Philips
Memory	16 MB	32 MB	64 MB
Form and Size	Deck of Cards	Like a Credit Card, Thin	Like a Credit Card, Thin
Ruggedness	Very Rugged	Very Rugged	Very Rugged
Color and Material	Face Plate	Face Plate	Face Plate
Rechargeable battery	Yes	Yes	No
FM	No	No	Yes
Voice recorder	No	No	No
LCD screen size	1 line	4 lines	4 lines
Sound enhancement	4 mode	4 mode	4 mode
Battery life	About 8 hours	About 4 hours	About 8 hours
Price	$125	$185	$245
Memory type	No internal Memory, comes with card	Internal, non-expandable	Internal, non-expandable
If you had to buy one, and these were the only available Flashplayer, which would you buy?	○ **Compaq**	○ **Rio**	○ **Philips**

Given your knowledge of the market, would you actually buy any of the MP3 players listed above? ○ **Yes** ○**No**

(Example 2)

In these two examples, the "None/Other" option is not explicitly included in the grid describing the alternatives, as is the case with more traditional CBC tasks.

In practice, the DR-2Max task is preferred to the DR-AnyMax task. Though perhaps obvious, the DR-2Max task has the advantage of providing greater respondent focus and simplicity. The DR-AnyMax task may seem awkward and harder to understand.

DR may be implemented with another task variation where the question order is reversed as compared to DR-AnyMax. The first question becomes a traditional CBC question and the second one a forced choice. The respondent is first asked to choose among all of the alternatives, including "None/Other," and the second task is a forced task among only the given, described alternatives.

Other than these differences in the task, few other differences remain between DR and standard CBC. DR requires the same experimental design considerations and the same reporting requirements of CBC. The only differences, other than the task layout, between DR and CBC involve the estimation data setup (construction of the estimation design) and choice of estimation algorithm.

Now that we have described DR, we will review why you might want to use DR in the first place. Following this description, the remainder of the paper will detail the data setup and estimation requirements (focusing on the DR-2Alt and DR-AnyMax task formats) to guide the practitioner in successful DR application.

Benefits of Using Dual Response

On the face, it appears that DR would take both more respondent time (two questions instead of one) and analyst effort (because the data layout is more complex than standard CBC). If this is the case, why would a researcher use DR?

DR provides greater efficiency and power than CBC. In contrast to standard CBC, with DR, we still obtain information about non-"None" alternatives when a respondent selects "None/Other" as the preferred alternative. The model can therefore capture and use more information about the relative value of the attributes and levels, in addition to the information that the "None/Other" alternative is more appealing than any of the given alternatives.

As the incidence of "None/Other" choices rises, the benefit of DR becomes greater. With 12% "None/Other" responses, using DR creates a 25% decrease in model error This is equivalent to the difference between using 180 respondents in CBC and only needing 130 respondents using DR to get the same level of model error. Likewise, with 33% "None/Other" response, DR provides a 55% decrease in error requiring only 130 respondents where 290 would be required with standard CBC (Uldry, *et al.* 2002).

The good news is that DR has not been found to significantly alter the underlying model estimates—with the exception of the coefficient for the "None/Other" alternative. Accounting for scale differences, using real-world and simulated data sets, aggregate parameter estimates do not differ significantly between DR and standard CBC. (Uldry *et al.* 2002).

Finally, researchers will also find that using DR provides higher resolution on preferences at the individual level. Because of less "missing" information when respondents select "None/Other" as the preferred alternative, more information is available for individual-level modeling.

The benefits of using DR should lead many researchers to consider applying it in a broader range of projects. As mentioned earlier, DR differs from CBC, not only in the task, but in the data setup and estimation requirements. Because data setup depends on the estimation approach, we will first examine the estimation requirements.

ESTIMATION ALGORITHM DIFFERENCES BETWEEN DR AND CBC

DR captures two responses instead of just one. Therefore, a different estimation algorithm is needed. Information from two responses must be combined to estimate one coherent model.

Typical CBC estimation uses the standard Multinomial Logit (MNL) theory. This is expressed below in Equation 1.

$$P(i \mid FC \,\&\, FC_0) = \frac{e^{\beta'_{FC} x_i}}{\left(\sum_j e^{\beta'_{FC} x_j} + e^{\beta_{FC0}}\right)}$$

Equation 1 – Typical MNL.

The MNL represents the probability that alternative i will be chosen from the alternatives in the combined set of FC and FC_0. FC (representing "forced choice") are the given alternatives in the choice set and FC_0 is the "None/Other alternative).

In DR, and specifically DR-2Max, the probability of choice can be expressed as a joint probability between choosing the most preferred given alternative and the choice that the most preferred alternative is chosen over "None/Other." The probability of choosing alternative i from the set of given alternatives, FC, can be expressed as Equation 2:

$$P(i \mid FC) = \frac{e^{\beta'_{FC} x_i}}{\left(\sum_j e^{\beta'_{FC} x_j}\right)}$$

Equation 2 – Forced Choice.

And the probability of choosing i over FC_0 is then represented by Equation 3.

$$P(i \mid 2nd) = \frac{e^{\beta'_{FC} x_i}}{\left(e^{\beta'_{FC} x_i} + e^{\beta_{2nd0}}\right)}$$

Equation 3.

If we assume that these probabilities are independent, they can be multiplied to form a joint-probability. The joint expression, the DR-2Max likelihood function, is denoted as Equation 4:

$$P\left(i \mid DR\right) = \frac{e^{\beta'_{FC}x_i}}{\left(\sum_j e^{\beta'_{FC}x_j}\right)} \times \frac{e^{\beta'_{FC}x_i}}{\left(e^{\beta'_{FC}x_i} + e^{\beta_{2nd0}}\right)}$$

Equation 4 – DR-2Max likelihood function.

Alternatively, the DR-AnyMax model can be expressed as the joint likelihood between the forced choice and the choice of "None/Other" over any of the given alternatives. Again, the forced choice can be represented as Equation 2 above.

The second choice can be represented as Equation 6:

$$P\left(FC \mid 2nd\right) = \frac{\left(\sum_j e^{\beta'_{FC}x_j}\right)}{\left(\sum_j e^{\beta'_{FC}x_j} + e^{\beta_{2nd0}}\right)}$$

Equation 6.

These likelihoods can then be multiplied to become Equation 7:

$$P\left(i \mid DR_2\right) = \frac{e^{\beta'_{FC}x_i}}{\left(\sum_j e^{\beta'_{FC}x_j}\right)} \times \frac{\left(\sum_j e^{\beta'_{FC}x_j}\right)}{\left(\sum_j e^{\beta'_{FC}x_j} + e^{\beta_{2nd0}}\right)}$$

Equation 7.

Which then reduces to Equation 8, and is the DR-AnyMax likelihood function.

$$P\left(i \mid DR_2\right) = \frac{e^{\beta'_{FC}x_i}}{\left(\sum_j e^{\beta'_{FC}x_j} + e^{\beta_{2nd0}}\right)}$$

Equation 8 – DR-AnyMax likelihood.

This expression can then be converted to a log-likelihood expression and embedded within either a traditional maximum likelihood estimation routine or into a hierarchical Bayes approach. Working in this manner would require building a custom estimation routine.

However, standard MNL routines may be used. Because the joint likelihood is multiplicative in nature, the log-likelihood is additive. As an additive function, the DR model does not need to be estimated simultaneously with both responses included in a joint likelihood function during estimation. Instead, the responses can be included sequentially (as two choice tasks) and summed. In this way, a typical MNL estimation algorithm can be used to estimate a DR model. The dual responses can be fed one at a time into the estimation algorithm. To do this, the estimation algorithm must be able to accept tasks with different numbers of alternatives within the same respondent.

An alternative to building a custom routine and to using a standard MNL is to use the DR capability in Sawtooth Software's CBC/HB v4 (Sawtooth Software 2005).

Using a standard CBC estimation routine would require a different or unique data setup, which is the focus of the next section.

DATA SETUP FOR TYPICAL CBC ESTIMATION

A standard CBC estimation algorithm can be used to estimate a DR model. The only requirement is that the routine allow for differing numbers of alternatives within each respondent.

The dual responses must first be split into two separate choice tasks and these tasks must be "stacked" together for each respondent in the estimation design data. Essentially, each response is treated as a separate choice task. The first response is set up as its own forced choice task. The second response is either set up as a choice between the chosen alternative and "None/Other" (DR-2Max) or as a choice among all alternatives, including "None/Other" (DR-AnyMax). We will illustrate the data setup using a DR-AnyMax estimation.

ID	Scen	Alt	Choice	ASC1	ASC2	ASC3	A1L1	A1L2	A2L1	A2L2
30001	1	1	**0**	1	0	0	1	0	1	0
30001	1	2	**1**	0	1	1	0	0	0	1
30001	1	3	**0**	0	0	0	0	0	0	0
30001	2	1	**0**	1	0	0	0	1	1	0
30001	2	2	**0**	0	1	1	0	0	0	1
30001	2	3	**1**	0	0	0	1	0	0	0
…	…	…	…	…	…	…	…	…	…	…
30001	1	1	**0**	1	0	0	1	0	1	0
30001	1	2	**1**	0	1	1	0	0	0	1
30001	1	3	**0**	0	0	0	0	0	0	0
30001	1	4	**0**	0	0	0	0	0	0	0
30001	2	1	**0**	1	0	0	0	1	1	0
30001	2	2	**0**	0	1	1	0	0	0	1
30001	2	3	**0**	0	0	0	1	0	0	0
30001	2	4	**1**	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…

These first three rows are redundant and should be removed

Table 1.

Table 1 above illustrates the data setup. The call-out box and circled task show an additional requirement in the data setup: redundant tasks should be eliminated from the estimation design.

In other words, if someone chooses one of the given alternatives, then the information in the second task (where the "None" alternative is included) fully captures the information. Sawtooth Software's CBC/HB v4 software employs this stacked data strategy, automatically stacking the data for the researcher and eliminating any redundant tasks prior to the estimation process.

While in theory, the simultaneous likelihood should provide the same results as the stacked-data / standard CBC estimation, we wanted to confirm this and explore the potential impact of differences in data setup. To this end, several empirical studies and one simulated dataset were employed to compare the difference between simultaneous and sequential estimation and the differences among several methods of stacking the data for sequential estimation.

TESTING OF DIFFERENCES AMONG DATA SETUP OPTIONS AND ESTIMATION APPROACHES

We tested several data setup options in the context of different potential estimation approaches to illustrate consistencies and provide direction for those who wish to employ DR. We conducted these tests on several data sets from actual projects and one simulated data set. The combinations of data setup and estimation approach are as follows:

- "Sawtooth Software"
    - Sawtooth Software's CBC/HB v4
    - Need to reformat .CHO file to include information regarding the Dual None response. Specifications for this are provided in the manual.
- "DR-Custom"
    - Custom function based upon DR-AnyMax likelihood function with simultaneous LL usage of first and second choice
    - Data setup tailored to the specific algorithm needs
- "MNL-AnyMax"
    - Typical MNL
    - Using stacked data, where the second choice is from all alternatives, including None/Other
    - This data setup, when used with a typical MNL estimation, replicates the DR-AnyMax likelihood function.
- "MNL-2Max"
    - Typical MNL with stacked data, where second choice is between the previously chosen alternative in the forced choice and the none
    - This data setup, when used with a typical MNL estimation, replicates the DR-2Max likelihood function.

As a practical note, while the DR-2Max task (from a respondent's perspective) is generally preferred over the DR-AnyMax task, the MNL-AnyMax estimation routine is preferred. The data from a DR-2Max task may be used in the MNL-AnyMax estimation. We will discuss this preference at greater length after describing the results of the tests. All of the comparisons in this paper have been completed using data generated using DR-2Max tasks.

The four conditions above will be compared using hold-out predictions (ability to predict individuals' choices) and MAE (Mean Absolute Error of predictions versus actual choices for the population).

The four conditions were tested using one simulated data set and data from two actual studies. The results are found below in Table 2.

		Sawtooth Software	DR-Custom	MNL-AnyMax	MNL-2Max
Simulated	Hit Rate(%)	63	63	62	62
	MAE	1.41	1.41	1.53	3.29
DataSet1	Hit Rate(%)	63	60	59	57
	MAE	3.21	3.59	3.00	6.09
DataSet2	Hit Rate(%)	65	65	62	65
	MAE	2.83	2.75	2.66	3.40

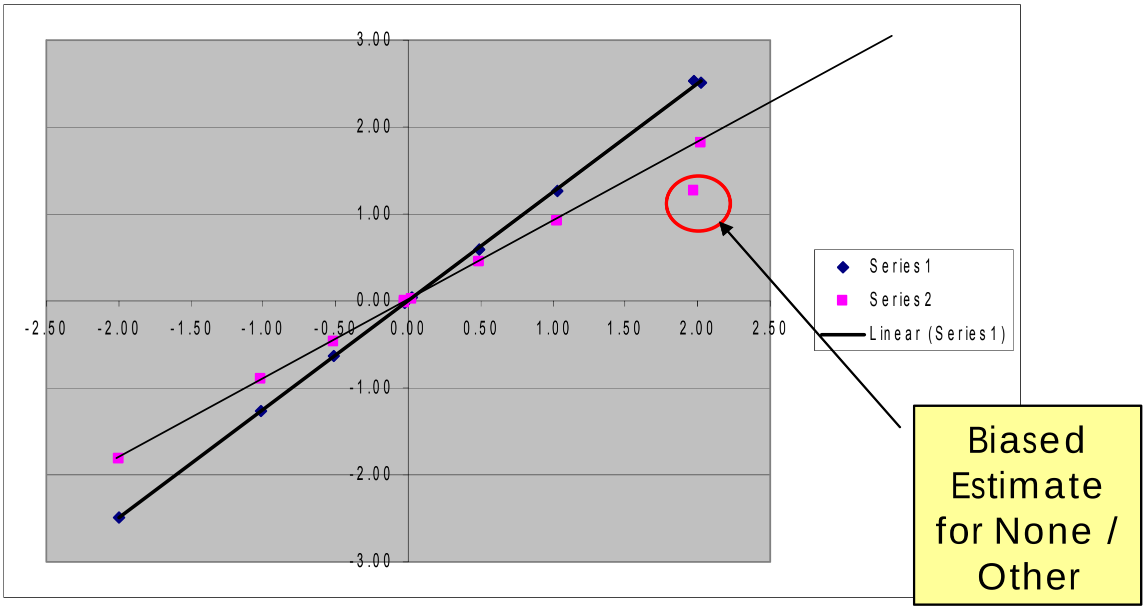
Table 2.

Our results show that DR-2Max performed more poorly in terms of MAE. However, in all other respects the different approaches performed similarly. The main finding for the purposes of this paper is that the researcher has a variety of estimation options for DR that all perform similarly well. Researchers can use their own custom functions, can stack the data and use the typical MNL or can use the DR feature in Sawtooth Software's CBC/HB v4.

Further Exploration and Explanation

To further explore the reason MNL-2Max performed more poorly, we plotted all of the model estimates together from each data set. We would expect that in a scatterplot, if each of the estimation approaches produced exactly the same estimates, the estimates would align creating a diagonal line at a 45% angle to the horizontal axis. If the different estimation techniques produced models that differed in scale, then the points would create a line that crossed through the origin but differed as to the angle against the horizontal axis. If any points deviated from the line, they would represent potentially significant bias in the model estimates between models.

We found, across datasets, that the MNL-2Max approach consistently produced a downward bias in the estimation of the effect for "None/Other." The plot for the simulated data is shown in Chart 1:



Biased Estimate for None / Other

Chart 1.

As the other estimation & data setup conditions utilized the DR-AnyMax likelihood function, the bias can be identified as a difference between using DR-AnyMax and DR-2Max. DR-2Max will consistently produce a lower estimate of "None-Other" than DR-AnyMax. But which is right?

We turned to the simulated dataset (and others that we created) to understand what the absolute bias may be. We found that the direction of this bias depends on the process of generating the data. If we create our simulated dataset using DR-AnyMax assumptions, it appears that DR-2Max underestimates the effect for None/Other. On the other hand if we use DR-2Max assumptions in generating simulated data, we find that DR-AnyMax estimation produces inflated estimates of None/Other.

The reason that DR-AnyMax estimates of None/Other are higher than DR-2Max comes from the second stage of the estimation—the one difference between the two approaches. With DR-AnyMax, for None to be consistently chosen at the same proportion as in the DR-2Max approach, it must have a higher utility because it is "competing" against more than just one other alternative. Each of the alternatives has its own random variation, and so, on average, the estimate for "None/Other" must be greater if it is competing against more alternatives and their independent variation.

The question then turns to whether the 2Max or AnyMax process more closely resembles the decision structure in the real world. In other words, for the given product category, do people ultimately make a buy/no-buy choice when looking only at the preferred alternative? Or, do consumers consider a set of alternatives? We believe it may be an empirical question. In either case, the only bias between the approaches appears to be in the "None/Other" estimate, which is already frequently the subject of post-model adjustment (or even ignored).

As an additional note, we performed many tests with regard to the specific coding of the alternative-specific constants (including that of "None/Other") in the data setup. We could not determine any differences between conditions with regard to model estimates and model accuracy. We tested the specific inclusion of a parameter for "None/Other." We tested the inclusion of a variable picking up the difference between the first and second response. We tested dummy codes versus effects coding. With these and several other tests we found no tangible differences in model estimates or accuracy.

CONSISTENCY BETWEEN TASK AND ESTIMATION APPROACH

Our experience suggests the appropriate use of the 2Max task with the AnyMax estimation, even though there appears to be a conceptual inconsistency. Practically speaking, the 2Max task is easier for the respondent and, we believe, should provide better information. From a rational viewpoint, the information respondents provide in the AnyMax task should be the same as that obtained in the 2Max task. However, this research effort did not specifically test the difference between using the two task formats. Additionally, irrespective of the way the data were collected, we perceive that the AnyMax estimation approach more accurately reflects the reality that None/Other is considered in the context of more than one alternative. Our experience and exploration here does suggest that there are no adverse effects of using the 2Max task to generate data for the AnyMax estimation approach or likelihood function.

SUMMARY AND CONCLUSION

The DR approach is now available to practitioners able to perform additional data processing for standard MNL software or within the Sawtooth Software .CHO file. Sawtooth Software plans to implement dual-response None as an automatic option within the next version of CBC/Web (version 6), in which case most any researcher will have access to DR. This paper sets forth the rationale and theory behind DR. Additionally, the research described in this paper shows that DR can be estimated using available software and standard algorithms and that differences between algorithms and data setup are largely inconsequential, meaning the approach is robust to variations in specific items of data setup or to whether you are using a simultaneous or sequential (data-stacking with typical MNL) approach.

To briefly review, past research has shown that DR offers several significant benefits. DR provides more efficient use of data, resulting in smaller design requirements, better prediction at a given sample size and greater statistical significance of parameter estimates. DR also provides potentially more realistic prediction of "None/Other," thus requiring less calibration. It has been shown that when using Dual Response, respondents choose "None/Other" much more often, reducing the need to scale "None/Other" when calibrating to actual market shares.

However, there are trade-offs to using DR and enjoying these benefits. These include a greater time requirement per scenario because there are two tasks instead of just one and the potential for annoying respondents by requiring a forced choice (this potential has not been empirically tested).

That it is a new approach and not well documented in the published literature is probably the largest obstacle and risk in using DR. Being a relatively new approach, Dual Response may raise more questions and may shift the burden of proof on your proposal in an area that should be an advantage.

In short, we have found many empirical benefits and a few drawbacks of the approach.

You, as a researcher, should consider using DR when:

- You expect to have a None/Other option in your choice sets and when you are doing a task that lends itself to a choice instead of some other behavior, like allocation
- You can spare a little more time for the tasks
    - However, if by using DR you can take advantage of the greater power, you can reduce the number of tasks and offset the extra time required to answer two tasks
- Sample size is an issue or you just want more robust estimates
- You are likely to have high number of "None/Other"
- You want to have less calibration of "None/Other."

References

Brazell, Jeff D.; Chris Diener, Ekaterina Karniouchina, William L. Moore, Valerie Severin, Pierre Francois Uldry, "The No Choice Option and Dual Response Choice Designs (March 2006), Accepted for publication with Marketing Letters.

Uldry, Pierre; Valerie Severin and Chris Diener. "Using A Dual Response Framework in Choice Modelling," 2002 AMA Advanced Research Techniques Forum, Vail, Colorado.

Sawtooth Software (2005), "CBC/HB v4 Manual," Sequim, Washington, USA.

"MUST HAVE" ASPECTS VS. TRADEOFF ASPECTS IN MODELS OF CUSTOMER DECISIONS

JOHN R. HAUSER
MIT
ELY DAHAN
UCLA
MICHAEL YEE
MIT LINCOLN LABORATORIES
JAMES ORLIN
MIT

This paper provides an applied managerial summary of the theory and empirical tests developed by the authors in a forthcoming *Marketing Science* article. For detailed derivations, proofs of all propositions, screen shots of the web-based data collection, details on the empirical analysis, detailed statistical tests, and other material please refer to:

Michael Yee, John Hauser, James Orlin, and Ely Dahan (2006), "Greedoid-Based Non-compensatory Two-Stage Consideration-then-Choice Inference," (April) forthcoming, Marketing Science.

The present paper is presented as a summary of the paper to be published in *Marketing Science*. Although we have endeavored to write new text, provide new figures, and organize the data in new tables specifically for this paper, should any conflict in copyright arise, it is to be resolved in favor of the Institute for Operations Research and Management Science (INFORMS), the publishers of *Marketing Science*.

"MUST-HAVE" FEATURES

There are over 300 make-model combinations of automobiles on the market, but the average consumer considers less than 10 make-model combinations. If an automobile manufacturer can determine how to entice consumers to consider its make-model combinations, e.g., a Ford Mustang, then the manufacturer can reduce the choice set from 1 in 300 to 1 in 10—a factor of 30. By designing cars that will be considered, an automobile manufacturer greatly increases its chances of making a sale. For example, General Motors (GM) believes that GM vehicles are better than consumers perceive them to be and that GM would gain share if consumers would be more willing to consider GM vehicles.

The consideration-set challenge is not limited to automobiles. A recent survey of websites selling personal digital assistants (PDAs) suggests that there are 21 models available at Circuit City, 25 at Staples, 27 at Microcenter, 30 at CompUSA, and 97 at Govconnection (MIT's approved vendor). It is the rare consumer who will evaluate all of the PDAs available before making a decision. More likely, the consumer will screen on some characteristics before evaluating a small subset of the available PDAs.

In each of these cases the managerial challenge is to identify the "must have" (or "must not have") features that determine the consideration sets of consumers. Must-have features are non-compensatory in the sense that a product with must-have features is preferred to a product without these must-have features even if the product without the must-have features is better on all other features.

The identification of must-have features is related to the conjoint-analysis goal of identifying the most important features, but there are differences. We seek to complement conjoint analysis methods. In particular, we address two issues: (1) We seek to infer directly non-compensatory decision processes, e.g., processes in which some products have must-have features that drive consideration or choice. (2) We explore methods that apply well if the respondent is asked to indicate those profiles which he or she would consider. This task can be used alone or in addition to the rank, rating, or choice tasks that are typical in conjoint analysis. The methods we summarize are practical. Empirical data suggests they often predict at least as well as traditional conjoint analysis.

Conjoint Analysis Assumes, Primarily, a Compensatory Model

Traditional conjoint analysis represents products by profiles of features and asks respondents to express their preferences among those profiles. Respondents might rank the profiles, provide ratings that express preference, or simply choose from sets of profiles. The basic preference model assumes that the preference for a profile can be expressed as a combination of the "partworths" of the feature levels. In most cases, the preference function is separable, meaning that a preference score is an additive or multiplicative combination of the partworths for the levels of the features that describe the profile. Although interaction terms are possible, they are not commonly used. For ease of exposition we ignore interactions, although this does not affect our basic arguments.

Figure 1 presents an additive model that illustrates how a respondent might evaluate two smartphones using a compensatory decision process. In Figure 1, the partworth for "Verizon" is larger than the partworth for "Cingular," but the other partworths of the Cingular phone, "flip," "price = $299," and "manufactured by Sony," compensate for the fact that, all else equal, the respondent would prefer "Verizon" to "Cingular." An additive conjoint model predicts that the respondent would choose the Cingular smartphone.

Figure 1
Illustration of a Compensatory Model

"Utility = sum of "partworths"

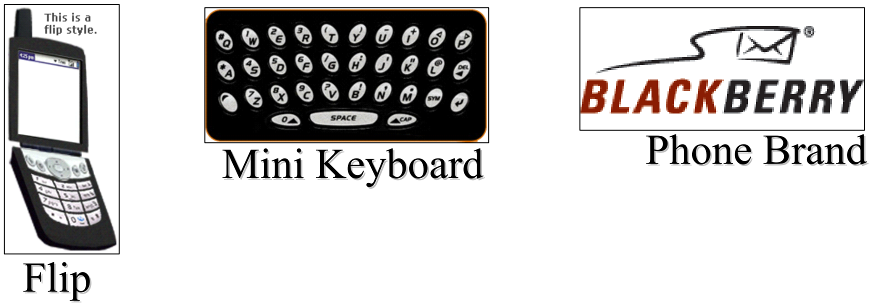
Verizon = 10	Cingular = 3
Brick = 2	Flip = 7
$499 = 1	$299 = 6
…	…
Nokia = 3	Sony = 5

"Flip" and "$299" can compensate for a lower partworth on "carrier."

But what if the respondent is faced with over twenty smartphones and adopts a different, more-realistic screening rule. As indicated in Figure 2, the respondent might consider only flip phones, with mini-keyboards, from Blackberry. If this were the case, smartphones without these characteristics would never be considered and would never be chosen. In a non-compensatory decision process, other features such as carrier, operating system, GPS capability, camera capability, and price cannot compensate for a smartphone that does not flip, have a mini-keyboard, or is from Blackberry.

Figure 2
Illustration of a Non-Compensatory Decision Process

"I will only consider flip phones, with mini-keyboards, from Blackberry"



In theory, an additive partworth model can represent such a process, but, in practice, such an additive model would put a strain on estimation. We illustrate this capability with binary features for ease of exposition. For binary features it is easy to identify a set of partworths that acts lexicographically. Lexicography is a non-compensatory process in which the respondent evaluates profiles by features, one at a time, accepting that profile only if it is better or equal to other profiles on the feature currently being evaluated. The respondent continues until all profiles have been sorted according to the task—full rank if traditional full-profile conjoint analysis, best of a set for choice-based conjoint analysis, or "considered" for a consideration task.

One function that acts lexicographically for *n* binary features is to assign 2^{n-1} to the first evaluated feature, 2^{n-2} to the next feature, …, and 1 to the last evaluated feature. For sixteen binary features, this means that the partworth for the first feature would be 32,768, the second feature 16,384, and the last feature 1. Clearly, partworths that vary in magnitude by a factor of over thirty-two thousand would put a strain on most estimation procedures and would be extremely sensitive to response errors. Furthermore, mimicking a lexicographic process by assigning values to partworths does not imply unique partworths, even to a positive linear transformation. Other combinations such as 3^{n-1}, 3^{n-2}, …, 1 would also work. In fact, still another set of "partworths," the denominations of US currency, acts lexicographically. A rational consumer would prefer 5¢ to 1¢, 10¢ to 5¢ plus 1¢, 25¢ to 10¢ plus 5¢ plus 1¢, etc. for 50¢, $1, $2, $5, $10, $20, $50, $100, all the way up to $10,000 – a form of currency no longer in circulation.

Consideration Task

The standard task in conjoint analysis asks the respondent to indicate preferences. Although most estimation procedures could be modified to deal with data in which the respondent simply expresses consideration, that has not been common. While we expect that non-compensatory processes are common for choice tasks, we expect they are even more common for consideration tasks. Thus, we want a method that handles consideration-set data naturally, as well as choice, ranking, or rating data.

Non-compensatory Models

In this paper we consider lexicographic models, an important class of non-compensatory models. As defined above for binary features, a respondent acts lexicographically if he or she first ranks the <u>features</u> of profiles. Those profiles with the first-ranked feature are preferred to those profiles without the first ranked feature. In the case of ties on the first-ranked feature, we move to the next profile continuing until the tie is broken. A profile with a higher-ranked feature is chosen even if another profile without the high-ranked feature is better on each and every lower-ranked feature.

For example, consider four binary features of smartphones: flip preferred to brick, small preferred to large, mini-keyboard preferred to no keyboard, and a Palm operating system preferred to a Microsoft operating system. Suppose the <u>features</u> are ranked flip, small, mini-keyboard, and Palm. Then if we represent profiles by indicting the features they have, this lexicographic ordering implies the following orders. We have underlined the critical comparison.

- {flip, large, no keyboard, Microsoft} $\succ$ {brick, small, keyboard, Palm}
- {flip, small, keyboard, Palm} $\succ$ {flip, large, keyboard, Palm}
- {brick, large, keyboard, Microsoft} $\succ$ {brick, large, no keyboard, Palm}
- {brick, large, keyboard, Palm} $\succ$ {brick, large, keyboard, Microsoft}.

The basic idea extends to multi-level features, but, in order to model consumer decision making, we must consider how consumers treat levels within features. Here we adopt Tversky's (1972) nomenclature and call each level of a feature an "aspect." Basically, an aspect is a binary feature, e.g., "flip vs. brick" or "Verizon vs. not Verizon." A multi-level feature is then a set of linked aspects. For example, the feature, "carrier," can have one of four aspects: Verizon, Cingular, Sprint, or Nextel.

For multi-level features we can have four different decision processes that vary on how consumers evaluate aspects (levels) within features. These are:

- lexicographic by features (LBF): the consumer first ranks the features and then ranks aspects within features
- acceptance by aspects (ABA): the consumer first ranks aspects and then accepts profiles if they have the aspects
- elimination by aspects (EBA): the consumer first ranks aspects and then rejects profiles if they do not have that aspect
- lexicographic by aspects (LBA): the consumer first ranks aspects and then either accepts a profile if it has an aspect or rejects a profile if it does not have an aspect

These four decision rules are illustrated in Figure 3 for playing cards. Note that (1) if all features are binary, then all four decision models can provide the same ordering of profiles, (2) if some features are multi-level, then the decision models provide different ordering of profiles, and (3) LBA nests all of the other decision models.[1]

[1] When we say that two decision models can provide the same ordering of profiles, we mean, there exists an aspect- or feature-ordering that orders the profiles the same. The aspect- or feature-ordering can be different in the two models being compared. For example, for a given ABA model there exists an EBA model that ranks profiles in the same order. To transform an ABA model into an EBA model, we reverse all binary aspects (e.g., change Verizon to not Verizon) and reverse the ranking of aspects. We can always represent an LBF model by an ABA model if we constrain aspects within a feature to be contiguous in a rank order, but we can identify ABA models that are not represented by equivalent LBF models. We can represent any LBF, ABA, or EBA model by an equivalent LBA model.

Figure 3
Four Lexicographic Heuristics Illustrated with Playing Cards

Totally Diverge

Simplifying Heuristic	TLA (*Three-Letter Acronym*)	Ranking Rule	First Choice	by 2nd Choice	3rd Choice	by 4th Choice	5th Choice	6th Choice	7th Choice	Last Choice
Lexicographic By Features	LBF	(♠ > ♥ > ♦ > ♣), (A > J)	A♠	J♠	A♥	J♥	A♦	J♦	A♣	J♣
Acceptance By Aspects	ABA	♠, A, ♥, ♦	A♠	J♠	A♥	A♦	A♣	J♥	J♦	J♣
Elimination By Aspects	EBA	~~♣~~, ~~J~~, ~~♦~~, ~~♥~~	A♠	A♥	A♦	J♠	J♥	J♦	A♣	J♣
Lexicographic By Aspects	LBA	~~♣~~, ♠, A, ♥	A♠	J♠	A♥	A♦	J♥	J♦	A♣	J♣

WHY NOT JUST ENUMERATE ALL LEXICOGRAPHIC MODELS?

A lexicographic model is a simple model and it is an easy decision rule for a consumer to use. Furthermore, while the set of all compensatory models is defined on an uncountable infinite set of partworth values ($\Re^{n-1}$ for n aspects), there are only finitely many lexicographic orderings of aspects. For example, if we know which aspects are favorable, e.g., the respondent prefers flip smartphones to brick smartphones, then there are $n!$ possible lexicographic orderings of the n aspects.[2] If we need to infer which aspects are favorable, then there are $2^n n!$ possible lexicographic orders.

If the number of aspects is small, then it is feasible to enumerate exhaustively all possible lexicographic orders and choose, for each respondent, the lexicographic order of aspects that best describes that respondent's profile ordering (rating or choice). For example, with 3 aspects we need only consider $3! = 6$ aspect orders for ABA or EBA and only $2^3 3! = 48$ aspect orders for LBA. With 4 aspects we need only consider $4! = 24$ and $2^4 4! = 384$ aspect orders for ABA and LBA, respectively.

With exhaustive enumeration we could use any reasonable metric to define "best." For example, we might maximize a Spearman rank correlation, a Kendall's τ rank correlation, or, perhaps, the number of paired violations.[3]

The challenge comes when the number of aspects grows. For example, Martignon and Hoffrage (2002) study a famous lexicographic problem and exhaustively enumerate all lexicographic orders for a 9-aspect problem. Their problem requires a total of $9! = 362,800$ orders to be evaluated. They report computations that required two days to complete. On

[2] If the model is constrained to LBF there are fewer possible orderings ($m_1!m_2!$ for m_1 features at $m_2!$ levels each).

[3] Minimizing the number of paired violations is equivalent to maximizing Kendall's τ.

today's computers we can do such computations much faster, perhaps minutes or even seconds. With 300 respondents and 9 aspects, exhaustive enumeration might still be feasible for 9 aspects. However, when we increase the challenge to a 16-aspect problem, then we must evaluate 16! aspect orders for ABA. However, 16!= 57,657,600 * 9! if we knew the preferred direction of each aspect. It is much worse if we must infer the preferred direction of each aspect: $2^{16}16! =$ 3,778,648,473,600 * 9!. If the Martignon and Hoffrage algorithm took just a single second to run, an LBA problem would take almost 120,000 years. Clearly, exhaustive enumeration of reasonably-sized problems will not be feasible any time soon, even taking Moore's Law of increased computing power into account.

GREEDOID METHODS DEVELOPED BY YEE, ET. AL. (2006)

Fortunately, the relationship between an aspect order and an induced profile order has special structure. Yee, et. al. (2006) establish that this relationship can be described by a greedoid language if the goodness of fit measure is to minimize the number of paired comparisons that are violated. Based on the mathematics developed for greedoid languages, Yee, et. al. prove the following propositions.

1. if there exists a lexicographic ordering on the aspects that induces a profile ordering that matches the respondent's profile ordering, then the aspect ordering can be found with a polynomial-time algorithm (a greedy algorithm).
2. if there is no lexicographic ordering on the aspects that induces a profile ordering that perfectly matches the respondent's profile ordering, then the best-fitting aspect ordering (or orderings) can be found with a dynamic program that runs in the order of 2^n steps.
3. the dynamic programming algorithm can be altered slightly to allow the paired orderings to be weighted differentially. A weighted algorithm also runs in the order of 2^n steps.

Yee, et. al. and, later, Yee (2006) prove, in addition, a series of propositions that speed the algorithm further. The net result is that there is a simple algorithm that can solve reasonably-sized problems in a second or less. (The simple algorithm requires only a dozen lines of pseudo-code. More complicated algorithms require more code, but run significantly faster.) Yee, et. al. further demonstrate that if the manager is only interested in the aspects that are ranked highly, then even large problems, say 50 aspects, can be solved in reasonable time.

With the greedoid-based dynamic programming algorithms it is now feasible to use realistic preference, choice, or consideration tasks to infer the best lexicographic ordering of the aspects.

BENCHMARKS

We consider two traditional benchmarks. The first is hierarchical Bayes and the second LINMAP. In our empirical test, we collect data with either a rank-order data or a consider-then-rank task, thus the appropriate hierarchical Bayes model is a ranked-logit model, which we label HBRL. LINMAP is a traditional model based on linear programming (Srinivasan and Shocker 1973). However, recently, Srinivasan (1998) recognized that the performance of LINMAP could be improved by requiring that partworths be chosen to assure strict rank orderings of the profiles. In previous versions, if the respondent ranked profile *i* above profile *j*, then LINMAP's goodness-of-fit measure would not penalize a set of partworths that rated profile *i* equal to profile *j*. The new version of LINMAP adds a penalty for partworths that allow ties. (See Srinivasan 1998 for details.) We provide comparisons between the two versions of LINMAP.

There is an important complication that we must take into account when using either HBRL or LINMAP as benchmarks. HBRL and LINMAP are additive models and, hence, have the capability to identify partworths that act as if the respondent were using a lexicographic model. (Review Section 2.) Although finding such partworths may not be practical when there is response error, such partworths are theoretically feasible. Any benchmark comparisons must take this (theoretical) nesting of models into account.

If the goal is simply to find the model that predicts best, then this complication does not matter. We place no constraints on either HBRL or LINMAP and choose the model that predicts best. However, we might have other goals. We might want to gather evidence to attempt to infer the decision rule that the respondent is using. In this case, we might take as evidence a finding that the best lexicographic model predicts better than the best compensatory model. In order to gather such evidence, we must define what we mean by compensatory.

In general, compensatory means that doing well on some aspects can compensate for doing poorly on other aspects. The least restrictive version of a non-compensatory decision rule is an additive partworth model in which the largest (aspect-based) partworth exceeds the sum of all other (aspect-based) partworths, plus, the second-largest partworth exceeds the sum of all remaining partworths, and so on to the smallest partworth (Kohli and Jedidi 2006). Unfortunately, imposing constraints based on this least restrictive definition is difficult computationally because such a set of constraints does not imply a convex set.

A simpler and more-realistic set of constraints is to define a *q*-compensatory model such that no partworth ratio is larger than *q*. (The largest partworth can be no larger than *q* times the smallest partworth.) This is an extension of the definition used by Bröder (2000) in psychology. (Bröder's comparison used $q = 1$.) Note that this is a definition. By the principle of optimality, imposing the *q*-constraints will lead to worse fit, although it might lead to better predictions. Thus, we cannot, in principle, write an algorithm to find the "best" *q*, unless we use holdout data to identify *q*, which would not be appropriate.

We compare models on three metrics. Our fit measure is the percent of ranked-pairs that are not violated. This measure is optimized by the greedoid-based dynamic program, although the program is searching over a finite set of orders compared to an uncountable infinite set of partworths that is evaluated by HBRL or LINMAP. Although the details vary, this fit measure is close to that which is optimized by LINMAP. This fit measure is not optimized by HBRL, although we expect that HBRL will do well on this fit measure.

In addition to our fit measure, we compare the performance of the greedoid-based dynamic program and the benchmarks on their ability to predict holdout data. The two measures of prediction are (1) holdout pairs that are not violated and (2) holdout hit rate.

EMPIRICAL DATA

Yee, et. al. tested greedoid methods with a 2x2 empirical experiment in which respondents evaluated 32 smartphones profiles that varied on 16 aspects which were grouped into seven features: carrier (Verizon, Cingular, Sprint, Nextel), manufacturer (Sony, Samsung, Nokia, Blackberry), price ($99, $199, $299, $499), operating system (Palm, Microsoft), form (flip, brick), keyboard (mini, none), and size (small, large). Half the respondents were asked to first indicate which smartphones they would consider and then to rank only those smartphones that they would consider. The other respondents ranked all 32 smartphones. In a full-crossed design, half of the respondents were allowed to presort smartphones by features and half were not. There was also a fifth cell in which respondents were presented with only 16 smartphones in the consider-then-rank, no-sort task.

Screen shots are given in Yee, et. al. Respondents were shown a fractional factorial of 32 profiles. Each profile (smartphones) was represented by icons which illustrated the features. The task and the questionnaire was pretested carefully and respondents understood the task, the icons, and the questions. To indicate consideration, respondents clicked on the smartphone icon. After the click, the considered smartphone was surrounded by a blue box. When respondents were finished indicating their consideration set, they clicked to continue. Smartphones that were not considered disappeared from the screen and only the considered smartphones were displayed. Respondents then clicked on their first choice, which disappeared from the screen. This continued until all considered smartphones had been ranked. Respondents in the rank-only task where not shown the consideration screen; they completed the rank task for all 32 smartphone profiles. In the two experimental cells in which respondents were allowed to presort the smartphone profiles, they were presented with three drop-down boxes in which they could choose features on which to sort. They could sort as many or as few times as they wanted.

In the holdout task, respondents were presented with two sets of four additional smartphone profiles, chosen from a second fractional factorial of the seven features. We designed the holdout task so that it had a different look and feel. In particular, respondents sorted the four profiles with a task similar to that which is used to sort slides in Microsoft PowerPoint. They then indicated which, if any, of the profiles they would consider.[4] Prior to completing the holdout task, and after the initial consider-the-rank or rank-only task, respondents completed a mini-IQ test to cleanse short-term memory (Frederick 2005).

Results (Summarized from Yee, et. al. 2006)

Table 1 summarizes the empirical comparisons reported in Yee, et. al. As expected, both LINMAP and the greedoid-based lexicographic-by-aspects (LBA) model do well on fitted pairs, although LINMAP is slightly (but significantly) better. HBRL, which does not optimize fitted pairs, does less well. Constraining the model to be q-compensatory reduces fit, more so for HBRL than for LINMAP. It appears that LINMAP can readjust its optimization to overcome the q-constraint.

[4] For an analysis of the holdout consideration data, see Yee (2006).

Table 1
Fit and Predictive Ability (from Yee, et. al. 2006)

	Fit (Pairs)	Holdout Pairs	Holdout Hit Rate
Lexico by aspects	95.5%	**74.5%***	**59.7%***
Lexico by features	82.8%	65.8%	48.1%
HBRL	87.1%	**74.3%***	54.9%
HBRL (*q*)	82.8%	69.0%	48.9%
LINMAP	**96.9%***	**73.6%***	54.9%
LINMAP (*q*)	95.7%	**73.6%***	**56.9%***

*Best or not significantly different than best at 0.05 level

The more interesting comparison is on the holdout data. LBA does better, albeit not significantly so on holdout pairs, than either HBRL or LINMAP. We are quite impressed with this result, which we interpret as "at least as well," because LBA is a much more highly constrained model than either HBRL or LINMAP. It certainly suggests further investigation.

When we constrain the additive models so that they are truly *q*-compensatory, then holdout predictions are significantly worse for HBRL (*q*).[5] One interpretation is that respondents may, in fact, be using simpler non-compensatory decisions. However, this interpretation can at best be treated as initial evidence and is subject to further empirical tests.

The comparison to a LINMAP-based *q*-compensatory model is quite interesting. Because LINMAP optimizes fitted pairs, there is a danger of over-fitting the data. Various constraints might help mitigate over-fitting. The non-compensatory constraints are one such set of constraints and they seem to improve holdout predictions at the expense of fit.[6] On the other hand, the *q*-compensatory constraints might also mitigate over-fitting. The improvement is slight relative to an unconstrained LINMAP, and not as much as LBA, but it is interesting and worth future research.

Finally, recall that LBF is nested within the more-general LBA. LBF is another form of constraints and may prevent over-fitting if respondents are truly using features rather than the more-detailed aspects to rank profiles. However, on average, predictions are not improved with LBF (relative to LBA) suggesting that most respondents are likely using aspects rather than features to order profiles. This result is intuitive in this category. Respondents might have a favorite carrier, say Verizon, which becomes an acceptance aspect. Once the consideration set is limited to Verizon, the respondent may prefer to rank on other aspects, say price or form, than rank on the remaining carriers.

[5] In this table we use $q = 4$. Yee, et. al. provide results for q that varies from $q = 2$ to $q = \infty$. The qualitative interpretations are relatively insensitive to the choice of q within reasonable ranges.

[6] The greedoid-based dynamic program searches over a finite set of aspect orderings rather than an uncountable infinite set of partworths. For every ordering there is a set of partworths that acts isomorphically.

Yee, et. al. also use the data to address a series of behavioral questions. We summarize the results here:

- the consider-then-rank task is rated significantly better on enjoyment, interest, and perceived accuracy and, for the no-sort cells, takes significantly less time to complete,
- about 2/3rds of the respondents' holdout predictions are at least as good with LBA as with a compensatory model and predictions are, on average, about 5 points higher,
- about 2/3rds of the respondents appear to be using aspects rather than features to sort profiles,
- Yee, et. al. obtain the same pattern of significance with either holdout hit rates or holdout pairs,
- giving respondents the opportunity to presort profiles does not appear to provide a significant difference in respondents' tendency to use lexicography,
- more respondents appear to be lexicographic if they are asked to complete a rank task than a consider-then-rank task,[7]
- for the consider-then-rank task, there does not appear to be a significant difference in the use of lexicography among those respondents who asked to evaluate 32 profiles as compared to 16 profiles.[8]
- when the analysis is replicated on a data set from Lenk, et. al. (1996), Yee, et. al. obtain a similar pattern of results as with the smartphone data. For example, LBA does just as well as HBRL on holdout hit rate. It does well, but not quite as well as HBRL on holdout pairs.[9]

Table 2
Additional Results on LINMAP and the Use of Self-Explicated Data

	Fit (Pairs)	Holdout Pairs	Holdout Hit Rate
Strict LINMAP w/o SEs	**96.9%***	**73.6%***	**54.9%***
Classic LINMAP w/o SEs	79.3%	67.2%	45.5%
Strict LINMAP w/ SEs	88.3%	**72.3%***	**53.1%***
HBRL w/o SEs	87.1%	**74.3%***	**54.9%***
HBRL w/ SEs	79.4%	69.3%	49.9%

*Best or not significantly different than best at 0.05 level.

[7] The number of pairs on which the model fit is optimized is less with the consider-then-rank task than with the full-rank task.. Although this applies equally to HBRL, LINMAP, and the greedoid-based dynamic program, we cannot rule out an interaction in the sense that one of the methods might be more sensitive than the others to the amount of information available.

[8] At first glance this null result appears to contradict the standard results in behavioral science such as in Payne, Bettman, and Johnson (1993). However, the majority of the standard results deal with a choice task rather than a consider-then-rank task. The Yee, et. al. data compare the percent of non-compensatory decision making for the consider-then-rank task. Most of the pairs in the consider-then-rank task result from consideration decisions by the respondent. The second rank-phase plays only a small role in estimation.

[9] LINMAP does not do as well as HBRL on the Lenk, et. al. data. The original Lenk, et. al. data were ratings data. HBRL, LINMAP, and the greedoid-based dynamic program are based on degrading the ratings data to retain only rank-order information. Interestingly, the holdout hit rate obtained with LBA and HBRL on the degraded data is not significantly different than the holdout hit rate that Lenk, et. al. obtained using HB on the metric data.

ADDITIONAL RESULTS

Yee, et. al. do not report a comparison of classic LINMAP to the new strict-pairs LINMAP. Nor do they report the details of a comparison in which the estimation methods use self-explicated data as well as the rank (or consider-then-rank) data. These results are presented in Table 2.

As predicted by Srinivasan (1998), the new strict-pairs LINMAP does significantly better than classic LINMAP on fit, holdout pairs, and holdout hit rate.

When self-explicated data (SEs) are added to LINMAP as constraints, the fit measure is degraded as is expected by the principle of optimality. However, LINMAP with SEs does as well as LINMAP without SEs on both holdout pairs and holdout hit rate. As recommended by Sawtooth (Sawtooth Software 2001), we expect HBRL with SEs as constraints to improve predictive ability. This does not seem to be the case with our data. See also discussion in Hauser and Toubia (2005) and Liu, Otter, and Allenby (2005).

SUMMARY

Greedoid-based methods provide a new, practical means to infer non-compensatory decision processes from the data that we normally collect for conjoint analysis. The data collected by Yee, et. al. were based on a full-profile rank task or a full-profile consider-then-rank task. The data collected by Lenk, et. al. (1996) were based on a full-profile ratings task. The greedoid-based dynamic program is also applicable to choice tasks and/or partial-rank tasks. With suitable modification, it should be applicable to partial-profile task.

The greedoid methods and related developments in the inference of non-compensatory decision making are new, but promising. Initial empirical tests suggest that the new methods fit or predict holdout data as well as less-constrained methods (see also Kohli and Jedidi 2006). The methods are easy to use, fast, and feasible for practical numbers of aspects. They can identify the features that respondents “must-have,” that is, features that are ranked highly in a lexicographic aspect order. We are optimistic and hope that other researchers continue to explore non-compensatory inference from conjoint-like data.

Greedoid-based methods are certainly not the only means to analyze non-compensatory decisions. Kohli and Jedidi (2006) used a modified greedy heuristic and report excellent results. Gilbride and Allenby (2004) use an HB specification to infer a screening process and also report excellent results.

REFERENCES

Bröder, Arndt (2000), "Assessing the Empirical Validity of the "Take the Best" Heuristic as a Model of Human Probabilistic Inference," Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 5, 1332-1346.

Frederick, Shane (2005), "Cognitive Reflection and Decision Making," Journal of Economic Perspectives. 19, 4, (Fall), 24-42

Gilbride, Timothy and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," Marketing Science, 23, 3, (Summer), 391-406.

Hauser, John R. and Olivier Toubia (2005), "The Impact of Utility Balance and Endogeneity in Conjoint Analysis," Marketing Science, 24, 3, (Summer), 498-507.

Kohli, Rajeev and Kamel Jedidi (2006), "Representation and Inference of Lexicographic Preference Models and Their Variants," Working Paper, Columbia University, New York, NY, April.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," Marketing Science, 15, 2, 173-191.

Liu, Qing, Thomas Otter and Greg M. Allenby (2006) "Investigating Endogeneity Bias in Conjoint Models." Working Paper, Ohio State University, Columbus, OH, (January).

Martignon, Laura and Ulrich Hoffrage (2002), "Fast, Frugal, and Fit: Simple Heuristics for Paired Comparisons," Theory and Decision, 52, 29-71.

Payne, John W., James R. Bettman, and Eric J. Johnson (1993), The Adaptive Decision Maker, (Cambridge, UK: Cambridge University Press).

Sawtooth Software (2001), "The ACA/Hierarchical Bayes Technical Paper," (Sequim, WA: Sawtooth Software, Inc.)

Srinivasan, V. (1998), "A Strict Paired Comparison Linear Programming Approach to Nonmetric Conjoint Analysis," in Operations Research: Methods, Models, and Applications, Jay E. Aronson and Stanley Zionts, eds., (Westport, CT: Quorum Books), 6-111.

______ and Allan Shocker (1973), "Linear programming Techniques for Multidimensional Analysis of Preferences, Psychometrika, 38, 3, (September), 337-369.

Tversky, Amos (1972), "Elimination by aspects: A theory of choice," Psychological Review, 79, 281-299.

Yee, Michael (2006), "Inferring Non-Compensatory Choice Heuristics," Ph.D. Thesis, Operations Research Center, MIT, Cambridge, MA. (June)

______, Ely Dahan, John Hauser, and James Orlin (2006), "Greedoid-Based Non-compensatory Two-Stage Consideration-then-Choice Inference," (April) forthcoming, Marketing Science.

EXTERNAL EFFECT ADJUSTMENTS IN CONJOINT ANALYSIS

BRYAN ORME AND RICH JOHNSON
SAWTOOTH SOFTWARE, INC.

BACKGROUND

The market simulator is generally the most practical and powerful deliverable of a conjoint analysis study. It lets the analyst or manager conduct an infinite number of "what-if" scenarios within the context of specific competitive products. The results are expressed in terms of shares of preference summing to 100%. These shares are probabilities, bounded by 0 and 1.0, and interpretable on a ratio scale.

Although many academics refer to simulated shares of preference as "market shares," practitioners usually avoid that label. It is not that those academics somehow believe that conjoint simulator shares are more accurate in predicting real world behavior than practitioners; they just don't have to make client presentations as often. Practitioners face the challenge of communicating the benefits of conjoint simulators without raising unrealistic expectations regarding the tool's ability to forecast sales volume and actual market shares. After all, there are often substantial differences between simulated shares of preference and actual market shares. Conjoint market simulators make a variety of assumptions, including:

- Equal distribution
- Equal awareness
- Perfect information
- Equal time on the market to reach maturity
- Equal effectiveness of sales force and marketing efforts
- That all attributes that influence product choice have been included in the model

Indeed, there are myriad elements that affect actual market shares beyond the fundamental fitness (utility) of the product concept vis-a-vis others on a level playing field. Even when conjoint simulators fail to predict actual market shares, they can be excellent tools for revealing strategic moves for improving market share. But, simulated shares of preference should be taken as *relative* indications of preference rather than indicators of *absolute* sales volume or market share. In the ideal world, managers would recognize the assumptions and accept conjoint simulators for what they do well—ignoring the fact that the shares of preference do not match the actual market shares precisely.

SHOULD WE ADJUST SHARES AT ALL?

We do not live in an ideal world. Many researchers have been able to take the "high road" when presenting simulators and simulation results, while others have faced immense pressure to calibrate the simulated shares to known benchmarks. (Using the word "calibrate" seems to convey a more scientific procedure than simply changing the shares to the desired result through "fudge factors.") Many researchers are paid a premium to apply conjoint data within a larger marketing forecast model, wherein they must account for additional external effects. Conjoint data certainly should be useful in this role.

The purpose of this paper is *not* to encourage or justify the widespread practice of adjusting conjoint simulators in an attempt to account for external effects. Researchers would do well to deliver the simulator as-is, and educate managers regarding the assumptions, proper interpretation, and use of the tool. However, there are cases where researchers must make adjustments to shares to accomplish a specific business purpose. Some types of corrections are perfectly legitimate (such as for distribution and scale factor) and should be done, given the proper data. Other types of corrections are theoretically less justifiable, though perhaps appropriate under certain conditions.

There are few if any published papers on adjustments to conjoint simulators to account for external effects. For many academics, the thought of adjusting conjoint data to match desired base case targets may not only seem heretical, but perhaps also too practical a problem to be of much interest. For practitioners, modifying model parameters is a potentially dangerous topic, but one that needs to be better understood.

REASONS WHY CONJOINT SIMULATORS FAIL TO PREDICT MARKET SHARES

Earlier, we listed some key assumptions of conjoint market simulators. To the degree that these assumptions do not hold in the real world for a specific product category and market, the shares of preference will deviate from actual market shares. There are many other reasons for disconnect between simulated shares and market shares:

Study Design Errors:

- Problems formulating proper attributes and levels
- Choice of conjoint method not compatible with real-world behavior
- Experimental design inadequate to support precise estimates of effects
- Respondents not properly instructed and/or confused by tasks

Data Collection Errors:

- Insufficient sample size
- Improper sampling of people (including modality and non-response bias)
- Time lag between survey data and market share measurement
- Inaccurate measurement of actual market share

Violations of Simulator Assumptions:

- Unequal distribution
- Unequal awareness
- Unequal marketing/sales effectiveness
- Unequal time in market to reach maturity[1]

[1] We should note that one can develop factors based on diffusion theory to adjust share for new products. There is quite a lot already available in the literature on this topic. We will not treat it further here, but recognize that such models can also be used in developing external effect factors.

- Simulator product specifications inconsistent with respondents' perceptions of performance[2]

Respondent Reliability:

- Respondents provide inconsistent answers ("Noise")
- Learning/context effects compromise parameter estimates

Respondent Validity:

- Respondents choose not to answer realistically
- Respondents are unable to answer realistically

Modeling Errors:

- The additive, compensatory assumption in conjoint analysis doesn't accurately capture buyers' actual decision processes
- Unresolved IIA problems/unrecognized heterogeneity
- Failure to account for significant within-concept interaction effects
- Failure to account for significant across-concept interaction effects
- Failure to use the proper market simulation model when converting part worths to shares of preference

SOME APPROACHES TO ADJUSTING SIMULATED SHARES:

Researchers have used various techniques to change the simulated shares of preference to match target base case market shares (in an attempt to account for unexplained variation due to external effects). Some approaches we're aware of are:

Adjustments for awareness

Adjustments for distribution

Scale factor adjustments

- A positive multiplicative factor applied to all utilities (also known as the "Exponent" in Sawtooth Software tools)

Aggregate shares adjustment (Sawtooth Software's Simple Approach)

- Positive multiplicative factors applied to aggregate shares, after which shares are re-normalized to sum to 100%

[2] One of the authors (Johnson) has had some experience with this problem. His experience is that when perceptual data are used within simulations, the perceptual data often have greater influence on simulated shares than the part worth utilities. This is not desirable.

Individual-level adjustments to shares/utilities

- Applied as positive multipliers to shares at the individual level, or as individual-level utility adjustments to product concepts

Respondent weighting

- A less widely-used procedure that we describe more fully in this article

ADJUSTMENTS FOR AWARENESS

Simulated shares of preference do not match actual market shares due in part to awareness. In the standard conjoint interview, respondents are permitted to choose among all alternatives, even those of which they had previously not been aware. This creates artificial awareness that may not be appropriate to model real world markets.

If we can assume that respondents will not buy brands/products that they are unaware of, then this offers a straightforward mechanism for adjustment. We can simply ask respondents which brands they are aware of and post-process the file of part worth utilities, setting any part worth for brands the respondent is not aware of to an arbitrarily low value (say -15) so that the simulated shares for those brands are zero for this respondent. (This essentially reflects a micro-level correction for product distribution/availability.)

Of course, one could extend this approach to the creation of the conjoint questionnaire itself, customizing the levels to include only the brands of which the respondent is aware. This leads to other opportunities and complexities beyond the scope of this paper.

Lack of awareness may not be such a barrier to purchase in many product categories, where the buyer learns about many options in the natural course of making a purchase. Therefore, using lack of awareness to set a share probability to zero may be too extreme a strategy in many situations.

ADJUSTING FOR UNEQUAL DISTRIBUTION

Unequal distribution is a common reason for simulated shares to deviate from actual market shares. Conjoint interviews typically make all brands and product options available to all respondents. One correction strategy for locally-purchased items is to ask respondents which zip code or other type of region they live in, and then set the part worth utility of any non-available product alternative to some arbitrary low value (such that the alternative will not receive any allocated share in simulations) for each respondent.

When a product has incomplete distribution, one might consider simply reducing its simulated share by some appropriate proportion. But that would assume that all other products gain equally when that product is not available. For example: Assume there are four products (or brands) in a market: A, B, C and D. Further assume that product A receives an overall simulated share of preference of 20% when in competition with B, C, and D for the entire sample. However, A is only available in half the markets. Assuming two equally sized regions, where A is only available in region 1, the proper share is:

Region 1: A=20%
B+C+D=80%
Region 2: A=0%
B+C+D=100%
Overall: A = 20/(100+100) = 10%

But what about the relative shares for B, C, and D? It might be tempting to deal with this situation just by cutting the share for A in half and redistributing ½ A’s share proportionally across B, C and D. However, A did not necessarily compete evenly with B, C and D. Perhaps A was perceived as a close substitute to B, and B would benefit more strongly (than C or D) by A not being available.

Two strategies for this simulation situation are to:

1. Run a simulation with all four products together. Record the shares. Run a separate simulation with only B, C, and D. Record the shares. Average the sets of shares from both runs (where A receives zero share when not available).
2. Replicate the sample. In the second replicate, set the utility of A to a very low value such that no share is allocated to A for any respondent[3]. Run the simulation with competitive set {A, B, C, D} across the total sample (both replicates).

Few modeling situations are as easy as the one described above with just two (or a very few) regions. It is typical to have data like these:

Product A available in 90 stores*
Product B available in 120 stores
Product C available in 70 stores
Product D available in 160 stores

*Note: we could as easily say regions or zip codes rather than stores.

It is tempting just to use the standard approach available in the Sawtooth Software market simulator to adjust the aggregate shares by multiplicative weights proportional to distribution, such as 0.9, 1.2, 0.70, and 1.6 (and re-normalize to 100%). However, these adjustments reflect proportional adjustments to aggregate shares. These products perhaps do not compete equally with one another. And, the patterns of availability of products in the stores (or regions or zip codes) may certainly matter.

If one has more complete data regarding the availability of products within specific regions, this permits the analyst to make adjustments for distribution that also account for unequal

[3] When using Sawtooth Software programs, it is probably most convenient to append a new attribute within the file of respondent part worths (the .HBU file) that includes a new level corresponding to each product to be modified. Use a data processing program script to read in the body of the original .HBU file, modify the data, and write out a new .HBU file. Level one of the new attribute contains an additive utility correction for A, level 2 for B, etc. The additive correction should be very low relative to logit-scaled utilities to force the share to zero for that product for each respondent, such as -15. In the market simulation, specify that A includes level 1 of the appended attribute, B includes level 2, etc. This convention is also useful for more sophisticated adjustments to account for distribution as discussed below. Adjust the "header section" of the .HBU file to account for the additional attribute and levels, and import the .HBU through the Run Manager into a new SMRT project.

substitution effects arising from different combinations of products being available within each region. Assume the following combinations of products available in various stores:

A and D	40 stores
B and D	20 stores
A, B, and D	30 stores
B, C, and D	50 stores
A, B, C, and D	20 stores

Summary:
Product A available in 90 stores
Product B available in 120 stores
Product C available in 70 stores
Product D available in 160 stores

The proportion of stores representing each available combination of products are:

A and D	40/160 = 25.00%
B and D	20/160 = 12.50%
A, B and D	30/160 = 18.75%
B, C and D	50/160 = 31.25%
A, B, C and D	20/160 = 12.50%

Given these data (but no information about which specific respondents shop in which stores), we can simulate respondent shopping trips by assigning respondents to randomly sampled stores offering different combinations of available products. A simple strategy to build such a simulation model starts by randomizing the respondent records in the data file of part worths. Recall above that A and D are the only products available in 25% of the stores. Therefore, for the first 25% of cases, set the utility of brands B and C to arbitrary low values (using a data processing procedure such as was described earlier). Set the utility of A and C to arbitrary low values for the next 12.50% of cases (where only B and D are available), etc. This procedure will result in lower precision for simulated shares because we are discarding some information for many respondents regarding their relative preferences for brands "not available" to them. One way to recover nearly all the benefit of the total sample with regard to precision of estimates is to replicate the sample multiple times (as many times as can be reasonably managed using your software tools), randomizing the order of cases across the replicates, and eliminating brands according to availability as before for different blocks of the expanded sample, proportional to the number of stores offering each unique combination of brand availability. Then, run the simulation across all replicates.

We can generalize the above procedure as a sampling and simulation exercise based on multiple draws per respondent. Within each draw, we simulate the respondent's purchase by random assignment to an available store (or region), where each store reflects the array of true-to-life available brands. If the stores do not capture roughly equal volume, one could easily extend the approach to account for different sales volume per location, by weighting the probability of assignment to a store proportional to the volume accounted for by that location.

From a data processing standpoint, it may be easier to think about each store (or region) as a separate array of zeros and ones, with number of elements equal to total brands in the study (available brand = 1, not available = 0). Replicate the separate files of stores and respondents multiple times, so there are at least as many stores as replicated respondent records. Randomize (independently) the separate files of stores and respondents, and assign the first store to the first respondent, the second store to the second respondent, etc., until all respondent records have been assigned. Adjust the utilities of non-available brands in each case by assigning an arbitrarily low utility of (say, -15). Read the new set of modified part worths into the market simulator. For example, with 800 original respondents, it might be appropriate to create 80,000 replicated records, so that each respondent is simulated for 100 shopping trips among available stores.

Adjustments based on distribution are quite justifiable. We'd suggest doing it prior to any adjustments described below, given you have access to good distribution data.

SCALE FACTOR ADJUSTMENTS

Consider a set of simulated and target market shares as follows:

	Simulated	Target
A	27%	19%
B	33%	31%
C	40%	50%

At first glance, it would seem that we are predicting brand B very well, but incorrectly predicting for A and C. Upon further inspection, we have correctly predicted C > B > A. Perhaps the errors in prediction are mainly due to *scale factor*.

Imagine that in reality, we had interviewed 500 human respondents and 500 monkeys. The human respondents actually supplied data that perfectly predicted the target market shares. But, the monkeys had answered randomly, with predicted shares of 33.3%, 33.3%, 33.3%. The net result of adding the monkeys to the human respondents is to add random noise, which "flattens" the simulated shares. Experienced conjoint analysts recognize that they can tune the scale factor (exponent) by multiplying all the utilities by a positive constant, prior to applying a market simulation rule such as logit, or Randomized First Choice. (Note that scale factor adjustments have no effect upon the first-choice simulation rule.)

The real world reflects a lot of random behavior, induced by variety-seeking, out-of-stock conditions, imperfect information, and satisficing. It is quite possible that the amount of random error present in the conjoint laboratory does not match (and is lower than) the amount of random variation in real buyers' decisions. Adjusting the scale factor for conjoint data is an appropriate and theoretically justified method for tuning simulated results to more closely fit actual market shares. Scale factor adjustments uniformly tip the shares to be "steeper" or "flatter." They preserve the original rank order of preference. After making appropriate corrections for awareness and distribution, we'd suggest next tuning the scale factor to best fit target market shares. Tuning for scale factor should occur prior to making further adjustments to shares (such as described below) to account for unexplained error.

Procedure for Accounting for External Effects via Aggregate Share Adjustment

If after making adjustments for awareness, distribution, and scale factor the shares still do not match target shares, then the analyst requiring a base case simulation matching target shares will need to make further adjustments. These external effect adjustments are developed once for the base case scenario, and retained in all subsequent simulations.

However, they are less defensible, and represent the attempt to somehow account for unexplained differences often without the use of any meaningful explanatory variables. One common approach that makes no pretense of actually explaining the differences involves changing the simulated shares using aggregate share adjustments. This has been offered (with strong cautions in the manual, we might add) in Sawtooth Software's market simulator for many years.

Assume three products in a simulation, with simulated and target shares as follows:

	Simulated	Target
A	0.10	0.15
B	0.30	0.25
C	0.60	0.60

Compute the ratio of target shares to simulated shares:

	Simulated	Target	Ratio
A	0.10	0.15	1.500
B	0.30	0.25	0.833
C	0.60	0.60	1.000

We simply apply the values in the final column above as multiplicative external effect adjustments to the average simulated shares of preference. These simulated shares of preference usually have resulted from shares first computed at the individual level (and then averaged across respondents). After multiplying the average simulated shares by the external effect adjustments and re-normalizing the shares to sum to 100%, the simulated shares exactly match the target shares.

"Reversal Anomaly" with Aggregate Adjustment Method:

In the course of doing technical support for Sawtooth Software's products, we sometimes receive reports of the software seeming to provide strange results. On a few occasions, users have shown us examples where making an isolated change to a product that decreases its utility not only decreases its share, but decreases the share of another (unchanged) competitive product as well. This is counterintuitive, as making one product worse might be expected to result in all other products gaining at least some share.

We'll illustrate the case with an example, based on a real data set (the "TV" data set) that ships with Sawtooth Software products. Consider the following base case, and the shares that result when Sony2 raises its price by $50:

	Base	Sony2	
	Case	+$50	Change
JVC1	9.49	9.71	+0.22
JVC2	18.65	19.96	+1.31
RCA1	9.06	9.94	+0.88
RCA2	28.52	29.84	+1.32
Sony1	14.39	15.54	+1.15
Sony2	19.89	15.00	-4.89

Assume the following external effect factors, to be applied as adjustments to aggregate shares:

JVC1	1.0
JVC2	1.0
RCA1	1.0
RCA2	1.0
Sony1	1.0
Sony2	0.5

After applying external effect factors, the new results for this base case and after the price increase for Sony2 are:

	Base	Sony2	
	Case	+$50	Change
JVC1	**10.53**	**10.50**	**-0.03**
JVC2	20.71	21.58	+0.87
RCA1	10.07	10.74	+0.67
RCA2	31.67	32.26	+0.59
Sony1	15.98	16.80	+0.82
Sony2	11.04	8.11	-2.93

JVC1's share decreases when Sony2 raises its price! This result is quite counterintuitive and can be disconcerting to analysts and managers alike. This anomalous result stems from the fact that the cross-elasticity between JVC1 and Sony2 is quite low (they are not very substitutable). After Sony2 raises its price, the redistribution of its smaller absolute share from Sony2 to JVC1 (than when Sony2 was $50 cheaper) through the external effect adjustment cannot make up for the small gain in JVC1's share due to the increase in Sony2's price. This result is described in more detail for the curious reader in Appendix A.

Procedure for Accounting for External Effects via Individual-Level Utility Adjustment

Suppose that rather than adjusting the aggregate shares, we want to find an adjustment to each brand's part worths, identical for each respondent and unique for each brand, so that the simulated shares become equal to the target shares. A simple iterative procedure for doing this might involve these steps: Observe the differences between target shares and unadjusted simulated shares. Considering either the ratios or differences for those two sets of numbers, add those ratios or differences to each respondent's brand part worth and re-simulate. The discrepancies between simulated and target shares will generally be closer than before, and the same procedure can be applied iteratively to make the discrepancies as small as desired.

This approach focuses on respondents on-the-cusp of choice behavior when shifting shares from one product alternative to another. Respondents with larger scale factor (respondents with higher certainty) are less affected.

We illustrate this procedure with the following example:

	Simulated Shares	Target Shares
A	0.101	0.150
B	0.190	0.250
C	0.061	0.100
D	0.293	0.200
E	0.150	0.100
F	0.204	0.200

Take the ratio of target shares to simulated shares, and zero-center those ratios, by subtracting off the average of the ratios.

Step 1

	Simulated	Target	Ratio	Zero-centered Ratio	
A	0.101	0.150	1.480	0.355	(1.480 – 1.125, etc.)
B	0.190	0.250	1.314	0.188	
C	0.061	0.100	1.628	0.503	
D	0.293	0.200	0.683	-0.442	
E	0.150	0.100	0.667	-0.458	
F	0.204	0.200	0.980	-0.145	
		Average:	1.125		

The zero-centered ratios[4] are added to each product concept and we re-simulate shares of preference given the adjusted product utilities. The new simulated shares (see **Step 2** table below) are closer to the target shares, but we need to make another step in the desired direction. We update the ratio of target to simulated shares, and zero-center those new ratios:

[4] It really isn't necessary to zero-center the ratios, since a positive constant may be added to all product utilities without changing the simulated shares under the logit model. This convention merely assures that the product utilities do not become so large that it affects the ability of the computer to deal with extremely large numbers that could result after exponentiation if many iterations are performed. Zero-centering also assures uniqueness of the solution (factoring out the arbitrary constant).

Step 2

	Simulated	Target	Ratio	Zero-centered Ratio
A	0.129	0.150	1.164	0.159
B	0.238	0.250	1.051	0.046
C	0.095	0.100	1.056	0.051
D	0.222	0.200	0.901	-0.104
E	0.117	0.100	0.858	-0.147
F	0.200	0.200	1.001	-0.004
		Average:	1.005	

We add the new zero-centered ratio to each product concept's previous utility, and re-simulate using the updated utilities. Therefore, the new utility adjustment for A after Step 2 for each individual is 0.355 + 0.159 = 0.514, etc.

The target and simulated shares after six steps of this procedure are:

	Target	Step0	Step1	Step2	Step3	Step4	Step5	Step6
A	0.150	0.101	0.129	0.141	0.146	0.148	0.149	0.150
B	0.250	0.190	0.238	0.247	0.249	0.249	0.250	0.250
C	0.100	0.061	0.095	0.099	0.100	0.100	0.100	0.100
D	0.200	0.293	0.222	0.206	0.201	0.200	0.200	0.200
E	0.100	0.150	0.117	0.106	0.102	0.101	0.100	0.100
F	0.200	0.204	0.200	0.201	0.201	0.201	0.201	0.200

After just six steps, the target shares are obtained to three decimal places of precision. Please note that this example is based on HB utilities at the individual level, which have much greater scale (larger range in utilities) than for latent class, logit, or ACA utilities. When using utilities with smaller scale factor, you may need to reduce the "step size" by multiplying the zero-centered ratios by a value such as 0.50 or 0.10.

This simple procedure may not be the most efficient method for reaching target shares in the fewest steps, but it seems to work pretty quickly. And, assuming you are not using too large a step size, will achieve the desired results. If using a logit-based simulation method, it will always achieve the desired results to as much precision as you desire, and the utility-adjustment solution is unique. In other words, there is only one such set of globally applied zero-centered utility adjustments that will result in the target simulated shares[5].

Procedure for Accounting for External Effects via Respondent Weighting

It is second nature to market researchers to weight respondents based on demographics to match known population characteristics. A related idea might be applied in conjoint simulators.

Sometimes differences between simulated shares and actual market shares may be due to the particular sample of respondents. For example, the sample may contain too few users of a particular product, or it may contain buyers who shop at outlets that don't carry that product.

[5] Sawtooth Software users can implement adjustments to product utilities in the market simulator using the same method of appending a new attribute (reflecting the utility adjustment, one level for each product to be adjusted) to the file of part worth utilities as was discussed earlier when dealing with adjustments for distribution.

Problems with sample may be due to the sample frame itself, or due to non-response bias. In such cases, it may be possible to improve simulated results by weighting respondents.

Rather than directly changing simulated shares or modifying model parameters, we seek a set of respondent weights that will produce simulated shares identical to the target shares. Of course, with only a few products to fit, and a very large number of weights that can be adjusted, there would be an infinite number of different solutions to the problem. We choose among those solutions by attempting to find the solution in which the respondent weights are as close to unity as possible.

We start with respondent weights of unity, and calculate a small change for each respondent in the direction that will minimize the sum of squared differences between simulated and target shares. (In optimization terminology, this is the "gradient" vector.) We continue this process through many iterations. In the end, we have computed a set of weights that make simulated shares match target shares exactly, and have done so in many small steps, each of which produced the largest improvement in matching of aggregate shares with the smallest change in the respondent weights. It is not guaranteed that the respondent weights will all remain acceptable—for example, some weights may become negative if the initial discrepancies between individual and target shares are too large. But if these weights are set to zero, the remaining weights will retain their least squares properties. The actual steps in the iterative approach are as follows:

Consider three products with the following simulated and target shares of preference for a sample of respondents:

	Simulated	Target
A	0.25	0.20
B	0.45	0.40
C	0.30	0.40

The simulated shares are based initially on all respondents having a weight of unity.

Consider the first respondent, with individual-level shares of preference (such as from a logit rule simulation model, BTL, or Randomized First Choice) as follows:

A	0.80
B	0.15
C	0.05

1. Subtract simulated shares of preference for the sample from target shares, obtaining the vector [-0.05, -0.05, 0.10]. This characterizes the direction and magnitude in which we would like to adjust simulated shares to match target shares.
2. For each respondent, subtract the target shares from the simulated shares. For the example respondent above, this results in the vector [0.60, -0.25, -0.35]. This vector describes how this respondent's shares compare to the target shares.
3. For each respondent, multiply the vectors from steps 1 and 2. For this example respondent, this results in the vector [-0.03, 0.0125, -0.035].

4. For each respondent, sum the elements within the vector obtained in step 3. For this example respondent, the result is -0.03 + 0.0125 + -0.035 = -0.0525.
5. For each respondent, add the value obtained in step 4 to the previous weight for this respondent. If any resulting weights are below zero (or another desired lower limit, such as 0.10), set them to zero (or the desired limit, such as 0.10).
6. Normalize the weights obtained in step 5 to sum to the number of total respondents, by multiplying each weight by the number of respondents divided by the sum total of weights across respondents from step 5. These are the new weights for respondents.
7. Update the simulated market shares for the sample, based on the new respondent weights.
8. Repeat steps 1 through 7 until weighted simulation results match target shares to the desired degree of precision.

These steps can be executed relatively easily in a spreadsheet program[6], and a macro can be written which copies the new weights over the old weights at the completion of each iteration (paste special + values). It often takes 100 or more iterations to match target shares to a high degree of precision, so it's imperative to automate the procedure using a macro.

In some situations, you may wish to constrain some products to match target shares, while not constraining other products. Simply zero-out any elements in the vector described in step 3 corresponding to non-constrained products.

There are many sets of weights that can be obtained for which the simulated shares match target shares. However, this iterative procedure finds the one unique solution where the weighted simulation returns target shares while also minimizing the standard deviation of weights across the sample (unless weights are constrained, as in step 5, in which case the standard deviation of weights is not minimized). If the target shares deviate considerably from simulated shares or if there is considerable homogeneity in the sample, no solution using positive weights may be possible that lead to a weighted simulation that matches target shares.

Often, adjusting simulated shares to exactly match target shares may result in quite extreme respondent weights. One can find a compromise between the variance of the weights and the fit to target shares by stopping after fewer iterations[7].

A Series of Tests:

In addition to adjustments for distribution and awareness, we've described three methods that have been used for adjusting simulated shares of preference to match target market shares. The remainder of this paper is dedicated to investigating how these methods perform in a series of tests. Following the tests, we will summarize with conclusions and recommendations.

[6] "Solver-Like" Excel plugins may also be used to find respondent weights such that a cell containing the mean squared error between simulated and target shares is minimized. However, as of the time of writing, the standard Excel "Solver" plugin was limited to 200 variables (in this case, respondent weights). So, data sets with more than 200 respondents would require more capacity (more capable programs are readily available on the market). Excel's Solver program produces the same results as the iterative procedure described above.

[7] A twist on this approach that significantly reduces the ratio of maximum to minimum weights (and avoids negative weights) is to assign respondents into groups according to the product with the highest utility (first choice rule). With a six-product simulation, we find only six weights for six respondent groups such that the simulated shares (under the logit, BTL, or Randomized First Choice) match target shares. While the range of weights decreases relative to the method of finding a unique weight for each respondent, the variance of the weights is increased. The results for all three tests shown here are nearly identical to the method of customized respondent weights.

Test #1: Reduction of Share for One Product: Substitution Effects

Consider the situation in which we want to reduce one product's share dramatically using external effects. We'll consider a base case scenario with six products:

	Base Share
JVC1	9.49
JVC2	18.65
RCA1	9.06
RCA2	28.52
Sony1	14.39
Sony2	19.89

We have constructed these products such that RCA2 and JVC2 are identical in all ways, except for brand. Thus, they should be highly substitutable. What happens if RCA2 raises its price by $50?

Effect of Price Increase for RCA2

	Base Share	RCA2 +$50	% Gain or Loss	
JVC1	9.49	10.97	+16%	
JVC2	18.65	23.58	+26%	(close substitute for RCA2)
RCA1	9.06	10.68	+18%	
RCA2	**28.52**	**17.11**	**-40%**	
Sony1	14.39	16.44	+14%	
Sony2	19.89	21.22	+7%	(not close substitute for RCA2)

As expected, simulated share of preference for RCA2 reduces and all other products gain share. But, since we are using individual-level (HB) utilities, the gains are not proportional. JVC2 gains the most (on a percentage basis) and Sony2 gains the least. Another way of stating this is that JVC2 is the closest substitute to RCA2 and Sony2 is the least substitutable. Those people seen as most likely to choose RCA2 are also quite likely to choose JVC2 and very unlikely to choose Sony2.

What would happen if we reduced RCA2's simulated share from 28.52 to 17.11 not through making it less desirable by increasing its price, but through various methods of adjusting for external effects? What would we expect to observe? Should the other products absorb RCA2's lost share in a similar pattern to the substitution effects we observed above? We'll test three different methods for adjusting for external effects, and demonstrate that the results are *very* different. We'll label the three methods as follows:

Agg Adj = Aggregate Adjustment (Sawtooth Software method)
Indiv UtilAdj = Individual-Level Utility Adjustment
Resp Wts = Respondent Weights

Change in Product Shares
When RCA2 Cut from 28.52 to 17.11
via Price Increase or External Effects

	Via Price	Agg	Indiv	Resp
	Increase	Adj	UtilAdj	Wts
JVC1	+16%	+16%	+14%	+18%
JVC2	+26%	+16%	+26%	–2%
RCA1	+18%	+16%	+17%	+14%
RCA2	**–40%**	**–40%**	**–40%**	**–40%**
Sony1	+14%	+16%	+13%	+21%
Sony2	+7%	+16%	+9%	+29%

The three methods of reducing RCA2's shares lead to *very* different results:

- The Aggregate Adjustment redistributes RCA2's share in proportion to the competitors' shares (each competitor gains equally, in proportion to their previous shares).
- The Individual Utility Adjustment redistributes RCA2's share very much in step with the previous substitution patterns that we saw earlier due to increases in RCA2's price.
- The Respondent Weights approach redistributes RCA2's share exactly *opposite* the patterns suggested by substitution effects.

Which is more correct? It depends on your beliefs regarding the need for adjusting RCA2's share to account for external effects.

Individual-Level Utility Adjustment is more appropriate:

- If the adjustment was due to RCA2's lack of full distribution (or awareness), we should expect JVC2 to satisfy those buyers who would have preferred RCA2 but did not find it available (or were not aware of it). JVC2 picks up RCA2's losses at a quicker rate. (However, we have already argued that corrections for distribution and awareness should be accounted for effectively *prior* to adjusting via external effects using one of these techniques.)
- If RCA2's share was too high because of overstated brand utility (equity), sales force effectiveness, or life stage maturity. Its close substitute (JVC2) should pick up share most rapidly when RCA2 is weaker on these fronts.

Respondent Weighting is more appropriate:

- If RCA2's shares are overstated because its performance features are not as desirable as the model predicts. If RCA2's share is overstated, then products (like JVC2) that share similar features might also be overstated. Thus, when RCA2's shares are reduced, products with similar performance aspects like JVC2 should also see a reduction.
- If RCA2's shares are overstated because the sample reflects too many kinds of respondents that gravitate toward RCA2 (and like products). When respondents who tend to like RCA2 are weighted downward, the share for RCA2 (and close substitutes) should also be reduced. Shares for products that appeal to different kinds of respondents (those that are not very substitutable) should increase more rapidly.

Test #2: Reduction of Share for One Product: Own Elasticity

Let's consider again the original base case scenario, with starting shares of preference:

	Base Share
JVC1	9.49
JVC2	18.65
RCA1	9.06
RCA2	28.52
Sony1	14.39
Sony2	19.89

We can easily estimate the price elasticity of demand for the products (%Δ Share/%Δ Price) by increasing the price of each by 10%. The (self) elasticities are:

	Elasticity
JVC1	-2.50
JVC2	-3.82
RCA1	-3.22
RCA2	**-3.21**
Sony1	-3.80
Sony2	-2.09

(Note: even though we are focusing on price elasticity of demand in this example, the same argument holds for sensitivity to other product changes beyond price.)

Rather than reduce RCA2's simulated share through a price increase, let's assume it loses share because it drops a desirable feature. In this case, let's assume it gives up its Picture-in-Picture capability (but continues to charge the same price). Simulated share for RCA2 would drop from 28.52 to 10.32. After this change in product formulation, RCA2's new price elasticity of demand is -4.01, instead of the original -3.21. This is in line with expectations, as a product with lower share and less desirable features should experience an increase in price elasticity.

What would happen if we reduced RCA2's simulated share from 28.52 to 10.32 not through the loss of Picture-in-Picture, but through various methods of external effects? How would the price elasticity be affected? We'll test the three different methods for adjusting for external effects, and again demonstrate that the results are different.

New Price Elasticity for RCA2, When Share Cut from 28.52 to 10.32 via Loss of PIP or External Effect Adjustment

	Loss of PIP	Agg Adj	[8]Indiv UtilAdj	Resp Wts
RCA2	**–4.01%**	**–3.73**	**-3.89**	**-4.29**

All of the methods for adjusting for external effects result in increases in self-elasticity for RCA2 relative to the original -3.21. But, the Respondent Weights method increases the self-

[8] Adjusting the share from 28.52 to 10.32 is a dramatic adjustment. The solution that results in the smallest standard deviation of non-negative weights across the sample would have led to 1/5 of the respondents having a zero weight. We constrained the solution such that the smallest respondent weight was 0.10.

elasticity of RCA2 most. Both the Aggregate Adjustment or Individual-Level Utility Adjustment methods increase the price elasticity as well, but at a lesser rate than the loss of PIP.

Which adjustment is more appropriate? Again, it depends on your assumptions regarding why the share should be adjusted, and how the change should be manifest in elasticity.

Aggregate or Individual-Level Utility Adjustments are more appropriate:

- If you expect that changing a product's shares should have a modest impact on the self-elasticity of the product. This is especially the case when corrections principally are due to distribution or awareness. As an example, if a product needs to have its share dramatically reduced because it is only distributed in 1/10th of the market, then its price elasticity should remain constant, despite the fact it has a much lower base. Again, however, adjustments for distribution can be handled in a more appropriate way prior to implementing adjustments for external effects.

Respondent Weights adjustment is more appropriate:

- If you expect that changing a product's share should have a stronger effect on the self-elasticity of the product than is seen with the other methods. If we are missing either some negative or very positive aspects related to the features of the product in our model, then perhaps its price sensitivity should dramatically change as well.

Test #3

In this test, we'll examine the effect of external effects on the relative relationships of self-elasticities and cross-elasticities among products when all product shares are adjusted. Let's consider the previous base case:

	Base Share
JVC1	9.49
JVC2	18.65
RCA1	9.06
RCA2	28.52
Sony1	14.39
Sony2	19.89

Using the market simulator, we can compute the self-elasticities and cross-elasticities resulting from a 10% increase in price for each product.

Elasticities—Effect of Increase in Row Product's Price on Column Product's Share

	JVC1	JVC2	RCA1	RCA2	Sony1	Sony2
JVC1	-2.50	0.35	0.77	0.17	0.27	0.07
JVC2	1.39	-3.82	0.52	1.18	0.77	0.44
RCA1	0.69	0.21	-3.22	0.27	0.42	0.26
RCA2	1.29	2.05	1.47	-3.21	1.18	0.55
Sony1	0.69	0.58	0.95	0.50	-3.80	0.73
Sony2	0.21	0.60	0.81	0.40	0.68	-2.09

The negative values along the diagonal represent self-elasticities. Off-diagonal elements are positive, and reflect cross-elasticities. For example, reading across the first row, an increase of 10% in JVC1's price results in a 3.5% increase in JVC2's share (a cross elasticity of +0.35). The higher the cross-elasticity, the more strongly two products referenced by that cell compete. Also, the larger the share of the row product, the larger the effect of its share changes upon the percentage increase in the column products' shares. Thus, RCA2 (with a base share of nearly 29%) has much higher cross-elasticities across its row than RCA1 (with a base share of around 9%). That is because when RCA2 raises price, it loses larger absolute share (which is redistributed among the other products) than when RCA1 increases price.

As stated before, JVC2 and RCA2 offer the same product features at the same price. When JVC2 increases its price by 10%, RCA2's share increases by 11.8% (a cross-elasticity of +1.18). When RCA2 raises its price by 10%, JVC2's share increases by 20.5% (a cross-elasticity of +2.05).

In this third test, we will apply external effect adjustments to all six products' shares, and compute a new cross-elasticity matrix. We'll compare the new self- and cross-elasticities to the original self- and cross-elasticities prior to adjusting for external effects.

The base case simulated shares and target market share values (hypothetical values used for this example) are as following:

	Base Share	Target Share	Adjustment
JVC1	9.49	14.00	+48%
JVC2	18.65	23.00	+23%
RCA1	9.06	6.00	-34%
RCA2	28.52	22.00	-23%
Sony1	14.39	23.00	+60%
Sony2	19.89	12.00	-40%

After adjustments, the resulting tables of cross elasticities are as follows:

Cross-Elasticities after Aggregate Adjustment

	JVC1	JVC2	RCA1	RCA2	Sony1	Sony2
JVC1	-2.41	0.47	0.90	0.29	0.38	0.19
JVC2	1.58	-3.72	0.70	1.37	0.95	0.61
RCA1	0.54	0.07	-3.31	0.13	0.28	0.12
RCA2	0.89	1.62	1.07	-3.45	0.78	0.17
Sony1	1.13	1.01	1.39	0.93	-3.55	1.17
Sony2	0.00	0.38	0.59	0.18	0.46	-2.25

(Note the near-reversal for Sony2 upon JVC1's share at the bottom-left corner. This is a near-example of the anomaly we discussed earlier.)

Cross-Elasticities after Individual Utility Adjustment

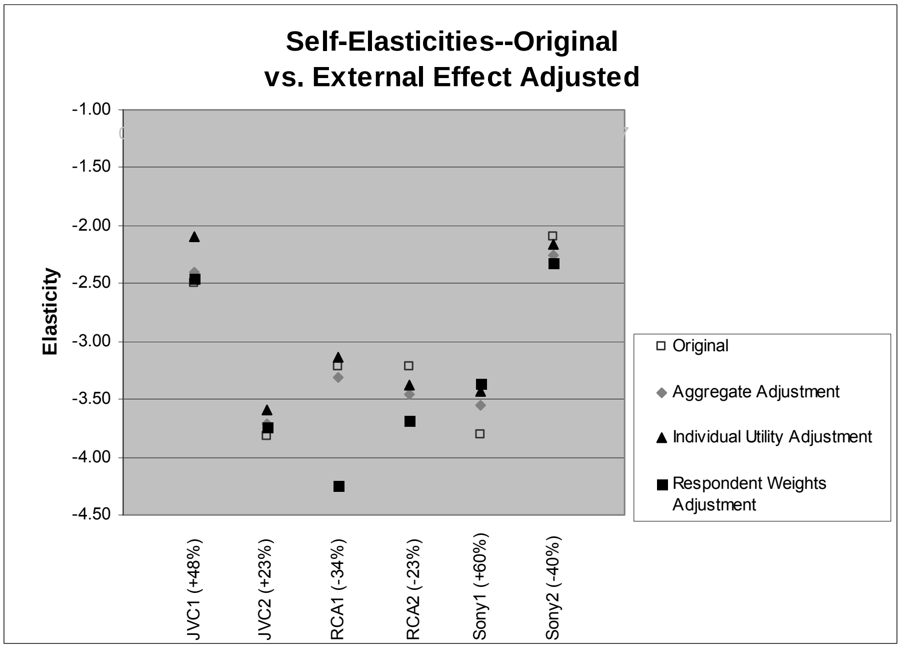
	JVC1	JVC2	RCA1	RCA2	Sony1	Sony2
JVC1	-2.10	0.43	0.94	0.21	0.38	0.09
JVC2	1.30	-3.60	0.69	1.40	0.83	0.55
RCA1	0.29	0.14	-3.14	0.19	0.25	0.19
RCA2	0.78	1.37	1.24	-3.38	0.86	0.41
Sony1	1.05	0.91	1.62	0.92	-3.43	1.19
Sony2	0.10	0.35	0.50	0.24	0.40	-2.16

Cross-Elasticities after Respondent Weights Adjustment[9]

	JVC1	JVC2	RCA1	RCA2	Sony1	Sony2
JVC1	-2.45	0.46	1.62	0.32	0.24	0.12
JVC2	1.47	-3.74	0.82	1.57	0.77	0.68
RCA1	0.58	0.16	-4.25	0.23	0.25	0.25
RCA2	1.02	1.49	1.45	-3.68	0.73	0.57
Sony1	0.76	0.81	1.81	0.80	-3.37	1.63
Sony2	0.13	0.37	0.65	0.29	0.32	-2.33

These tables present a large amount of information that is difficult to grasp unless we dissect the pieces. We'll begin by comparing the self-elasticities in the chart below, before and after the external effect adjustments.

[9] Respondent weights to adjust base case shares to match target shares were as follows: Max weight=2.41, Min Weight=0.28, Standard Deviation=0.77. The ratio of respondent weights from maximum to minimum is 8.6X, which is quite extreme. A clear disadvantage of Respondent Weighting is that the weights may become quite extreme whenever shares need to be adjusted considerably or when the sample is relatively homogeneous.



Note that the elasticities have changed very little from the original baseline values when applying Aggregate Adjustment and Individual Utility Adjustments. The Respondent Weights adjustment increases the elasticity quite dramatically for the product whose resulting share after adjustment was smallest (RCA1).

Again, which adjustment method is preferred depends on the reasons why the analyst believes the correction is needed. Adjustments to share due to distribution or awareness should not change elasticities. If the product's share is adjusted because the brand equity or desirability of the features is higher or lower than the model predicts, then changes in share should also result in changes to elasticities. In our example, we have decreased RCA1's share from 9.06 to 6.0 (a 34% decrease in share). The Respondent Weights approach leads to much higher price elasticity for the 6% share RCA1.

Let's now turn our attention to the off-diagonal data entries, the cross-elasticities. The patterns of cross-elasticities are quite similar across all four tables: prior to external effect adjustments, and after three methods for adjusting shares of preference. The correlations among the methods are:

Correlations among Cross-Elasticities

	Prior to Adjustment	Agg Adj	Indiv UtilAdj	Resp Wts
Prior to Adjustment	1.00			
Agg Adj	0.81	1.00		
Indiv UtilAdj	0.78	0.96	1.00	
Resp Wts	0.76	0.90	0.94	1.00

The Aggregate Share adjustment cross-elasticities are most similar to the original cross-elasticities, prior to adjustment. However, we should probably expect that cross-elasticities should change due to the relatively large shifts in shares we imposed upon the products. So, this is not necessarily a mark of success.

The Aggregate and Individual Utility Adjustment methods are highly correlated at 0.96. These approaches indeed result in very similar cross-elasticities after adjustments to shares.

ECONOMIC MODELING VIA REGRESSION ANALYSIS:

It is difficult to judge from the tables of cross-elasticities and from the correlation analysis which external effect adjustment makes most sense. We'll apply a multiple regression model to investigate how consistent the methods are with respect to economic theory. The percentage change in Product j's share due to a change in Product i's price is a function of the shares of products i and j and the amount of similarity between i and j. More formally,

$$\text{Cross\_Elasticity}_{ij} = f\,(\text{Share}_i + \text{Share}_j + \text{Product\_Similarity}_{ij})$$

Here, we define the product similarity of products i and j as the total number of attribute levels shared in common.

Here are the results of the regression equation for the original model (prior to correction with external effects), and after each of the external effect adjustments we've been investigating:

Original Shares (Prior to Adjustment)

R-Squared	0.819		
Constant	-0.079		
	Coeff.	Std Err	T-Value
Share$_i$ (Row Share)	0.040	0.00569	7.11
Share$_j$ (Column Share)	-0.015	0.00569	-2.67
Similarity$_{ij}$ (# Shared Levels)	0.196	0.02902	6.76

Aggregate Share Adjustment

R-Squared	0.800		
Constant	-0.129		
	Coeff.	Std Err	T-Value
Share$_i$ (Row Share)	0.042	0.00674	6.17
Share$_j$ (Column Share)	-0.013	0.00674	-1.91
Similarity$_{ij}$ (# Shared Levels)	0.189	0.03415	5.55

Individual Utility Adjustment

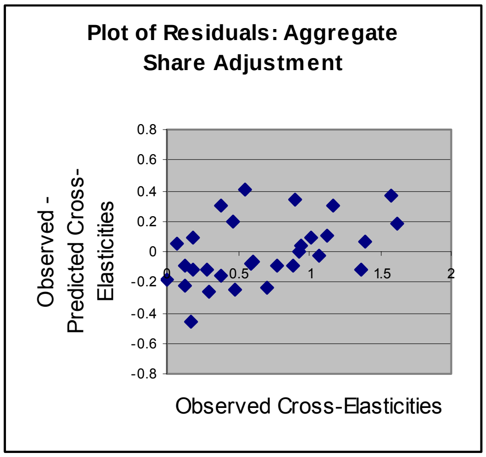
R-Squared	0.827		
Constant	-0.061		
	Coeff.	Std Err	T-Value
Share$_i$ (Row Share)	0.042	0.00599	6.99
Share$_j$ (Column Share)	-0.016	0.00599	-2.73
Similarity$_{ij}$ (# Shared Levels)	0.171	0.03038	5.62

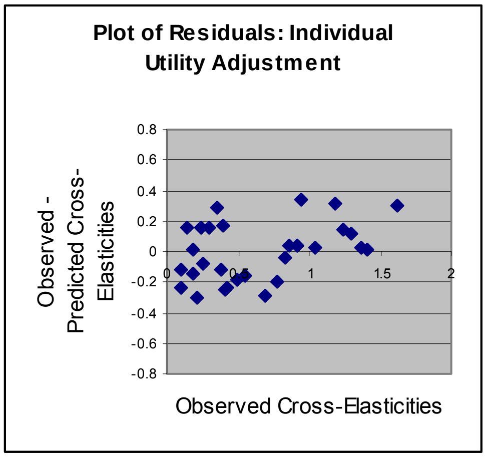
Respondent Weights Adjustment

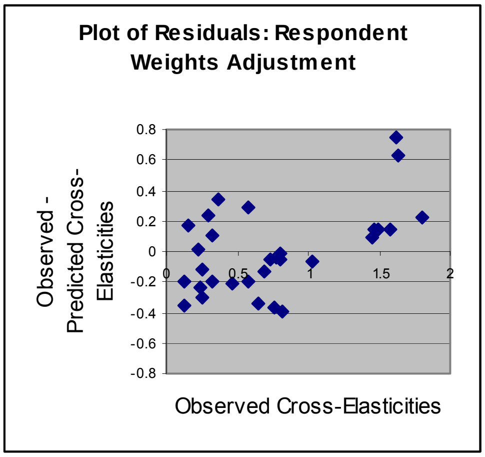
R-Squared	0.720		
Constant	0.357		
	Coeff.	Std Err	T-Value
$Share_i$ (Row Share)	0.038	0.00889	4.24
$Share_j$ (Column Share)	-0.034	0.00889	-3.83
$Similarity_{ij}$ (# Shared Levels)	0.190	0.04502	4.22

The overall model fit of the best models is quite good (over 80% of the variance in the cross-elasticities explained). In all models, the coefficients have the expected directions and are all significant (t>1.96), with one exception, at the 95% confidence level. However, the model fit is noticeably decreased in the case of Respondent Weights adjustments. This analysis would suggest that the other methods more closely align with economic theory—especially the Individual Utility Adjustment method. If the adjustments to product shares (and resulting respondent weights) weren't so extreme, we'd expect the Respondent Weights method to perform better. But to test the comparative qualities of the approaches, it's useful to try more extreme cases.

Below, we have plotted the residuals from the multiple regressions.







The cases with the largest observed cross-elasticities seem to be slightly overstated, relative to the economic theory regression model. This is especially the case with the Respondent Weights adjustment, with two data points having residuals > 0.6 (associated with the smallest Column share products RCA1@6% and Sony2@12%). The Respondent Weights adjustment dramatically overstates the cross-elasticity in these cases.

RECOMMENDATIONS AND CONCLUSIONS:

We began by discussing why simulated market shares often do not match target market shares. We covered methods for adjusting for awareness and availability, with special emphasis on a simulation and sampling method for simulating repeated respondent "shopping trips" within sampled "stores/regions," each reflecting available product offerings. We're confident about this approach, and suggest analysts use it whenever the data permit and prior to other external effect adjustments.

Adjustments based on scale factor should always be investigated, given proper market share data, after appropriate corrections have been made for distribution (and possibly awareness). This is a legitimate and often necessary adjustment to control for noise and consistency between two samples and choice environments.

This research suggests that there is no clear winner among the three methods for adjusting shares via external effects. The method you choose certainly depends on the modeling situation and by how much the shares need to be adjusted. But we can characterize some general strengths and weaknesses of the approaches.

- The method of **Aggregate Adjustment** as currently offered in the Sawtooth Software simulator has drawbacks (such as the "reversal anomaly") and is naïve in the way it proportionally re-distributes share. But, it is easy to implement and makes only modest adjustments to self- and cross-elasticities.

- **Individual Level Utility Adjustment** is quite similar in spirit to the Aggregate Adjustment, but leverages the power of individual-level data, focusing changes on respondents on-the-cusp of choice. Shares are not simply re-distributed in proportion to previous shares, as with the Aggregate Adjustment, but are reapportioned according to expected switching patterns among products with varying similarity and correlations in preference. This approach also avoids the "reversal anomaly" that we demonstrated. It also makes only modest adjustments to self- and cross-elasticities, in the expected direction depending on whether the product's share is increased or decreased, which in many cases is the desired outcome.

- **Respondent Weighting** is theoretically best when corrections need to be made due to improper sample composition. However, one should watch for extreme weights, when some respondents are weighted much more heavily (especially 5x or more) than others. In some cases, no solution is possible using positive weights. Respondent re-weighting is not appropriate for corrections dealing with distribution and awareness (which should be handled in other ways). Self-elasticities can become relatively more extreme for adjusted products through respondent weighting. Results from the cross-elasticity test and regression modeling suggest that certain cross-elasticities (substitution effects) may be exaggerated beyond the expectations supported by the economic model we imposed. The variance of the respondent weights directly affects the performance of this adjustment approach. When the variance of respondent weights is not as extreme as in our test case, the results should be greatly improved.

In general, we suggest researchers avoid adjusting simulated shares to match target market shares. But, if you must, we suggest corrections be made in this order:

1. Corrections for availability (do not allow respondents to allocate share to product alternatives that are not available to them).
2. Corrections for awareness, if achieving awareness is a significant hurdle to purchase consideration. Adjustments for awareness are not as straightforward, and we suggest caution in applying them.
3. Scale Factor. Such adjustments are quite appropriate and defensible (given a good overall conjoint model and that the previous steps are undertaken).
4. Product share adjustments, either through Individual Utility Adjustment or a modest degree of Respondent Weighting, depending on what qualities you believe are lacking in the model and the degree to which shares need adjustment. Hopefully, if you have designed your study well and have properly accounted for distribution, the needed adjustments will be modest.

APPENDIX A

AGGREGATE SHARES CORRECTION ANOMALY

In the body of the paper, we showed how the standard External Effect adjustment as applied in the Sawtooth Software simulator can cause JVC1's share to actually *decrease* when Sony2 raises its price. This example is taken from an HB run for a real data set (the "TV" example data set and accompanying HB run that ships with Sawtooth Software's market simulator). In case the reader is interested to replicate the results using the "TV" data set, the base case scenario is as follows:

Base Case Product Specifications:

	Brand	Screen	Sound	Channel		
	Size	QualityBlockout			PIP	Price
JVC Low End	1	1	1	2	2	300
JVC High End	1	2	2	2	2	375
RCA Low End	2	1	2	1	1	325
RCA High End	2	2	2	2	2	375
Sony Low End	3	1	2	2	1	350
Sony High End	3	3	3	1	1	400

We'll repeat the example below:

	Base	Sony2	
	Case	+$50	Change
JVC1	9.49	9.71	+0.22
JVC2	18.65	19.96	+1.31
RCA1	9.06	9.94	+0.88
RCA2	28.52	29.84	+1.32
Sony1	14.39	15.54	+1.15
Sony2	19.89	15.00	-4.89

Assume the following external effects:

JVC1	1.0
JVC2	1.0
RCA1	1.0
RCA2	1.0
Sony1	1.0
Sony2	0.5

After applying external effects, the new results for this base case and price increase for Sony2 are:

	Base Case	Sony2 +$50	Change
JVC1	10.53	10.50	-0.03
JVC2	20.71	21.58	+0.87
RCA1	10.07	10.74	+0.67
RCA2	31.67	32.26	+0.59
Sony1	15.98	16.80	+0.82
Sony2	11.04	8.11	-2.93

JVC1's share decreases when Sony2 raises its price! This result is quite counterintuitive and can be disconcerting to analysts and managers alike.

How does this happen? Let's consider the original simulation, prior to applying external effects. The respondents who prefer Sony2 are not very likely to switch into JVC1 when Sony2 raises its price. Note that JVC1 only gains +0.22 in share with Sony2's price increase (9.71 – 9.49 = 0.22).

Under the external effect adjustment, Sony2's shares are reduced by a multiplicative factor of 0.5. The lost share from Sony2 is then distributed among the other offerings in proportion to their original shares:

	Base Case		Ext. Eff.			Renormalize to 100%
JVC1	9.49	x	1.0	=	9.49	9.49 / 90.06 * 100 = 10.53
JVC2	18.65	x	1.0	=	18.65	18.65 / 90.06 * 100 = 20.71
RCA1	9.06	x	1.0	=	9.06	9.06 / 90.06 * 100 = 10.07
RCA2	28.52	x	1.0	=	28.52	28.52 / 90.06 * 100 = 31.67
Sony1	14.39	x	1.0	=	14.39	14.39 / 90.06 * 100 = 15.98
Sony2	19.89	x	0.5	=	9.95	9.95 / 90.06 * 100 = 11.04
			Totals:		90.06	100.00

Thus, the loss in share for Sony2 due to the 0.5 external effect adjustment (see last row of the table above) is redistributed according to the proportions of the competitive products, and JVC1 picks up 1.04 share points (from 9.49 to 10.53).

After Sony2 increases its price, its unadjusted share decreases from 19.89 to 15.00. The shares are redistributed according to the External Effect multipliers as below:

	Sony2 + $50		Ext. Eff.			Renormalize to 100%
JVC1	9.71	x	1.0	=	9.71	9.71 / 92.50 * 100 = 10.50
JVC2	19.96	x	1.0	=	19.96	19.96 / 92.50 * 100 = 21.58
RCA1	9.94	x	1.0	=	9.94	9.94 / 92.50 * 100 = 10.74
RCA2	29.84	x	1.0	=	29.84	29.84 / 92.50 * 100 = 32.26
Sony1	15.54	x	1.0	=	15.54	15.54 / 92.50 * 100 = 16.80
Sony2	15.00	x	0.5	=	7.50	7.50 / 92.50 * 100 = 8.11
			Totals:		92.50	100.00

Prior to the price increase by Sony2, a larger absolute share from Sony2 was shifted to the other brands due to external effect adjustments, resulting in JVC1 picking up 10.53 – 9.49 = 1.04 share points. After Sony2's price increase, Sony2 of course loses overall share. When Sony2's new lower share is shifted to the other brands due to the external effect factors, it results in JVC1 picking up 0.79 share points (10.50 – 9.71), which is 0.25 fewer points than before due to the external effect adjustment. This loss for JVC1 cannot counteract its relatively small gain in share of +0.22 (9.71 – 9.49) due to Sony2's price increase prior to applying external effects. The net change to JVC1 is 0.22 – 0.25 = -0.03.

CONFOUND IT! THAT PESKY LITTLE SCALE CONSTANT MESSES UP OUR CONVENIENT ASSUMPTIONS

JORDAN LOUVIERE
UNIVERSITY OF TECHNOLOGY, SYDNEY
THOMAS EAGLE
EAGLE ANALYTICS, INC.

INTRODUCTION

Since the early 1990s there has been much progress in understanding and taking into account preference heterogeneity in probabilistic discrete choice models (e.g., Wedel and Kamakura 1999; McFadden and Train 2000). The vast majority of models applied in marketing and applied economics try to represent heterogeneity as some type of discrete or continuous distribution of preferences. These relatively new types of statistical models have done well in comparisons against simpler model forms like conditional multinomial logit in terms of in- and out-of-sample fits, with fit performance often assessed against so-called "hold-out" sets. It is fair to say that these models are long on statistical theory, but short on behavioral theory; the latter aspect is the focus of this paper.

In particular, scientists typically spend time a) formulating hypotheses and/or theory, b) testing deductions from theory and/or hypotheses, or c) testing assumptions underlying theory and/or hypotheses. The focus of this paper is on testing certain key assumptions about error distributions implicit in all limited dependent variable models. It is rare to see these assumptions tested, but it is important to do this to allow researchers to know if the assumptions are satisfied, and to discuss the consequences if they are not.

To preface our discussion and results, we note that all statistical models in which the dependent variable is latent confound estimates of model parameters with error variability. Thus, if error variances are not constant across individuals and choice sets, estimated model parameters will vary with differences in error variances. Moreover, if error variances are not constant one cannot estimate unconfounded discrete or continuous distributions of preferences, and the consequences of not satisfying this assumption can be serious. This paper focuses on the error variance (or "unobserved heterogeneity") associated with probabilistic discrete choice models. We first discuss the role of the error variance in simple and more complex choice models, then we discuss and illustrate the confound between model parameters and error variance, and then we present and review academic research that demonstrates that it is unlikely that error variances are constant, but instead it is much more likely that the error variance is systematically related to a number of factors that we outline and discuss. Then we propose and discuss ways to deal with systematic variation in error variances. We conclude with some cautions about making claims based on current models, and summarize the key points in the paper.

The Scale Parameter in MNL

The familiar MNL model is a random utility model in which the errors are assumed to be independent and identically distributed (iid) as Type 1 Extreme Value random variates. That is, recall the familiar axiom of random utility theory:

$$U_{jn} = V_{jn} + \varepsilon_{jn}, \qquad (1)$$

where U_{jn} is the latent utility that individual n associated with choice option j; V_{jn} is the systematic or mean utility that individual n associates with option j; and ε_{jn} is the random or stochastic component of the utility of option n for individual n, which is assumed to be distributed as Extreme Value Type 1. If individual n seeks to maximize her utility, we can model the probability that she chooses option j as follows:

$$P(j|C_n) = P[(V_{jn} + \varepsilon_{jn}) > (V_{in} + \varepsilon_{in})], \text{ for all j options in choice set } C_n, \qquad (2)$$

where all terms are as previously defined, except for $P(j|C_n)$, which is the probability that option j is chosen by a randomly chosen individual n facing choice set C_n. If the errors are iid Extreme Value Type 1, expression (2) leads to a closed-form expression for the choice probabilities called the MNL (multinomial logit) choice model (McFadden 1974). The consequences of these error assumptions are that individuals are "preference clones" who share the same fixed set of preference parameters. Variability in choices arises due to analysts' misspecification of true utility functions, inability to account for all relevant factors in choice, and other omissions and commissions.

These assumptions imply that the main diagonal of the error variance-covariance matrix is constant, and all off-diagonal error covariances equal zero. The vector of unknown model parameters can be expressed as a generalized regression function:

$$V_{jn} = \Sigma_k\ \beta_k X_{kn} + \varepsilon_{jn}, \qquad (3)$$

where all terms are as previously defined, except for β_k and X_{kn}. β_k is a vector of empirical parameters associated with a vector of factors that underlie choices, X_{kn}. The vector of parameters is subscripted only with respect to the factors ("attributes") because they are fixed for all individuals. The vector of factors is doubly subscripted to indicate that factors vary over k dimensions but also (potentially) over people. The confound between error variance and the estimated parameters can be expressed as follows:

$$V_{jn} = \Sigma_k\ \lambda\ \beta_k X_{kn} + \varepsilon_{jn}, \qquad (3)$$

where all terms are as previously defined, except for λ, which is a "scale parameter". Formally, in the case of MNL, $\lambda = \text{SQRT}(\pi^2 / 6\ \sigma_\varepsilon^2)$, where π is the natural constant 2.141…, and σ_ε is the standard deviation of the error distribution. We say that λ "scales" the vector of parameters because each parameter, β_k, actually is $\beta_k/\sigma_\varepsilon$.

IMPLICATIONS OF SCALE IN MNL

The confound of scale and model parameters creates a fundamental identification problem, with the consequence that MNL model parameters cannot be identified unless λ is fixed to some constant (almost always 1.0). Thus, the parameters "are identified up to scale", which means that they can be identified once a constant is selected. The confound has no impact on predicted probabilities in MNL if the error assumptions are satisfied. This confound is not new, and discussions can be found in many sources like Ben-Akiva and Lerman (1985), Swait and Louviere (1993) and Louviere, Hensher and Swait (2000). What is new is that more researchers now realize that it is unlikely that error variances (and hence, scale) are constant in empirical data; instead, it is more likely that error variances systematically vary with attribute levels varied in choice experiments and real markets, as well as differences in individuals. For example, we later show that error variances systematically vary with levels of attributes, and there is evidence that survey respondents with low literacy skills have higher error variances than those with higher skills (See Louviere, Hensher and Swait 2000, Chapter 13).

A consequence of the confound is that it impacts magnitudes of estimated model parameters, and by implication, statistical inference. Specifically, smaller error variances (large scales) lead to larger model parameters, while larger error variances (small scales) lead to smaller model parameters, all else equal. Not surprisingly, standard errors of parameter estimates also are impacted - smaller scales lead to less precise estimates. Discussions of such issues can be found in workshop reports published in Marketing Letters since the advent of the modern Invitational Choice Symposia that were proposed and initially organized by one of the present authors (1989, Banff, Alberta, Canada). For example, discussion of the scale parameter ands its implications are in Louviere, et al, (1999), Louviere, et al (2002) and Louviere, et al (2006). We now turn our attention to the consequences if error variances are not constant.

WHAT IF SCALE IS NOT CONSTANT?

It is surprising that researchers have largely focused on preference heterogeneity to the exclusion of most other likely sources of unobserved variability. There has been little research into variance component models for discrete choices (for an exception, see Cardell 1997) that explicitly recognize that errors can be decomposed into systematic components associated with differences within and between individuals (the latter can be preference heterogeneity, but may be related to other factors that differ between individuals), environmental and context differences, temporal and spatial differences and other sources. Louviere, et al (1999) discuss these issues in detail, so we only briefly summarize them here.

They suggest that scale is impacted by many factors that can be summarized in the following general expression:

$$Y \mid X, Z, C, G, T, \qquad (4)$$

where Y = behavioral outcomes of interest; X = directly observable or manipulated variables; Z = characteristics of the individuals; C = factors that vary over conditions, contexts, circumstances, or situations; G = geographical, spatial or environmental factors that are relatively constant in one place, but may vary from place to place; and T = time varying factors. Thus, it is highly unlikely that scale is constant; it is much more likely to be systematically impacted by a wide array of factors. Also, error variability is unlikely to be unidimensional, but probably varies a) within consumers; b) between consumers; c) with measurement instruments; d) with market and environmental differences; and e) with many other sources.

Any one data source confounds many of the above sources of error variability, as discussed by Louviere, et al (1999, 2002). So, most researchers have samples of size = 1, but claim meaningful results! Single data sources limit generalizeability in many cases where potential sources of – say – temporal and spatial variability are constant, making it unclear how to generalize to past or future time periods or spatial locations. Thus, much more research into combining sources of preference and choice data is needed.

SCALE OR PREFERENCE HETEROGENEITY?

During the past decade many published choice models assumed that individuals are heterogeneous in their preferences. Very few publications suggested that individuals also might be heterogeneous in their error variances (or scales). We earlier noted that model parameter estimates and scale are confounded; hence, if scale varies across individuals, distributions of preference parameters will be confounded with distributions of scales. Later we provide evidence that this occurs in empirical data, but we note that because model estimates actually are $\beta_k/\sigma_\varepsilon$, it should be clear that one can have a distribution of true β_k if and only if σ_ε is constant. If β_k and σ_ε vary across individuals, observations, contexts, time and space, one cannot estimate a distribution of β_k without separating β_k and σ_ε. Of course, one can try to capture these effects by estimating higher moments of assumed distributions, but adding more statistical complexity in the form of additional latent effects does not seem warranted in the absence of better behavioral theory.

That is, little behavioral theory is evident in recent choice modeling papers; most authors rely on statistical theory. Limitations of single data sources suggest that without theory to suggest how components of variance differ by individuals, markets, contexts, experiments, etc, adding higher moments to choice models is probably a bad idea unless these distributions are constant over such sources of variation. One only needs to consider predicting infrastructure projects like toll roads or bridges years into the future to see that estimating higher moments is a bad idea as they also have to be forecast into the future. A better way forward is to develop theory and methods to capture variability differences. We now show how easy it is to confuse heterogeneity in model parameters and scale, which should give researchers reasons to think before estimating higher moments.

Consider two cases involving ten people in Table 5: 1) everyone has identical preferences for travel times and costs of public bus systems, with scales equal to 1.0; only intercepts differ; 2) scales vary across people, holding everything else constant. A key takeaway is that if scales vary across people, they will seem heterogeneous in preferences even if they differ only in scale. The last two columns show the processes are equivalent.

Table 5: Consequences of Scale Varying Across Individuals

Condition 1: Only intercepts vary **Condition 2: intercepts & scale vary**

Person	Inter 1	time1	cost1	scale1	u1	Inter 2	time2	cost2	scale2	u2	u1*scale2
0	-1.00	-1.5	-1	1	1.50	-0.20	-0.30	-0.20	0.20	0.30	0.30
1	-0.75	-1.5	-1	1	1.75	-0.60	-1.20	-0.80	0.80	1.40	1.40
2	-0.50	-1.5	-1	1	2.00	-0.70	-2.10	-1.40	1.40	2.80	2.80
3	-0.25	-1.5	-1	1	2.25	-0.25	-1.50	-1.00	1.00	2.25	2.25
4	0.00	-1.5	-1	1	2.50	0.00	-3.00	-2.00	2.00	5.00	5.00
5	0.25	-1.5	-1	1	2.75	0.15	-0.90	-0.60	0.60	1.65	1.65
6	0.50	-1.5	-1	1	3.00	0.20	-0.60	-0.40	0.40	1.20	1.20
7	0.75	-1.5	-1	1	3.25	1.35	-2.70	-1.80	1.80	5.85	5.85
8	1.00	-1.5	-1	1	3.50	1.20	-1.80	-1.20	1.20	4.20	4.20
9	1.25	-1.5	-1	1	3.75	2.00	-2.40	-1.60	1.60	6.00	6.00

The above example assumes that time1 = -1 and cost1 = -1

That is, the only difference in the left-hand and right-hand side parameters is scale. That is, if one multiplies the left-hand side parameters by scale2, one obtains the right-hand side parameters. The last two columns show that multiplying u1 (overall utility for a bus described by time = -1 and fare = -1 in condition 1) by scale2 produces the same outcome as u2 (overall utility for a bus described by a time = -1 and a fare = -1 in condition 2).

Differences in scales across people can be consequential, as we now show. For example, let person 0 from Table 5 face a choice among three buses as shown below:

Bus #	Time	Cost	Utility from left-hand side	Choice Prob	Utility from right-hand side	Choice Prob
1	-1	+1	-0.5	0.506	-0.1	0.367
2	+1	-1	-1.5	0.186	-0.3	0.301
3	0	0	-1.0	0.307	-0.2	0.332

If person 0 significantly differs from a scale of 1.0, which in this case is a scale of 0.2 (a smaller scale, or larger error variance), choice probabilities can differ a lot even though the person's preferences for time and cost do not change. Thus, it is not possible to make meaningful statements about preference parameter distributions without taking scale differences into account at the same time.

How and Why We Know That Variances (Scales) Differ

There has been a fair amount of research on scale. For example, reviews of research in this area can be found in Louviere, Hensher and Swait (2000), Louviere, et al (2002), Louviere, et al (2006); empirical work on size and scopes of error assumption violations can be found in Train and Weeks (2005), Louviere and Islam (2004), Louviere (2004), DeShazo and Fermo (2002), Swait and Adamowicz (2001a,b), Louviere, et al (2002), Dellaert, Brazell and Louviere (1999) and Ohler, et al (2000), to name only a few. The preceding discussion and cited sources discuss several consequences if error variances are not constant, such as:

1. Estimates of price sensitivities (elasticities) or other policy effects may be misleading;
2. Willingness-to-pay (WTP) or other policy estimates may be misleading.
3. Forecasts may be biased and may depart significantly from reality.
4. Randomness in parameters confounds scale and real preference heterogeneity.
5. Hyperparameters in HB and/or MIXL are affected, and Latent Class model differences may be misleading.

In particular, differences in people may be reflected in parameter estimates, scale differences and/or some combination of both. The table below shows that individuals can be in the same part of the probability distribution with small parameters and a large scale or small scale and large parameters. These outcomes are observationally equivalent, which implies that more research is needed to separate parameters and scale. The table also suggests that researchers need to recognize that scale can be related to factors varied in choice experiments, factors that managers control, individual differences, environment, context, temporal and/or spatial differences. We say this again to emphasize that ignoring variability sources and/or assuming them away is dangerous and limits generalizeability. It also begs the question of how to deal with scale; later we discuss two ways to deal with scale in choice experiments.

	True Preference Parameters	
Scale	**Small**	**Large**
Small	Complimentary Processes	Potentially Observationally Equivalent
Large	Potentially Observationally Equivalent	Complimentary Processes

As noted by several authors, such as Louviere, et al (1999), one must combine multiple sources of data to make real progress in separating components of variance. This is not merely an academic issue because one cannot correctly predict choices in real markets if variance components differ between model estimation data sources and the market(s) to be predicted. So, instead of being seen as a convenient (or annoying) statistical assumption, researchers should begin to recognize that scale (or error variance) is a behavioral phenomenon that needs to be understood in its own right.

We now discuss how scale varies with choice experiment design. Example one is from an honours thesis by Chelsea Wise (Louviere and Wise 2004). She designed four conditions to vary attributes: 1) only brand and price, 2) brand, price + four less salient attributes, 3) brand, price + four more salient attributes; and 4) brand, price + all attributes. Figure 1 graphs the relative variability from each of these conditions. The least variability occurs for the least complex task (only brand and price), followed by a task with brand, price and more salient attributes. Higher variability occurred for tasks with less salient attributes and all attributes. These results are not "surprising". It is surprising that inequality of variances in choice data has received so little research attention.

Figure 1: Relative Variance Differences in Experimental Conditions

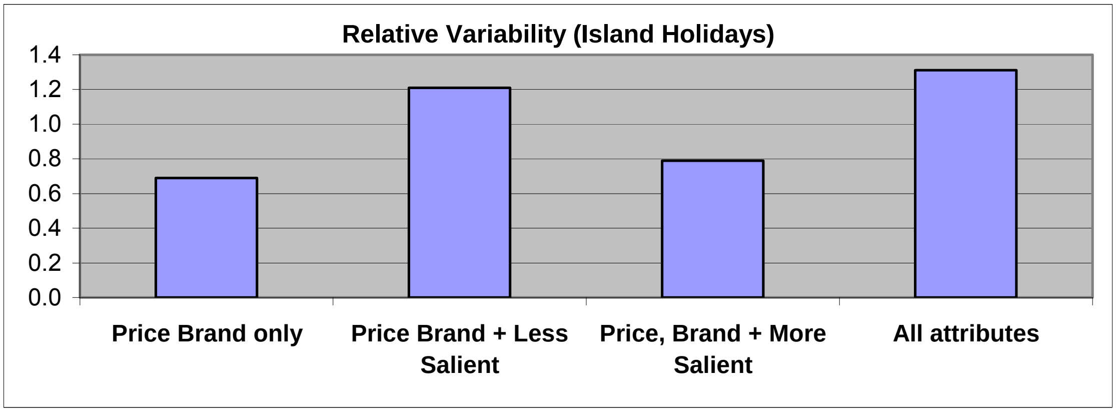


Figure 2 is the conditional (mean) price response curves estimated from Wise's four conditions, which show that price slopes differ systematically with relative variability in each condition. Thus, the four conditions exhibit different choice probabilities, and the differences are due to differences in choice variability.

Figure 2: Differences in Price Response Curves by Experimental Condition

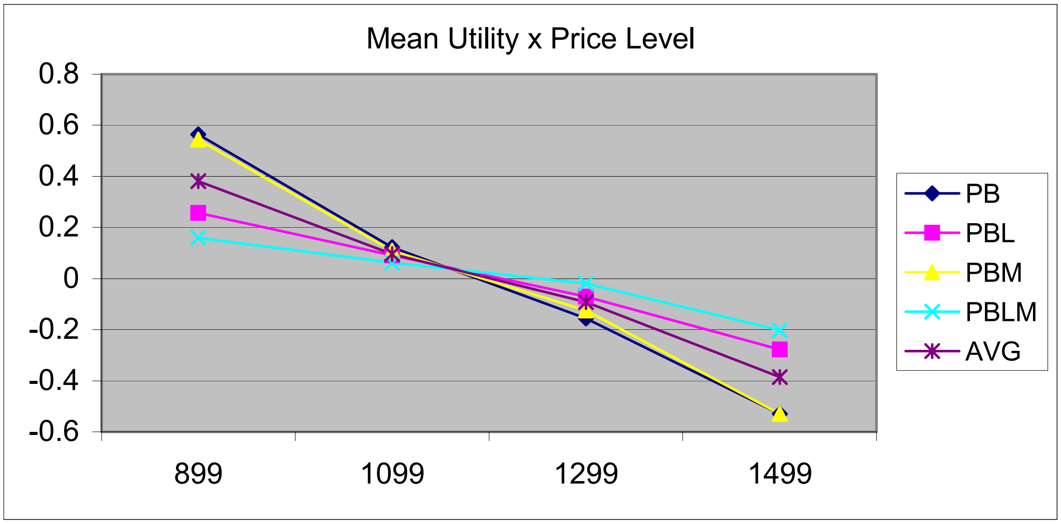
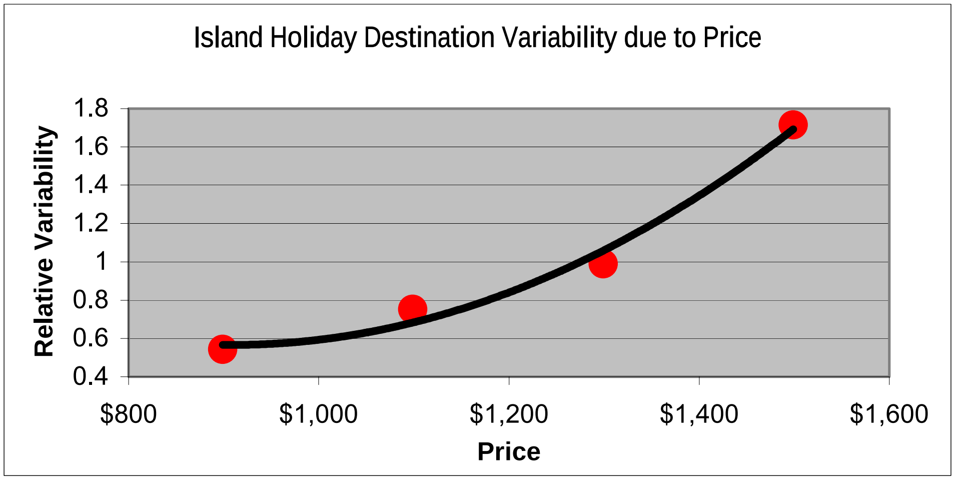


Figure 3 shows a relationship between relative variability and price levels in Wise's data. Each point is an estimate of the variability for each price level, analogous to a random coefficients result with a different standard deviation for each price level. The graph shows choice variability is not constant, but varies systematically with price levels.

Figure 3: Relationship between Relative Choice Variability and Price Levels



Islam and Louviere (2004) provide another example. They designed 64 experiments to vary attribute presence/absence. Brand and price were always present; the design dictated other attributes that vary. Figure 4 shows relative choice variability associated with each level of percent real fruit juice or the juice price. In the case of percent juice, there also is an estimate of the effect of this attribute being "missing".

Figure 4: Conditional Means for Percent Juice and Associated Variability

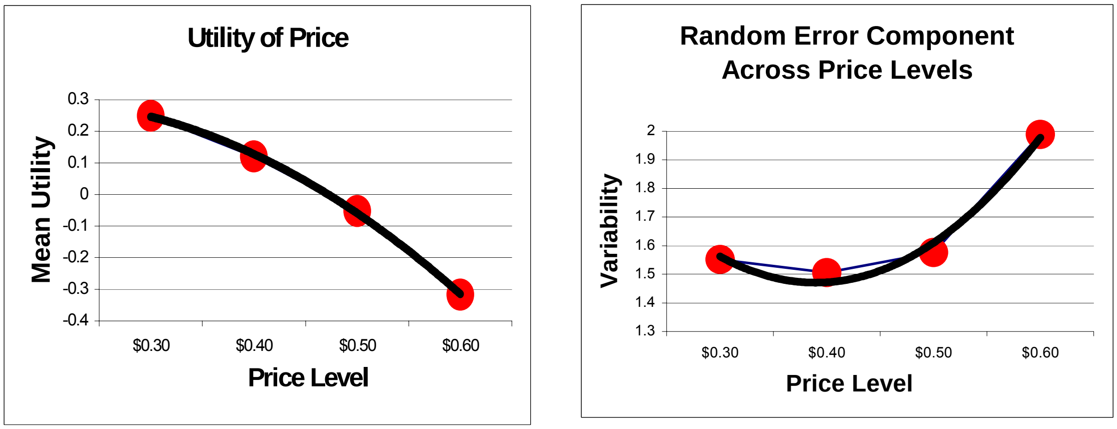
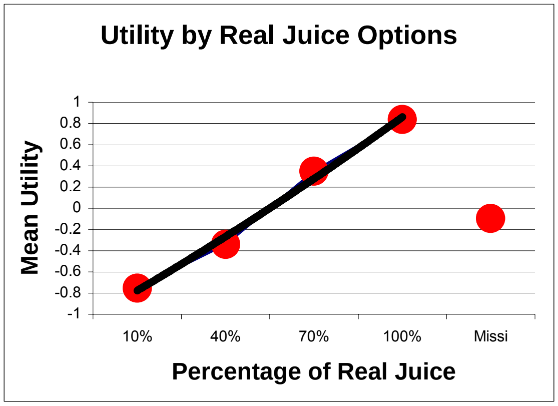
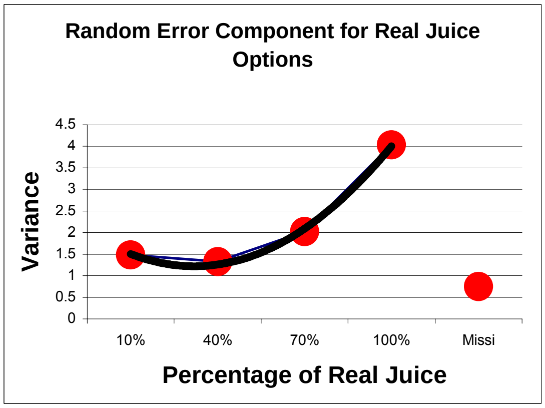


Figure 4 shows that the relative variance is an inverse-U shaped function of the percent real juice and price levels. Estimated effect of “missing” for percent real juice is nearly zero, implying that when absent, it doesn’t affect choices. Also, when absent, the relative variance decreases. The relative variance increases when attributes are present, implying choice variability increases/decreases as one “adds”/“subtracts” more attributes.





We cited literature and reviewed empirical evidence that error variability in choice experiments is not constant but varies systematically with many factors. Now we discuss two ways that one can try to capture and control this variability in models.

Can We Model Scale?

To model scale, one must specify scale as a function of observables. “Observables” are factors that can be identified and measured (preferably also forecast into the future over space, contexts, etc). For example, embedding attitudes in a model might “explain” some individual differences, but it is hard to forecast them into the future. So, including them in models requires strong assumptions about invariance over people, space, time, contexts, etc. This assumption is unlikely to hold, so this is not a particularly good idea.

Covariance Heterogeneity Models (CHMs) are growing in popularity in applied economics and marketing; examples include Swait and Adamowicz (1997; 2001a,b); Hensher, Louviere and Swait (1999); DeShazo and Fermo (2002) and Delleart, Brazell and Louviere (1999). Attractive CHM properties include: a) simple and interpretable, b) closed form, c) captures attribute interactions, even if not specified in mean components, and d) captures nonlinear effects even if main effects are linear. CHMs mimic random coefficient models if one models error variability and conditional response means jointly as a function of attributes varied in experiments. We can express this CHM as follows:

$$P(i|C) = \{\exp[\exp(\alpha_0 + \alpha_1 X_i)(\beta_0 + \beta_1 X_i)]\} \,/\, \Sigma_{j \in C}\ \{\exp[\exp(\alpha_0 + \alpha_1 X_j)(\beta_0 + \beta_1 X_j)]\}, \qquad (5)$$

where α_0 is an intercept for the scale function, constrained to be positive (scale must be > 0); α_1 is a vector of parameters associated with a design matrix of attribute effects, X_i and X_j; β_0 is an intercept for the mean (systematic) function; $\beta_1 X_i$) is a vector of parameters associated with a design matrix of attribute effects, X_i and X_j.

We used CHMs to analyze 44 experiments designed by the UTS Centre for the Study of Choice (funded by the Australian Research Council). Experiments combined attributes (4, 8, 12, 16), numbers of attribute differences (nested under attributes), and relative design efficiency (37% to 100%). Subjects were recruited from an opt-in panel, with 100 randomly assigned to each experiment. The mean component was specified as a function of the effects-coded factors; the scale component as a function of logarithms of numbers of attributes and design efficiency. Estimation results are in Table 6.

Table 6: Results for CHMs Estimated from Pizza and Island Holiday Choices

Island Holiday Choices				Delivered Pizza Choices			
Effects	Estimate	T-Stat	P(T)	Effects	Estimate	T-Stat	P(T)
Price	-0.082	-9.890	0.000	Type of Pizza	-0.002	-0.330	0.740
Destination type	0.035	7.790	0.000	Price	-0.104	-7.180	0.000
Airline	-0.008	-2.500	0.013	Quality	-0.177	-7.380	0.000
Length of stay	0.126	10.590	0.000	Del time	-0.036	-6.040	0.000
Meals included	-0.125	-10.540	0.000	Crust	0.032	5.080	0.000
Tours available	-0.048	-8.800	0.000	Sizes	-0.053	-6.390	0.000
Season	0.000	0.110	0.909	Temp	-0.144	-7.410	0.000
Hotel type	-0.199	-10.790	0.000	Open hours	-0.023	-4.450	0.000
Length of Trip	-0.011	-2.890	0.004	Del Charge	-0.093	-6.870	0.000
Cultural activities	-0.037	-7.210	0.000	Store Type	0.064	6.250	0.000
Dist to attractions	-0.049	-7.920	0.000	Baking Method	-0.040	-4.830	0.000
Swimming Pool	-0.052	-8.140	0.000	Manners	-0.011	-1.620	0.105
Helpfulness	-0.026	-4.300	0.000	Vegetarian option	-0.064	-5.150	0.000
Type of Holiday	-0.045	-6.670	0.000	Delivery Guarantee	-0.056	-4.980	0.000
Beach availability	-0.068	-7.990	0.000	Distance	-0.041	-3.820	0.000
Brand	-0.001	-0.090	0.927	Variety	-0.039	-3.420	0.001
Intercept (8 atts)	0.002	0.230	0.815	Intercept (8 atts)	0.034	2.740	0.006
Intercept (12 atts)	-0.020	-1.950	0.051	Intercept (12 atts)	-0.014	-0.990	0.324
Intercept (16 atts)	0.029	3.290	0.001	Intercept (16 atts)	-0.040	-2.770	0.006
Scale Intercept	3.377	22.490	0.000	Scale_Intercept	4.437	23.600	0.000
Log (No. of Attrib.)	-0.571	-14.630	0.000	Log (No. of Attrib.)	-0.908	-23.390	0.000
Log (Efficiency/10)	-0.715	-20.230	0.000	Log (Efficiency/10)	-0.878	-18.250	0.000

The key takeaway in Table 6 is scale varies systematically with the number of attributes and design efficiency. Prior to estimating the model in Table 6, we graphed estimated scale against attributes and efficiency; this suggested both were approximately logarithmically related to scale. Efficient designs maximize attribute differences; so the results imply that more efficient designs that vary more attributes will increase choice variability at a decreasing rate, all else equal. The example shows that CHMs can capture systematic relationships between unobserved variability and factors varying within and between experiments, across individuals, contexts, etc. Now we discuss another way to capture unobserved variability, namely estimating choice models for single individuals.

Can We Estimate Choice Models For Individuals?

The Australian Research Council funded a team in the Centre for the Study of Choice at UTS to develop theory and design methods to estimate choice models for single people (Louviere, Marley, Street and Burgess 2004). We now briefly describe some empirical work that focused on potential constraints on sizes of problems that can be handled with this approach. Specifically, we designed 66 experiments to vary a) 11 combinations of two- and four-level attributes ranging from a 2^3 x 4^3 to a 4^8 x 2^4, b) number of options per choice set (3, 4 or 5) and c) category (delivered pizza or cross country flights). Subjects were web panelists recruited from an Australian opt-in panel; approximately 20 subjects were randomly assigned to each of the 66 conditions.

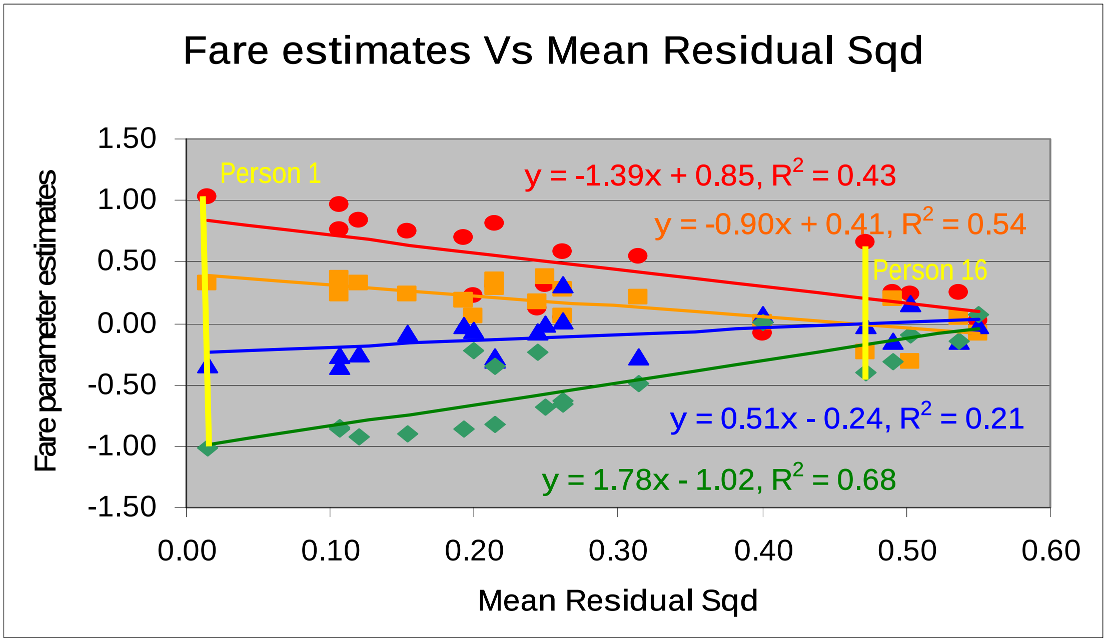
Below we report the results of one condition with 32 choice scenarios (sets), and four options described by 10 attributes (3^3 x 2^7). To date we have analyzed the smallest and largest of the 66 conditions, and have been able to estimate models for all individuals in these conditions. This one condition is one of the larger conditions, but otherwise is not unique in any way; the model estimation results are summarized in Table 8, which can be interpreted like the results of a random coefficient model. Table 8 contains the summary statistics for effects-coded MNL models for the individuals in the sample. One does not need to make assumptions about distributions of preference parameters in this case because (by definition) one has the empirical distribution for this sample population. A key takeaway from Table 8 is that standard errors for numerical attributes like fare are not constant; they are systematically related to the attribute levels.

Table 8: Summary Statistics for Individual Models in One Condition

Effect	N	Mean	StdErr	StdDev	T-Stat
ASC1	20	0.0205	0.0511	0.2286	0.4005
ASC2	20	0.0409	0.0530	0.2369	0.7718
ASC3	20	-0.0134	0.0500	0.2237	-0.2675
Flying time1	20	0.1507	0.0562	0.2511	2.6845
Flying time2	20	0.0306	0.0236	0.1057	1.2925
Flying time3	20	0.0115	0.0187	0.0834	0.6173
Fare1	20	0.4650	0.0750	0.3352	6.2037
Fare2	20	0.1550	0.0434	0.1943	3.5678
Fare3	20	-0.1013	0.0392	0.1753	-2.5842
Checkin	20	0.0226	0.0179	0.0801	1.2597
Airline1	20	0.1263	0.0450	0.2010	2.8103
Airline2	20	-0.0289	0.0335	0.1500	-0.8633
Airline3	20	-0.0175	0.0399	0.1786	-0.4395
Meals	20	-0.0423	0.0127	0.0570	-3.3222
Entertainment	20	-0.0151	0.0097	0.0434	-1.5535
Wait time for Bags	20	0.0678	0.0293	0.1312	2.3119
Frq Flyer rewards	20	-0.0059	0.0165	0.0737	-0.3584
Number of Stops	20	0.0153	0.0092	0.0410	1.6686
%OnTime departures	20	0.0442	0.0185	0.0826	2.3945
Free Alcohol	20	0.0138	0.0145	0.0648	0.9511
Free Drinks	20	-0.0629	0.0174	0.0776	-3.6273

Experimental subjects were asked to choose most and least preferred options in each scenario; questions were repeated to order options, which also yields frequencies of choices in each scenario. This allows calculation of individual error variances (sums of squared residuals). Figure 6 is a graph of mean squared residual sums (MRSs) versus four fare level estimates across individuals. Random utility theory predicts four straight lines converging to zero as the MRS increase. Each person is represented by four points on the graph; vertical lines on left- and right-hand sides indicate two subjects.

Figure 6: Mean Square Residuals Versus Fare Parameter Estimates

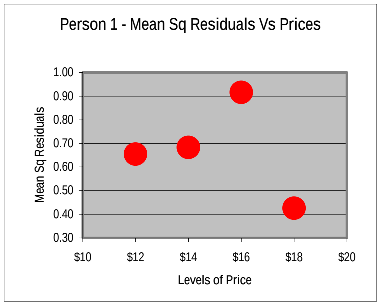


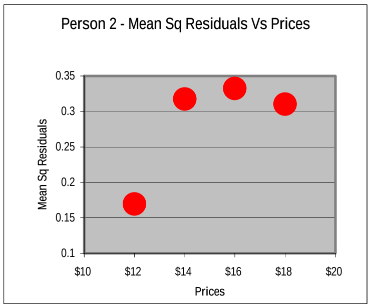
A key takeaway in Figure 6 is that variation in estimated fare parameters is due largely to MRS differences. That is, Figure 6 implies that a “random scale” model (a distribution of individual scales) should fit as well as (if not better than) a fare preference distribution model. We can test this hypothesis by interacting the effects-coded design columns associated with each four level attribute with the individual MRS values. If the MRS values contribute nothing to the model fit, all interactions should be non-significant.

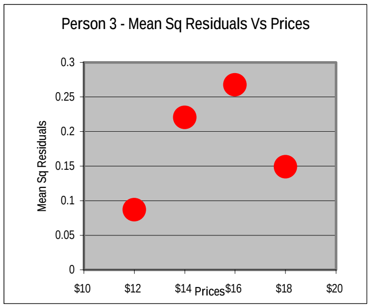
Table 8: Testing MRS by Attribute Level Effects

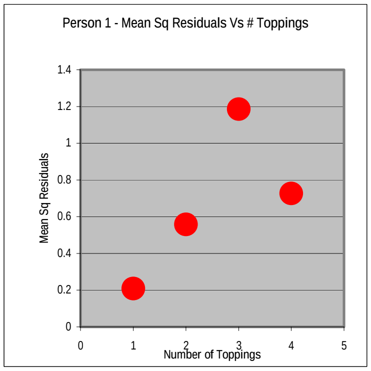
Effect	Estimate	StdErr	T-Stat	P(T)
asc1	-0.212	0.050	-4.256	0.000
asc2	-0.151	0.049	-3.061	0.002
asc3	-0.127	0.050	-2.529	0.011
asc1 x MRS	0.912	0.138	6.612	0.000
asc2 x MRS	0.753	0.138	5.462	0.000
asc3 x MRS	0.447	0.145	3.082	0.002
Flying time1	0.286	0.037	7.805	0.000
Flying time2	0.032	0.039	0.838	0.402
Flying time3	-0.062	0.039	-1.609	0.108
FT1 x MRS	-0.475	0.117	-4.062	0.000
FT2 x MRS	-0.091	0.120	-0.760	0.447
FT3 x MRS	0.169	0.120	1.411	0.158
Fare1	0.768	0.034	22.490	0.000
Fare2	0.319	0.039	8.166	0.000
Fare3	-0.198	0.044	-4.480	0.000
Fare1 x MRS	-1.247	0.108	-11.576	0.000
Fare2 x MRS	-0.646	0.123	-5.267	0.000
Fare3 x MRS	0.373	0.132	2.831	0.005
Airline1	0.058	0.039	1.479	0.139
Airline2	-0.059	0.038	-1.545	0.122
Airline3	0.020	0.011	1.861	0.063
Aline1 x MRS	0.330	0.116	2.846	0.004
Aline2 x MRS	-0.293	0.124	-2.365	0.018
Aline3 x MRS	0.178	0.119	1.494	0.135
Meals	-0.032	0.011	-3.011	0.003
Entertain	-0.015	0.011	-1.364	0.173
GetBags	0.054	0.011	4.948	0.000
FrqFlyer	-0.006	0.011	-0.551	0.582
#Stops	0.012	0.011	1.042	0.298
%OnTime	0.036	0.011	3.325	0.001
Alcohol	0.009	0.011	0.791	0.429
Drinks	-0.058	0.011	-5.311	0.000
Checkin	0.029	0.038	0.767	0.443

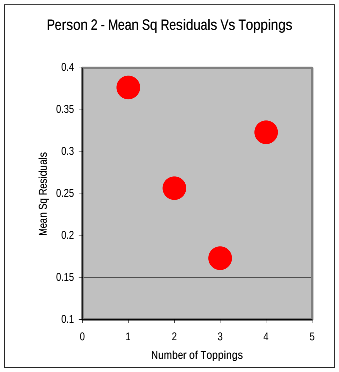
MRS x attribute level interactions are bolded in Table 8. Many interactions are highly significant; hence, individual-level error variance differences significantly impact model results. The results also suggest that individuals differ in levels of error variability; hence, constant error variability assumptions are wrong; the results also are not unique to this experimental condition. Constant variances within individuals also are unlikely. The graphs below show that individual-level MRSs values are systematically related to levels of prices and pizza toppings for the first three individuals in the Table 8 dataset. Again, there is nothing unusual about these three individuals; we chose only three to save space. The graphs suggest that error variances are systematically related to the attribute levels for each person. Thus, error variances are not constant within individuals, either.

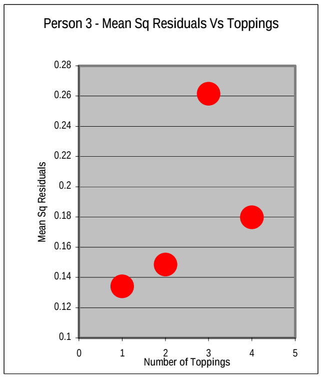

Person 1 - Mean Sq Residuals Vs Prices


Person 2 - Mean Sq Residuals Vs Prices


Person 3 - Mean Sq Residuals Vs Prices


Person 1 - Mean Sq Residuals Vs # Toppings

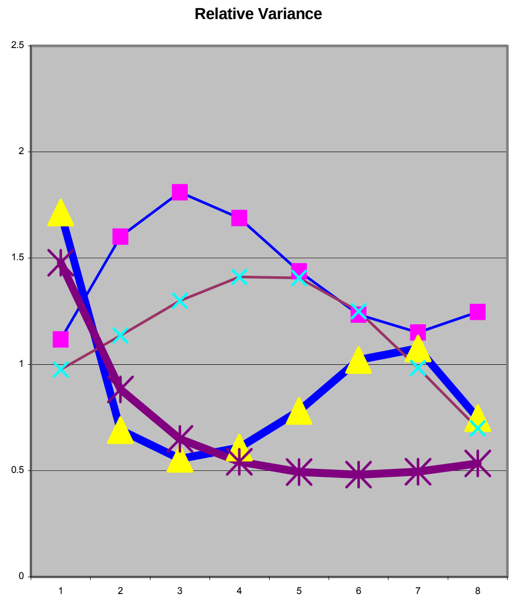

Person 2 - Mean Sq Residuals Vs Toppings

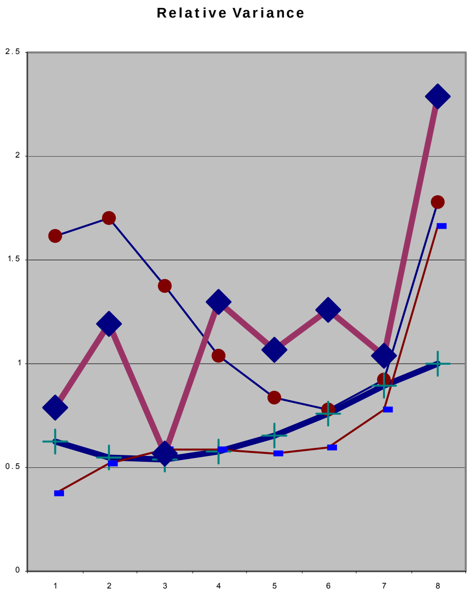

Person 3 - Mean Sq Residuals Vs Toppings

DOES ERROR VARIABILITY DIFFER BY SCENARIO ORDER?

A team from the UTS Centre for the Study of Choice helped design a UK experiment to test several hypotheses about welfare measures and choice experiments (led by Ian Bateman, currently editor of EARE). We tested order effects on choice variability by designing a target 2^3 factorial that described changes in water quality and costs relative to a status quo (other designed scenarios are not germane to the test). The 8 targets were shown as the first or last eight scenarios; we also varied whether subjects saw/did not see a glossary with all attributes and levels before the scenarios and whether a first (non-target) scenario described a large or small change in quality. We used a latin square to control for order, which created 8 more conditions that allow us to estimate models for each order condition. Individuals were randomly assigned to a particular order condition x first/last condition x glossary x quality change.

We estimated CHMs from the data for the glossary x quality change conditions for the first and second set of eight scenarios, allowing separate scale parameters for each order. Below we graph these results for each condition. Scenario order is on the X-axis; points on the graph are estimates of scale for each of the four conditions. The left-hand side graph represents the condition where the target scenarios appeared as the first 8; the right-hand graph represents the condition where the target appeared as the second eight. These graphs show that variance systematically varies across scenario orders.





Discussion and Conclusions

All latent choice models confound scale and parameter estimates. The confound poses particular problems in complex models like random coefficients models and latent class models. If one merely wants to predict choice probabilities from choice models estimated from experiments (seemingly the majority of applications), the models will predict well but will be biased and incorrect. That is, it is likely that a) many random coefficient models are over-fit, b) error distribution assumptions are not satisfied, and c) as Train and Weeks (2005) note, some distribution combinations make little theoretical sense.

We began by noting the scientists care about assumptions that underlie models, and this paper was about such concerns. That is, we were concerned about assuming that errors are iid, a widespread assumption in random coefficient models. We noted that model parameter estimates are "scaled" by standard deviations of error distribution, and unless this standard deviation is constant for all observations, distributions of model parameter estimates will be confounded with distributions of error variances. Evidence was provided to show that it is very unlikely that errors are constant; instead, they are likely to be systematically related to different factors that we noted in the paper.

So, the bottom line is that one cannot estimate individual-level parameters from complex choice models unless one can separate scale and model parameter estimates. We discussed two potential ways to do this: using various forms of covariance heterogeneity models and developing ways to estimate models for single persons. We illustrated both with empirical results that reveal systematic relationships between the attribute levels manipulated in choice experiments and error variability. Thus, it is highly likely that random coefficient models are biased and misleading. Moreover, it is well-known that virtually all choice models fit to choice experiment data will fit the data well (See, e.g., Dawes and Corrigan 1974), and so good fits and predictions to sets of hold-out choices are largely useless to test the validity of choice models estimated from experiments.

The field needs research that will lead to new models that can capture both scale and systematic component (mean) effects. The field also would benefit from research that leads to better and more useful behavioral theory. What the field definitely does not need is more complex statistical models, and it would be beneficial for academics in marketing to admit that there are serious issues associated with these classes of models, and that just because one can formulate and estimate complex statistical models does not mean that one should in fact do this. So, it is now time for marketing researchers to acknowledge these issues and to move on to more promising and less obviously empirically incorrect methods and models.

References Cited

Ben-Akiva, M. and Lerman, S.R. (1985) Discrete Choice Analysis. Cambridge: MIT Press.

Cardell, N.S. (1997) “Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity,” Econometric Theory, 13, 185-213.

Dellaert, B., Brazell, J. and Louviere, J.J. (1999) ”The Effect of Attribute Variation on Consumer Choice Consistency,” Marketing Letters 10(2), 139-147.

DeShazo, J.R. and G. Fermo (2002) “Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency,” Journal of Environmental Economics and Management, 44(1), 123-143.

Hensher, D., Louviere, J.J., Swait, J. (1999) “Combining Sources of Preference Data,” Journal of Econometrics, 89, 197-221.

Louviere, J.J. (2001) “What if Consumer Experiments Impact Variances as Well as Means? Response Variability as a Behavioral Phenomenon,” Journal of Consumer Research, 28, 506-511.

Louviere, J.J. (2004) "Complex Statistical Choice Models: Are the Assumptions True, and If Not, What Are the Consequences?" CenSoC Working Paper No. 04-002, Centre for the Study of Choice, University of Technology, Sydney, http://www.business.uts.edu.au/censoc/papers/index.html

Louviere, J.J. and Islam, T. (2004) “To Include or Exclude Attributes in Choice Experiments: A Systematic Investigation of the Empirical Consequences,” Wiley, J & Thirkell, P. (eds.), Proceedings of the Australian and New Zealand Marketing Academy Conference, Victoria University of Wellington, Wellington, New Zealand, 2004.

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

Louviere, J.J., Burgess, L., Street, D. and A.A.J. Marley (2004), "Modeling the Choice of Single Individuals By Combining Efficient Choice Experiment Designs with Extra Preference Information," CenSoC Working Paper No. 04-005, , Centre for the Study of Choice, University of Technology, Sydney, http://www.business.uts.edu.au/censoc/papers/index.html

Louviere, J.J., Street, D., Carson, R., Ainslie, A., DeShazo, J.R., Cameron, T., Hensher, D., Kohn, R. and A.A.T. Marley (2002) "Dissecting the Random Component of Utility," Marketing Letters, 13, 3, 177-193.

Louviere, J.J., Meyer, R.J., Bunch, D.S., Carson, R.T., Dellaert, B., Hanemann, M., Hensher, D.A. and J. Irwin (1999) "Combining Sources of Preference Data for Modelling Complex Decision Processes," Marketing Letters, 10, 3, 187-204.

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T.A., Carson, R.T., DeShazo, J.R., Fiebig, D., Greene, W., Hensher, D. and Waldman D. (2006) "Recent Progress on Endogeneity in Choice Modelling," Marketing Letters, 16, 3-4.

McFadden, D. (1974) "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), Frontiers in Econometrics, 105-142, Academic Press: New York.

McFadden, D. and K. Train (2000). "Mixed MNL Models for Discrete Response," Journal of Applied Econometrics, 15, 447-470.

Ohler, T., Le, A., Louviere, J.J. and J. Swait (2000) "Attribute Range Effects in Binary Response Tasks," Marketing Letters, 11, 3 (August), 249-260.

Swait, J. and Adamowicz, W. (2001a) "Choice Complexity and Decision Strategy Selection," Journal of Consumer Research, 28, 135-148.

Swait, J. and Adamowicz, W. (2001b) "Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice," Organizational Behavior and Human Decision Processes, 86, 2, 141-167.

Swait, J. and Louviere, J.J. (1993) "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," Journal of Marketing Research, 30, 305-314.

Train, K. and Weeks, M. (2005) "Discrete Choice Models in Preference Space and Willingness-to-Pay Space," in A. Alberini and R. Scarpa, (Eds.) Applications of Simulation Methods in Environmental Resource Economics, Springer Publishers: Dordrecht, The Netherlands, Chapter 1, pp. 1-17,

Wedel, M. and Kamakura, W. (1999) Market Segmentation: Conceptual and Methodological Foundations, Dordrecht: Kluwer Academic Publishers.

Wise, C., Louviere, J.J. (2004) "The Impact of Varying Amounts of More and Less Salient Product Information Upon Consumer Willingness-to-Pay." In Wiley, J. & Thirkell, P. (eds.), Proceedings of the Australian and New Zealand Marketing Academy Conference, Victoria University of Wellington, Wellington, New Zealand.

ESTIMATING ATTRIBUTE LEVEL UTILITIES FROM "DESIGN YOUR OWN PRODUCT" DATA—CHAPTER 3

JENNIFER RICE AND DAVID G. BAKKEN
HARRIS INTERACTIVE

ABSTRACT

We describe a method for estimating utilities from single exposure *design your own product* data using Sawtooth Software's CBC/HB program. Results are compared to choice-based conjoint results using the same attributes and levels and choice data gathered from the same individuals.

BACKGROUND AND INTRODUCTION

The variety of methods known collectively as "conjoint analysis"—full profile (e.g., Sawtooth Software's CVA), adaptive (ACA), and choice-based (CBC)—are the most popular marketing research techniques for understanding the decision processes that buyers employ when evaluating competitive alternatives. These methods have three elements in common. First, they are derived from a theory of *random utility maximization*, which holds that buyers seek to maximize a utility function such that they will select the course of action that offers the greatest *expected* utility among the different actions available at the time of the decision. Second, these methods attempt to *simulate* the decisions that buyers make, usually by presenting simulated product or service offerings to a representative sample from the target market and asking survey respondents to evaluate these simulated offers. Evaluations may take the form of paired comparisons (typical of ACA), preference ratings, ranking, or choice from a set of offers. Third, these are *repeated measures* methods, with multiple simulated decisions for each respondent.

Utility maximization has proved to be a robust framework for understanding consumer evaluations of *multi-attributed* alternatives, which describes the majority of offers found in a variety of product and service markets. One reason for this robustness lies in the fact that the probability of choosing an alternative under utility maximization can be expressed mathematically in terms of the ratios of the utilities of the alternatives. This in turn leads to statistical methods that can be used to "decompose" the observed choices into a set of *utilities* or *part-worths* for each of the component attributes of the choice alternatives.

Among the conjoint methods, *choice-based* conjoint provides the most realistic simulation of the decision process, and has become increasingly popular due to commercially available software for the design of choice tasks and *individual-level* estimation of utilities using hierarchical Bayesian methods. As experience with this method has grown, researchers have a better understanding of the best applications for these methods and some of the limitations.

Common marketing applications for choice-based conjoint include new product (or service) design, product portfolio optimization, pricing research, and brand equity assessment. As an example, a consumer electronics marketer might wish to determine which features to include in a portable DVD player. Most often, decisions about which features to include in a product represent engineering or design trade-offs for the manufacturer. For example, including more memory in MP3 players may make them heavier, reduce battery life, add cost, and so forth. In order to determine which design trade-offs to make, marketers often turn to consumers to determine which features or capabilities consumers value most. For example, would buyers of MP3 players prefer more memory or longer battery life? The answer to this type of question helps the marketer to make the appropriate trade-off to maximize the return on product development investments, especially when the manufacturing requirements favor one or at most a few “optimal” offers.

Offering a fixed portfolio of products with different features is one way to satisfy diverse customer needs while keeping the costs of manufacturing complexity down. Another strategy is based on “mass customization”—providing a basic product or service *platform* and enabling consumers to select among optional features that can be provided at relatively low marginal cost in terms of manufacturing complexity. With disaggregate estimation of utilities, choice-based conjoint can, in theory, deal with mass customization as well as with identification of an “optimum” product portfolio. Disaggregate estimation reveals each buyer’s *ideal point* for each attribute in the model. By using a *first choice* rule for the market simulator, we can identify a set (of *j* alternatives) of offers for a portfolio that will satisfy the requirements of the maximum number of buyers. Similarly, we can identify the combinations, in a mass customization offer, which are most likely to be ordered or requested by customers by, in effect, determining each respondent’s *ideal* and then counting the number of different combinations as well as the frequencies with which each of the optional components would be chosen.

If the number of optional features is large, implementing a choice-based conjoint exercise for a mass-customization application may be difficult. Alternatives such as adaptive conjoint or partial-profile choice-based designs might be applied in such cases.

The optional features offered via mass customization usually add some incremental amount to the total product price. For example, a consumer might choose to substitute a 17” LCD monitor for the 19” CRT monitor in the desktop computer package she is buying, at an incremental cost of $250. The computer system marketer would like to know the level of price elasticity for the monitor upgrade. Perhaps many customers are willing to pay more than $250 for this upgrade, while others may be willing to pay only $150. The ideal way to determine this “willingness to pay” might be to incorporate a price variable for the optional feature into a choice-based conjoint design. For example, when the 17” LCD monitor is included, the system price is increased by $150, $200, or $250[1]. If there are many optional features with incremental pricing, this leads to a very large design.

A more common approach is to use sensitivity to the overall product or service price to derive willingness to pay for a selected optional feature. Survey respondents do not receive information about the incremental price of any one feature. This approach requires running market simulations where the overall price and the presence/absence (or level) of the feature are varied at the same time. For products or services where the incremental *utility* of any one optional feature is large relative to the total utility of the offer, this may provide an adequate estimate of the incremental willingness to pay for the feature. However, our experience indicates that for many products, such as automobiles and computers, the differences in incremental cost for optional features often are obscured by the differences in overall price.

[1] This could be implemented by creating a 4-level attribute for the monitor: 19” CRT; 17” LCD @ $150; 17” LCD @ $200; 17” LCD @ $250. Other implementations are possible.

Menu-based approaches to choice-based conjoint have been employed to address this shortcoming. For example, Leichty, Ramaswamy and Cohen (2001) employed a menu-based approach in which consumers could select from a list of *a la carte* features. The prices of the features were varied from one choice scenario to the next. In this case, each feature was treated as a separate binomial choice. This permitted estimation of price sensitivity for each of the features.

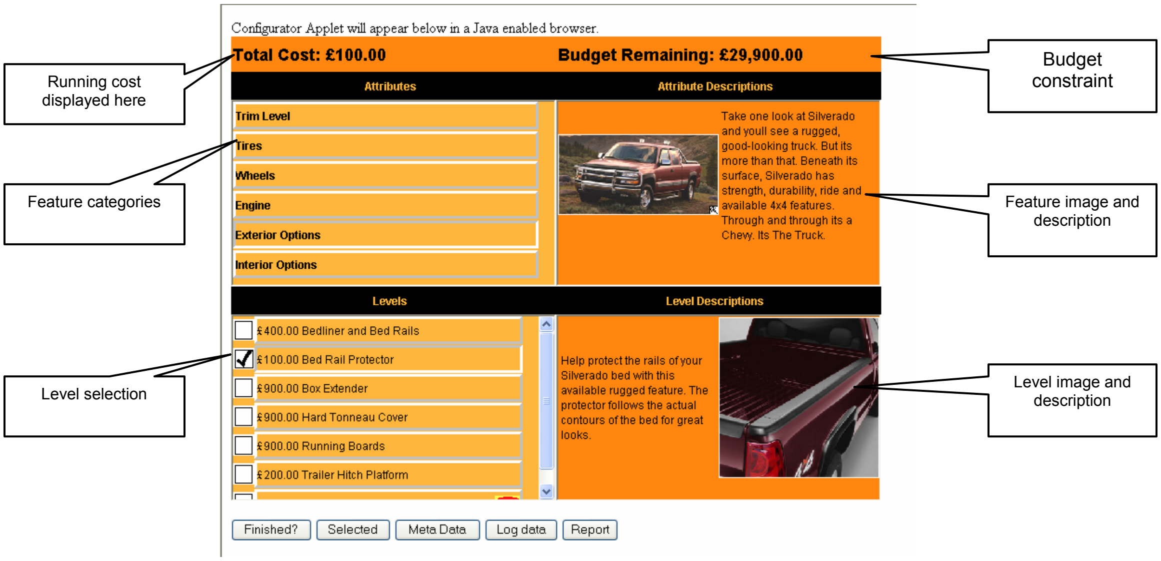
THE *DESIGN YOUR OWN PRODUCT* ALTERNATIVE TO CHOICE-BASED CONJOINT

Design your own product questions have been used in survey research for a number of years and predate computer-assisted and web-interviewing methods. However, as manufacturers have introduced web-based tools that allow customers to configure products to their own specification, interest in emulating these configuration tools for research has grown. Design your own product questions are appealing for several reasons. In most cases, they require less interview time and present respondents with a series of relatively simple choices. Choice-based conjoint tasks, in comparison, often require respondents to evaluate several alternatives across six or more dimensions at a time.

Chapters 1 and 2

We previously have described our user interface for the design your own product (DYOP) task (Bakken & Bayer, 2001; Bakken & Bremer, 2003). Figure 1 presents an example of a DYOP task. In our first application ("Chapter 1"), we specifically looked at the difference in predicted selection rates for features between choice-based conjoint and the respondent's self-configured product. We found that the choice-based conjoint model appeared to over-estimate demand for higher priced options. For example, in an automotive study where one of the options included a navigation system available in two different forms (printed instructions versus a detailed map), the choice model predicted equal adoption rates for the two different navigation modes. In other words, customers who wanted a navigation system appeared to be indifferent to the type of system. In the design-your-own-vehicle exercise, when the price difference between the two modes was revealed, nearly twice as many respondents chose the less expensive system. We concluded from this study that design-your-own-product exercises are a useful adjunct to choice-based conjoint, especially when we want to learn something about sensitivity to optional feature pricing. We also observed that, in order to support conjoint-like simulations, it would be necessary under aggregate estimation to show a variety of prices for each feature level to different respondents and, for disaggregate estimation, to have multiple DYOP tasks for each respondent with prices varying across the tasks.

Figure 1.
Design Your Own Product "Configurator" Image



Configurator Applet will appear below in a Java enabled browser.

Total Cost: £100.00 Budget Remaining: £29,900.00

Trim Level
Tires
Wheels
Engine
Exterior Options
Interior Options

Take one look at Silverado and youll see a rugged, good-looking truck. But its more than that. Beneath its surface, Silverado has strength, durability, ride and available 4x4 features. Through and through its a Chevy. Its The Truck.

☐ £400.00 Bedliner and Bed Rails
☑ £100.00 Bed Rail Protector
☐ £900.00 Box Extender
☐ £900.00 Hard Tonneau Cover
☐ £900.00 Running Boards
☐ £200.00 Trailer Hitch Platform

Help protect the rails of your Silverado bed with this available rugged feature. The protector follows the actual contours of the bed for great looks.

Because DYOP or "build your own" exercises have many apparent advantages, for Chapter 2 we attempted to estimate conjoint-like utilities from this type of data (Bakken & Bremer, 2003). We were interested in estimating price sensitivity for the features, rather than just observing differences in selection rates between higher and lower-priced options. We faced two key challenges: 1) we had only one observation per feature for each respondent, and 2) the incremental price for any given feature was the same across all respondents. We hypothesized a set of conditional relationships between utility and price that allowed us to specify a model where the probability of choosing a specific feature was a function of the *intrinsic* value of the feature, the value of the competing features, the price of the feature, and the prices of the competing features. We further assumed that the utility obtained by spending a dollar on one feature is equal to the utility obtained from that same dollar spent on any other feature. As a result, we used the price of the feature relative to the total price of the configured product as the price variable. This served to introduce variability in the price of any feature *across* respondents. In this preliminary effort, simulation results using the DYOP utility estimates were similar (in order and relative magnitude) to results using a choice-based conjoint model estimated from data obtained from the same respondents for the same products and feature sets.

Chapter 3

We consider "Chapter 2" as a "proof of concept." We determined that it was possible to estimate utilities from the limited data obtained with a design-your-own-product exercise. However, we were not entirely satisfied with the effort. We set two objectives for "Chapter 3"—better evaluation of the DYOP results vis-à-vis choice-based conjoint, and making the process more "user-friendly" by employing a commercially available software solution (in this case, CBC/HB from Sawtooth Software).

The hierarchical Bayesian method of estimation utilized by CBC/HB assumes that the individual decision process is described by multinomial logit, while the heterogeneity in individual preferences or utility is distributed multi-variate normal. In order to apply HB estimation to our design-your-own-product data, we hypothesized a conditional decision process such that we could treat the choices across different feature groups as repeated observations from a given individual. The basic model specification is as follows:

$$\text{Prob}(j) = f(U_j + \text{RelPrice}_j + \text{Appeal}_j + \varepsilon)$$

Where:

J = specific feature

U_j = *intrinsic* utility of j (unobserved)

RelPrice_j = covariate of j reflecting incremental price, expressed relative to the total price of the product as configured

Appeal_j = stated appeal for feature J, measured outside the design your own exercise

ε = all other unobserved components

Our estimation objective is to isolate $\mathbf{U_j}$ from the other components driving the probability of choice. As previously noted, we hypothesize a conditional choice process such that, if the chosen feature also has the highest appeal (as measured in a rating exercise), that becomes a sufficient reason for choosing the feature. However, if a less appealing feature is chosen, then the price of that feature relative to alternatives must enter into the decision process. By testing for these conditions using Bayesian inference, we should, in theory, be able to determine the probability that a choice is based on appeal or on price. In effect, we look at the appeal rating and the relative price and use the conditional hypotheses to infer the intrinsic utility of the feature.

In order to make CBC/HB work in this case, we treat each feature decision as a separate choice task where the same conditional hypotheses apply to all feature choices. We should note that this is only one way of organizing the choice data. We could frame the DYOP as one huge choice task, where every possible combination represents a separate alternative. This approach was first proposed by Ben-Akiva and Gershenfield (1998) and also tested, with simulated choice data, by Johnson, Orme and Pinnell (2006). We might also take a menu-based approach, in which selection of each feature is treated as a binomial choice. This is similar to the estimation method introduced by Leichty, Ramaswamy and Cohen (2001). Johnson, *et al.* also analyzed the simulated choice data by treating each attribute as a separate choice task and found that the estimated utilities were almost identical to those obtained using the one big choice task method.

The Case Study

The data for this analysis were obtained from a study of automotive preferences. Approximately 500 consumers who were either current owners of a specific vehicle “segment” or “intenders” for that segment were recruited from the Harris Interactive Online Panel (HPOL) and completed a survey that included a self-explicated measure of feature appeal, a choice-based conjoint exercise, and a “Configurator” design your own product exercise. There were 64 different feature possibilities, making for 106 billion possible “orderable” combinations in the DYOP exercise. For the choice-based conjoint tasks, some of these features were grouped into packages, such as “luxury options package.”

We should note that we incorporated feature-specific pricing into the choice-based conjoint exercise. We included an overall price attribute which determined the *base-level* price (the price before options); this varied randomly across four levels (intervals between levels were in the range of $3,000 to $5,000, depending on the make of the vehicle). Each *extra cost* feature was assigned a single price value. Once the attribute levels for each concept in a choice task was set by the design, we added the price of each extra cost optional feature to the base price level for that concept; that price was presented to the respondent as the total "MSRP" for that concept in the choice task. This resulted in a price "variable" comprised of a randomly determined base price and a feature price component that was determined by the (randomly selected) feature levels. Thus, even though each optional feature level had only one possible price, the total price was a random combination of these prices plus the randomly determined base price. Respondents had no information about the individual features prices.

We employed a partial profile design for the choice-based conjoint experiment, with 11 (of 15) attributes appearing in each scenario.

Model Estimation and Analysis

As we noted previously, a primary objective of this analysis was the estimation of price elasticity for the optional features. We also wanted to make as direct a comparison as possible between the DYOP and choice-based conjoint models. For the DYOP data, we estimated a single model with linear price terms for each optional feature (as well as for the base price). We compare the results to three different models for the CBC data: 1) base price estimated as a part-worth, with feature price effects reflected (if at all) in the feature-level part-worth estimates; 2) linear price estimation for the *total price* (which includes the base price and the sum of the feature prices); and 3) linear price terms for each feature and the base price.

For the DYOP model, feature price levels were calculated based on the *relative price* of the selected feature level to the total price of the vehicle as configured by the respondent. Our intent was to introduce some variability into the price levels, in effect simulating the situation where different respondents see different prices for each feature.

The CBC exercise included 20 training and 5 holdout tasks which were interspersed throughout the exercise. Aggregate holdout prediction is summarized in Table 1. Mean absolute deviation is similar across all three CBC models. The first and last holdout tasks were identical, and test/retest reliability for these two tasks is 63.3%. We should point out that, like the training tasks, each holdout task included only 11 of the fifteen attributes, and the subset was different for each of the non-replicate holdout tasks.

Table 1.
Choice-Based Conjoint Model Validation

	Mean Absolute Deviation				
	Holdout 1 (T7)	Holdout 2 (T14)	Holdout 3 (T21)	Holdout 4 (T24)	Holdout 5 (T25)
Part Worth Price Term	7.31%	2.35%	2.69%	2.59%	3.12%
Linear Price Term for Total Price	6.30%	1.97%	2.64%	2.72%	3.53%
Linear Terms for Relative Feature Prices	7.01%	1.00%	3.20%	1.87%	2.82%

We evaluated hit rates for the DYOP model, which are summarized in Table 2 (actual features have been concealed). The very high hit rates are to be expected, given that we had only one observed choice per feature for each respondent.

Table 2.
Design Your Own Hit Rate

% Correct Prediction of Individual Feature Level Choices				
99.6%	BODY TYPE (SEDAN/COUPE/CONVERTIBLE)		100.0%	TRACTION CONTROL
99.6%	MAKE		99.6%	LEATHER TRIMMED STEERING WHEEL
100.0%	ENGINE		99.6%	ADJUSTABLE PEDALS
99.0%	DRIVE TYPE		100.0%	XENON HEADLAMPS
100.0%	ELECTRONIC SUSPENSION CONTROL		100.0%	ADAPTIVE HEADLAMPS
100.0%	INTERIOR STORAGE		99.2%	REAR AIRBAGS
99.6%	LEATHER SEATING SURFACES		100.0%	AUTOMATIC TROUBLE REPORTING
100.0%	FOLD DOWN REAR SEAT		99.8%	ADAPTIVE CRUISE
99.6%	SUNROOF/CONVERTIBLE OPTIONS		100.0%	POWER TRUNK
99.4%	AUDIO SYSTEM		99.4%	HEATED AND COOLED FRONT SEATS
99.8%	REAR SEAT DVD PLAYER		99.2%	REFRIGERATOR
99.6%	NAVIGATION SYSTEM		98.8%	SATELLITE RADIO
99.4%	RUN FLAT TIRES		99.6%	EXTENDED WARRANTY
100.0%	REARVIEW VIDEO DISPLAY		99.7%	Average

We compared estimates of price elasticity. For two of the CBC models (part-worth and total linear price models) we predicted preference shares for selected features at different price levels. One example, for electronic suspension control, is presented in Figure 2. For the CBC model with feature-level price terms and the DYOP model, we compare the linear price coefficients. Looking at Table 3, we see that eight of ten coefficients estimated from the DYOP data are larger than the coefficients estimated from the CBC data.

Figure 2.
Price Elasticity for Electronic Suspension Control

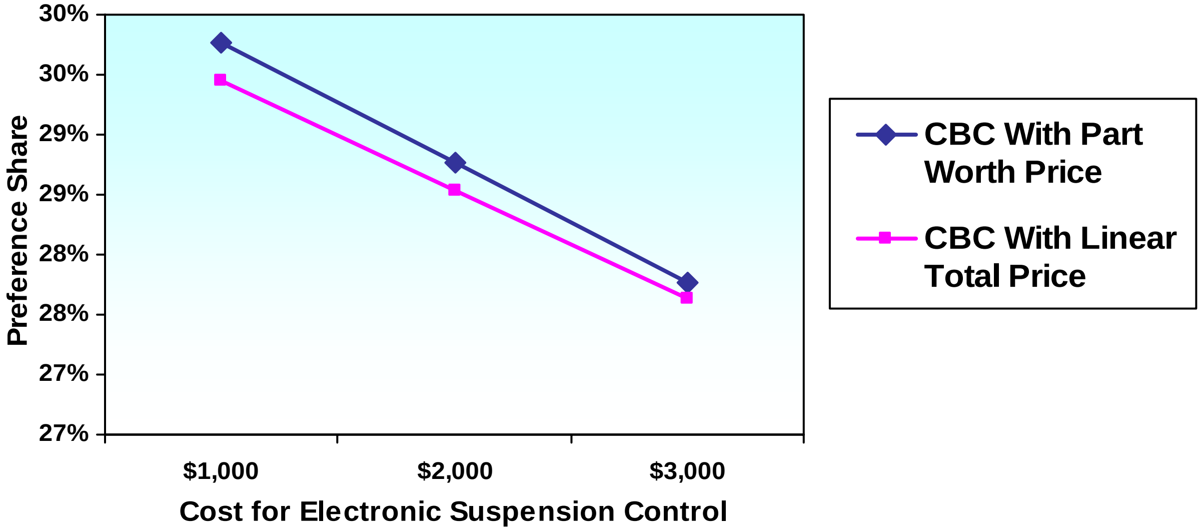


Table 3.
Comparison of Models with Relative Price Terms

Coefficients for Linear Feature Price Terms		
	CBC Model With Relative Price Terms*	Design Your Own Model
MAKE	-1.35	-8.03
ENGINE	-1.55	-2.67
DRIVE TYPE	-1.19	-1.88
SUNROOF/CONVERTIBLE OPTIONS	-1.26	-1.92
INTERIOR STORAGE	-0.64	-2.20
ELECTRONIC SUSPENSION CONTROL	-0.54	-1.09
AUDIO SYSTEM	-1.02	-1.77
REAR SEAT DVD PLAYER	-0.44	-1.39
NAVIGATION SYSTEM	-1.84	-0.46
EXTENDED WARRANTY	-1.01	-0.89

Given the differences in the price coefficients, we would expect differences in simulated preference shares between the DYOP and CBC models. Table 4 summarizes predictions for four simulated vehicles. Predictions from all three CBC models are nearly identical, but we see a big difference in the DYOP predictions for the third and fourth vehicles. Vehicle 4 includes the most expensive level of "electronic suspension control." The DYOP price coefficient for this level is twice as large as that for the multiple linear price CBC model.

Table 4.
Comparison of Simulated Preferences across Models

	Vehicle 1	Vehicle 2	Vehicle 3	Vehicle 4
CBC Model with Part Worth Prices	30.84%	11.95%	26.87%	30.34%
CBC Model with Linear Total Price	27.73%	10.74%	29.66%	31.87%
CBC Model with Linear Relative Prices	27.27%	10.61%	28.57%	33.55%
Design Your Own Model	23.61%	9.69%	44.31%	22.40%

SUMMARY

We continue to be enthusiastic about design your own product questions for understanding consumer choices among multi-attributed alternatives where some degree of customization for each respondent is allowed. At a minimum, DYOP data can inform decisions about production capacity for the different optional features.

We have demonstrated that it may be possible to estimate price sensitivities for optional features from DYOP data, even when there is only one DYOP observation per respondent and the nominal feature prices are the same for all respondents. Anecdotal evidence from clients suggests that choice-based conjoint underestimates price elasticity for individual features. We found that, for most features, the elasticities estimated from the DYOP data are larger than elasticities estimated from CBC data. Our approach might be applied to actual product configuration data obtained from manufacturers' websites or similar automated ordering mechanisms.

We have tested only one model of the data generating process for DYOP. It may be that other model specifications will produce different and better results.

Johnson, *et al.* (ibid.) have made an important contribution to our understanding of DYOP questions. In particular, they demonstrate that a simple counting process for analyzing DYOP data and approximating aggregate logit utilities may be as good as more involved estimation methods.

REFERENCES

Bakken, D. G. and L. R. Bayer (2001), "Increasing the Value of Choice-based Conjoint with 'Build Your Own' Configuration Questions." 2001 Sawtooth Software Conference Proceeedings.

Bakken, D. G. and J. Bremer (2003), "Estimation of Utilities from Design Your Own Product Data," A/R/T Forum, June, 2003, Monterey, California.

Ben-Akiva, M. and Sl Gershenfeld (1998), "Multi-Featured Products and Services: Analyzing Pricing and Bundling Strategies," Special Issue of the Journal of Forecasting, Vol. 17, Issue 3/4.

Johnson, R., B. Orme and J. Pinnell (2006), “Simulating Market Preference with ‘Build Your Own’ Data,” 2006 Sawtooth Software Conference, Delray Beach, Florida.

Leichty, J., V. Ramaswamy and S. H. Cohen (2001), “Choice Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand with an Application to a Web-based Information Service,” Journal of Marketing Research, Vol. 38, Number 2.

SIMULATING MARKET PREFERENCE WITH "BUILD YOUR OWN" DATA

RICH JOHNSON AND BRYAN ORME
SAWTOOTH SOFTWARE, INC
JON PINNELL
MARKETVISION RESEARCH

INTRODUCTION

Choice-Based Conjoint analysis (CBC) has become popular in recent years. Although CBC does a good job of identifying the specific combination of features that will be most appealing to each respondent, CBC's method of questioning is not efficient. The respondent must answer several choice tasks which do not provide precisely what he or she may be looking for. The particular combination of features desired by that respondent is inferred through a complicated method of analysis. This leads naturally to the question, "Why not ask the respondent directly what he or she would like?"

Researchers have considered this possibility throughout the history of conjoint analysis, and some have preferred more direct "self-explicated" measures of preference. The "Build Your Own" (BYO) method of preference elicitation has been used in mass customization applications where the seller needs to learn what specific product features to assemble for a specific customer. In that context it is important to provide a task that the customer considers relevant and efficient. The fact that manufacturers are successful in using BYO tasks to sell products suggests that buyers do not find them difficult or objectionable.

The possibility of using BYO tasks in more general marketing research applications has been tantalizing for several years. The first work of which we are aware was by Ben-Akiva and Gershenfeld in 1998 (*Journal of Forecasting*). Further work was reported in 2001 by Liechty, Ramaswamy & Cohen (*Journal of Marketing Research)* and in 2001 by Dahan & Hauser (*The Journal of Product Innovation Management*). The first examination of this possibility at a Sawtooth Software conference was by Bakken and Bayer, also in 2001. In a recent survey of Sawtooth Software customers (2006), 10% of respondents reported having fielded BYO tasks in the last 12 months.

As in CBC, the typical BYO questionnaire considers a product as consisting of a collection of attribute levels. However, rather than having the respondent choose among product concepts expressed as bundles of attribute levels, BYO presents the attributes one at a time. The levels of each attribute are displayed, often with specific incremental prices. Somewhere on the screen the total price is shown for the product as currently configured. The respondent is asked to choose a level for each attribute and may also back up to revise previous choices. In the end, the respondent has presumably described his or her own ideal product, after consideration of price.

The BYO interview seems appealing as a way of eliciting customer preference data, because it appears to be efficient and relevant to the respondent. It captures each respondent's ideal, given the prices. However, the prices displayed are usually the same for all respondents, so there is no information about how respondents would have reacted with different prices. Lacking such information, BYO data cannot support market simulations involving price sensitivity like those of conjoint analysis. Bakken and Bayer concluded that to support such simulations, for aggregate estimation it would be necessary to show a variety of prices to different respondents, and for individual estimation it would be necessary to have multiple tasks for each respondent with varying prices. We believe both conclusions were correct.

WAYS TO ESTIMATE PARTWORTHS WITH BYO

1. One Enormous Choice Task: Although the respondent's task in BYO questionnaires appears quite different from the task in CBC questionnaires, there is a fundamental similarity. The BYO respondent chooses one specific alternative from among the set of all possible alternatives. Suppose there were 10 product attributes, each with three levels. Then there would be 3^{10}, or 59,049 possible alternatives, from which the respondent chooses the most preferred.

The usual BYO questionnaire therefore may be regarded as consisting of one very large choice task. Given data from many respondents, one could estimate aggregate partworths for a sample of respondents, using a multinomial logit model. But there is a practical problem: the number of choice alternatives provided to the respondent can be so great that no conventional multinomial logit software could handle the data.

The first step in this investigation consisted of creating a special-purpose multinomial logit program to estimate aggregate partworths from BYO data. In Sawtooth Software's regular MNL estimation routines, the specifications of choice tasks are saved in a large file on the hard disk. For each iteration of the estimation procedure, those specifications are read into memory as they are needed. For BYO applications, such a file could be of enormous length, making that procedure infeasible. In this special-purpose program, the specifications for each task are rebuilt each time they are needed. This is not conceptually difficult because the alternatives consist of all possible combinations of attribute levels. The only data needed to be read for each respondent are the description of the alternative that was chosen.

2. A Separate Choice Task for Each Attribute: Another approach to estimating partworths from BYO tasks is to assume independence among attributes, considering the BYO questionnaire not as a single choice task, but rather consisting of several independent choice tasks, one for each attribute. The alternatives in each task would consist of the different levels of that attribute. This approach can be handled by standard multinomial programs, and is much simpler than the preceding one, although there may be doubt about whether the assumption of independence of attributes is warranted.

3. Working Directly with Aggregate Counts: An even simpler approach would be to avoid multinomial logit estimation altogether, basing the analysis on what are called "counts" data in the Sawtooth Software community. We might just observe the number of respondents choosing each level of each attribute, and attempt to produce aggregate partworths by processing those counts in some way. Consider how aggregate partworths are used to estimate shares of preference in the case where there is a one-to-one correspondence between partworths and counts: we estimate shares of preference by exponentiating the partworths and then percentaging the results. One natural way to estimate partworths from counts would be to reverse the process: take logs of the counts, and then zero-center the logs. The zero-centering is to conform to the zero-sum property of effects-coded logit coefficients.

In what follows we explore these three approaches, first with artificial data and then with real data collected for that purpose.

SIMULATION WITH ARTIFICIAL DATA

This simulation used actual partworths estimated by CBC/HB for 194 individuals from a recent study of hotel amenities.[1] After deleting attributes in that study concerned with brand and price, there were 7 attributes left that described various hotel amenities. Two attributes had 4 levels, three had 3 levels, and two had 2 levels, for a total of 21 levels. Some attribute levels were accompanied by incremental prices, as would be the case in a BYO questionnaire.

In generating artificial choice data, each respondent was assumed to choose the alternative with highest utility. For the full multinomial model, this means the alternative with highest utility, after summing partworths over all seven attributes. For the simpler logit model, as well as for the counts data, this means the respondent was assumed to choose the level of each attribute having the greatest partworth. The results of this exercise were interesting and perhaps surprising.

First, aggregate partworths estimated by methods (2) and (3) were identical to at least five decimal places. In retrospect, this result should not have been surprising. Attributes in that logit model appear independently of one another, so the model is equivalent to a set of separate logit models, one for each attribute. Those separate models are trying to fit the aggregate counts, and there are as many parameters for each attribute as there are separate counts frequencies. Since the logit coefficients are zero-centered within attributes, performing a similar operation on logs of the counts must produce identical results, as it did.

Second, average partworths from these two simpler methods were identical to those from the full multinomial logit model to at least three decimal places. We were surprised by this result. The full logit model does represent an orthogonal design, so if we were doing a *linear* optimization rather than a logit estimation, one might expect the two models to produce similar results. But with the logit model each alternative is weighted by the square root of its probability of being chosen, which might have destroyed orthogonality.

The similarity of these three results represents a considerable bonus for the researcher. Apparently, BYO data do not require complex estimation programs. At least for these data, either of the multinomial logit models produces results nearly identical to what one can get just by taking logs of the counts of individuals choosing each attribute level. This is potentially very good news for prospective users of BYO data. Although we have not yet shown that the results are valid, correct, or useful in any way, at least we know that they are not difficult to obtain.

We now turn to an experiment that was used to confirm these findings, and to compare BYO and CBC partworth estimates.

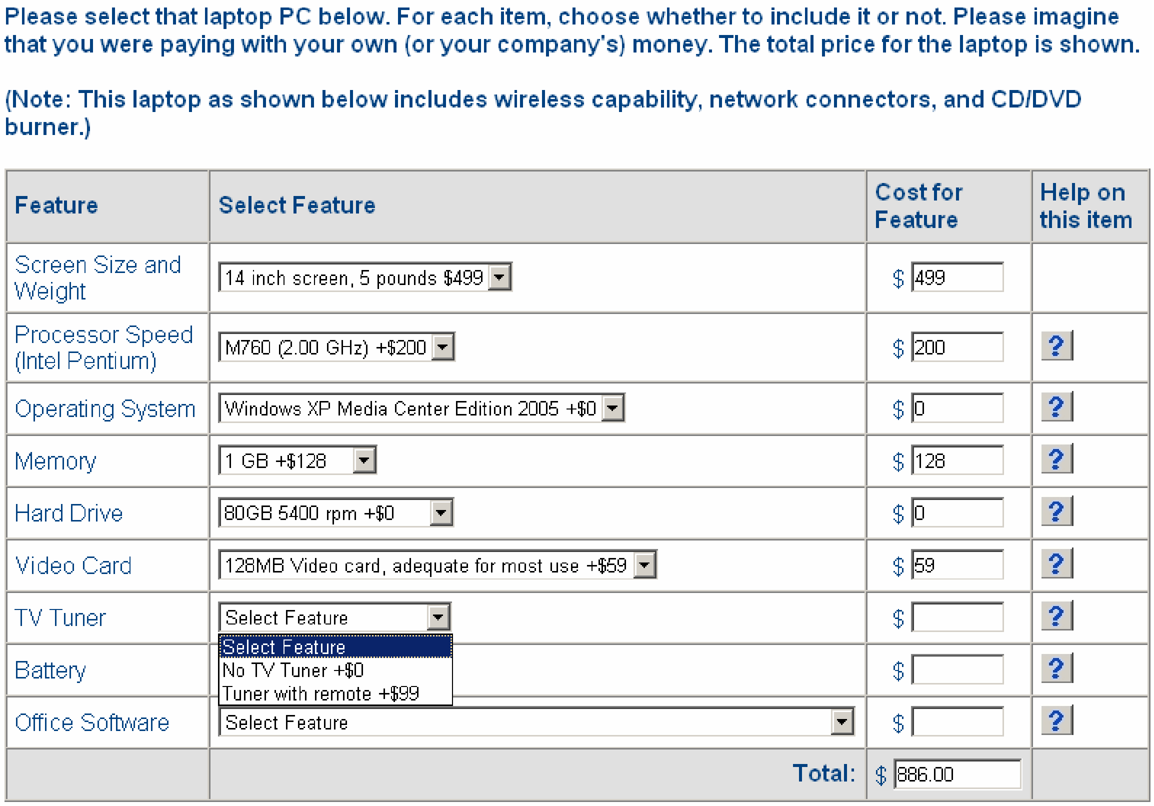
AN EXPERIMENT

In January 2006 we conducted an experiment to obtain relevant BYO and CBC data. We used a sample of 605 individuals from GMI's panel, who were interviewed on the subject of

[1] The paper, "Testing Adaptive Choice-Based Conjoint Designs" by Johnson, Orme, Huber, and Pinnell, is available at SawtoothSoftware.com.

laptop computers.[2] The computers were described using 9 attributes, each having between 2 and 4 levels. Each respondent did 1 BYO task. The layout is illustrated in the following figure:

Figure 1: Screen Format for the BYO Task



Please select that laptop PC below. For each item, choose whether to include it or not. Please imagine that you were paying with your own (or your company's) money. The total price for the laptop is shown.

(Note: This laptop as shown below includes wireless capability, network connectors, and CD/DVD burner.)

Feature	Select Feature	Cost for Feature	Help on this item
Screen Size and Weight	14 inch screen, 5 pounds $499	$ 499	
Processor Speed (Intel Pentium)	M760 (2.00 GHz) +$200	$ 200	?
Operating System	Windows XP Media Center Edition 2005 +$0	$ 0	?
Memory	1 GB +$128	$ 128	?
Hard Drive	80GB 5400 rpm +$0	$ 0	?
Video Card	128MB Video card, adequate for most use +$59	$ 59	?
TV Tuner	Select Feature	$	?
Battery		$	?
Office Software	Select Feature	$	?
	Total:	$ 886.00	

The default product had a basic level of each attribute and a base price. Upgrade levels were offered for each attribute, at ascending incremental prices. There were three sets of upgrade prices for each attribute. Each respondent saw one set of those prices randomly chosen for each attribute. As an example of a three-level attribute, Memory had these incremental price levels

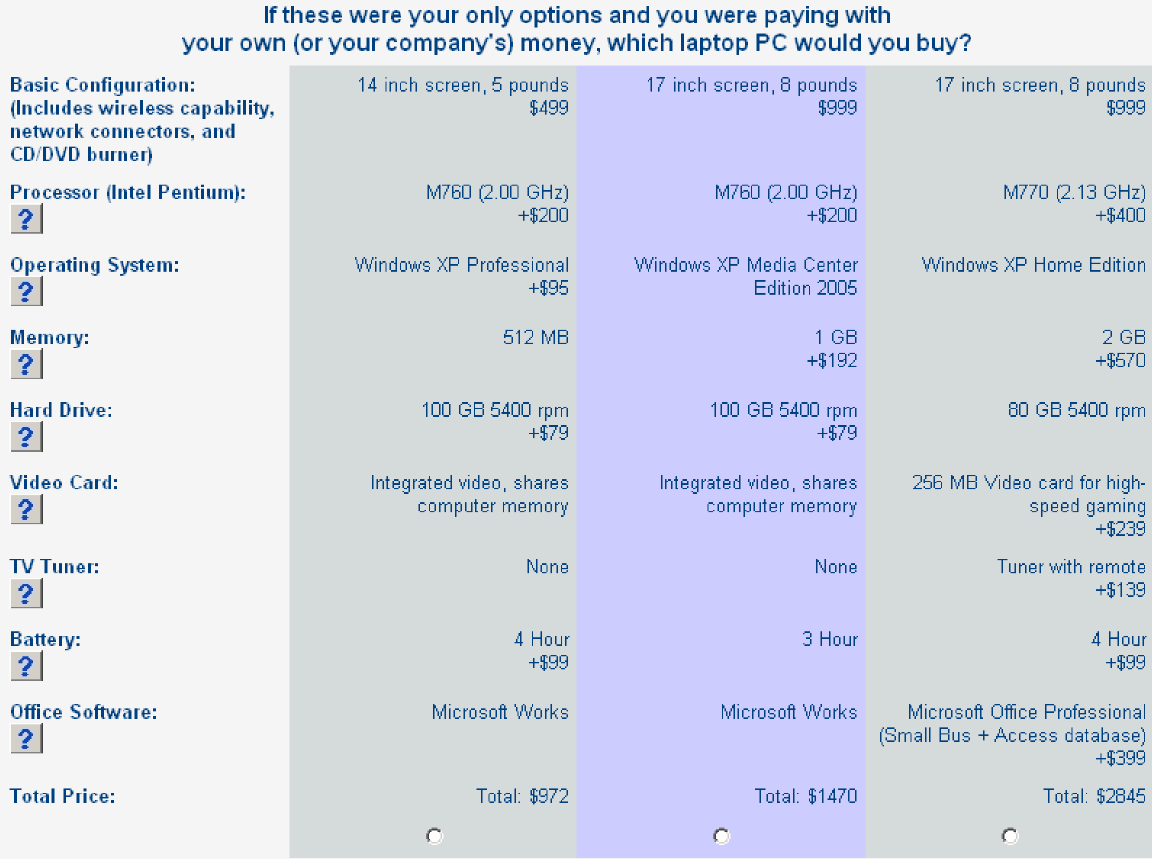
	512MB	1GB	2GB
Low	$0	$128	$380
Medium	$0	$160	$475
High	$0	$192	$570

As the respondent altered the specifications of the product being selected, the total price at the bottom of the screen was automatically updated to reflect the features selected. We planned an aggregate analysis, with price sensitivity information to be available from between-respondent effects.

[2] Our appreciation to GMI (Global Market Insite), and especially to Jim Weisfield, Amy Tung, and Miroslava Mitova from GMI who managed the data collection effort smoothly and without difficulties of any kind.

Each respondent did 4 CBC choice tasks, to provide comparative CBC data. Figure 2 shows the screen format of those CBC tasks

Figure 2: Screen Format for CBC Tasks:



If these were your only options and you were paying with your own (or your company's) money, which laptop PC would you buy?

Basic Configuration: (Includes wireless capability, network connectors, and CD/DVD burner)	14 inch screen, 5 pounds $499	17 inch screen, 8 pounds $999	17 inch screen, 8 pounds $999
Processor (Intel Pentium):	M760 (2.00 GHz) +$200	M760 (2.00 GHz) +$200	M770 (2.13 GHz) +$400
Operating System:	Windows XP Professional +$95	Windows XP Media Center Edition 2005	Windows XP Home Edition
Memory:	512 MB	1 GB +$192	2 GB +$570
Hard Drive:	100 GB 5400 rpm +$79	100 GB 5400 rpm +$79	80 GB 5400 rpm
Video Card:	Integrated video, shares computer memory	Integrated video, shares computer memory	256 MB Video card for high-speed gaming +$239
TV Tuner:	None	None	Tuner with remote +$139
Battery:	4 Hour +$99	3 Hour	4 Hour +$99
Office Software:	Microsoft Works	Microsoft Works	Microsoft Office Professional (Small Bus + Access database) +$399
Total Price:	Total: $972	Total: $1470	Total: $2845
	○	○	○

We showed only 4 CBC tasks because we wanted to compare BYO results with CBC results based on approximately similar amounts of interview time. The BYO task went more quickly than we anticipated, requiring an average of 68 seconds. Although time for the CBC tasks was not measured, previous research suggests that four tasks would have taken longer than that.

For the CBC tasks, as for the BYO tasks, prices for levels of each attribute were drawn randomly for each task. However, unlike BYO tasks, this meant that the respondent usually saw different prices for the same features in different tasks. Therefore, measurement of price sensitivity for CBC was a within-respondents treatment. For BYO, where the respondent saw each attribute only once, the measurement of price sensitivity was a between-respondents treatment.

There were 8 fixed holdout tasks, each repeated later in the interview. The holdout tasks also had the CBC format illustrated in Figure 2. Respondents were divided randomly into four groups, and each group was shown only two of the holdouts. Thus we have reliability measures for 8 distinct holdout tasks, each having been answered twice by approximately 150 respondents, and we can predict shares of choice using both CBC and BYO data.

Analysis & Results

Although each respondent saw a potentially unique set of prices in the BYO task, the number of levels shown for each attribute was the same for everyone:

	Feature Levels
Screen Size	2
Processor	4
Op. Sys.	3
Memory	3
Hard Disk	4
Video	3
TV Tuner	2
Battery	3
Software	4

The number of potential alternatives in each respondent's choice task was the product of these numbers of levels, 2 * 4 * 3 * 3 * 4 * 3 * 2 * 3 * 4 = 20,736.

We computed four sets of aggregate partworths:

1. Multinomial Logit partworths from BYO data, with **one enormous choice task** for each respondent.
2. Multinomial Logit partworths from BYO data, with **9 choice tasks** for each individual, each based on one attribute.
3. Approximate Logit partworths from BYO data, obtained by taking **logs of "counts"** (the percentage of times each level was offered that it was selected) and then zero-centering the logs within each attribute.
4. Multinomial Logit partworths from 4 x 605 = 2420 **CBC choice tasks**.

BYO partworths for "enormous" choice tasks and BYO partworths from separate choice tasks for each attribute were virtually identical. Their correlation was greater than .999999, and differences among them were less than 0.0003. If we had used tighter convergence criteria they would have been even closer. This verified our earlier finding from the analysis of artificial data. This is good news for users of BYO data, because it means that no special multinomial logit program is required for analysis.

Approximate partworths estimated from logs of counts were nearly identical to logit partworths. They were correlated .9898 with one another. Differences were due to the fact that respondents were randomly assigned to different price groups for each attribute, so the design was not precisely balanced. If it had been balanced the results should have been identical.

As an example of how partworths can be computed from counts, we provide the computation for Screen Size, which had two levels:

Attribute Level	Shown	Chosen	Ratio	Log	Centered	MNL
14 Inch Screen	605	345	.57025	-0.56168	.14143	.14143
17 Inch Screen	605	260	.42975	-0.84455	-.14143	-.14143

The column labeled "Centered" contains zero-centered logs of the ratio for each level of the number of times it was selected divided by the number of times it was shown. The final column, labeled "MNL," gives multinomial logit estimates from the "enormous" choice task model. Note that the estimates are identical to five decimal places. This is another favorable outcome. BYO data can be analyzed (at the aggregate level) without having to run multinomial logit at all. Simply taking logs of choice percentages and centering within attribute produces the same results.

BYO partworths and CBC partworths were rather different, and were correlated only 0.537. We describe and comment on those differences below.

We turn now to the quality of the holdout data which we are interested in predicting with BYO and CBC partworths. Each holdout task was evaluated by 150 respondents, who answered it twice. Test-retest reliability of individual choices averaged 72.1%, within the range usually seen for such statistics. When choice shares for the first administration of the holdouts were compared to shares for the second administration, we found a root mean square (RMS) error of 5 share points.

With an aggregate analysis, we are using partworths based on all 605 respondents to predict holdout data produced by subsets of 150. Therefore we can expect some additional prediction error just due to sampling error in the presence of heterogeneity. The standard error of a proportion of 1/3 with a sub-sample of 150 is about 3 share points. Therefore, with a perfect prediction, we can expect RMS errors of about

$$\sqrt{(3^2 + 5^2)} = 6 \text{ share points.}$$

Prediction of Holdout Choice Shares

With aggregate-level analysis, our test of validity is restricted to assessing prediction of choice shares, rather than individual hit rates. We used Sawtooth Software's RFC method of simulation, with default settings. The simulated market simulations were too extreme for each set of data. We found the constants with which to rescale each set of partworths which resulted in optimal share predictions, measured by root mean square (RMS) error.[3]

RMS Error for Predictions of Holdout Choice Shares

10.71	share points for BYO
11.24	share points for CBC

BYO predicted slightly more accurately than CBC, but the difference was small. More important, neither method was very successful. With "perfect" predictions we would expect RMS error of about 6, providing a rough lower bound on our error rate. Simply predicting shares of 33.333 for each alternative within each holdout task would produce a RMS error of about 13, providing a reasonable upper bound. Both methods produce results closer to 13 than to 6. We turn now to an examination of why both methods of prediction were less successful than we might have hoped.

[3] These constants were less than unity, and the BYO data required more "damping" than the CBC data. The optimal scale factors were 0.15 for BYO and 0.30 for CBC.

1. Price sensitivity is measured poorly by both methods: The same feature at a higher price should receive a lower partworth. We varied price on 17 attribute levels, each receiving three prices. There were a total of 51 pairwise relationships where we expect one partworth to be greater than another.

For CBC, 19 of these relationships, or 37% were violated.

For BYO, 23 of these relationships, or 45% were violated.

A look back at Figure 2 will show that respondents were burdened with a lot of information. The multiplicity of prices presents much more complexity than in a typical CBC task where there is only one price for each product alternative. Previous research has shown that when burdened with too much information, respondents tend to employ non-compensatory strategies to simplify their choices, such as looking for a particular feature. Such behavior can lead to high test-retest reliability, but impede prediction by a compensatory model such as logit.

We presented the price information by attribute in the choice tasks to make them as much like the BYO task as possible. In retrospect, it seems likely that respondents would have displayed more predictable choice behavior if we had made the holdout tasks simpler.

Also, BYO probably had worse violations of price relationships than CBC because for BYO price was a between-respondents treatment, while for CBC price was a within-respondent treatment, and therefore probably better measured.

For designs where price is varied within attribute, it may be necessary to have much larger sample sizes to get stable measurement of price sensitivity. Poor measurement of price sensitivity doubtless contributed to our prediction errors.

2. There are large context effects in BYO Data: Below are aggregate partworths for BYO data, after averaging over prices, and multiplying by 100. The first level is offered at the base price and subsequent levels typically provide greater capability at higher prices. For attributes with more than two levels, the second level nearly always has the highest partworth, and later levels always have negative partworths.

Average Partworths (*100, averaged over prices)

	Lev1	Lev2	Lev3	Lev4
Screen Size	14	-14		
Processor	-26	35	-21	-6
Operating Sys	-50	26	-8	
Memory	-85	42	-13	
Hard Disk	29	40	-6	-43
Video	14	45	-49	
TV Tuner	123	-41		
Battery	22	10	-18	
Software	32	50	-25	-36

This is not true for CBC data, where the first level is chosen more often than the second. Because this effect was present in the BYO data, but not in the holdout choices, it represents a relative disadvantage for BYO in predicting the holdout tasks.

The presence of this effect raises an interesting question: Should we try to build such effects into our predictions, or should we try to eliminate them? If actual buyer choices include such effects, then perhaps we should include them in our predictions. Some products, such as computers, are often purchased in a BYO environment. If CBC is used to estimate partworths for those products, perhaps the test should be whether CBC can predict BYO holdouts, rather than using BYO to predict CBC holdouts, as we have done in this study.

3. BYO and CBC may be measuring different things: Below are average BYO and CBC partworths for ***No*** TV Tuner (***presence*** of TV Tuner had negative partworths):

	No TV Tuner
BYO Partworths	123
CBC Partworths	17

This is a large difference that can affect share predictions. In CBC, where tradeoffs are **between attributes**, presence of a tuner may be relatively unimportant. But in BYO, where every attribute is traded off only with its price, even a relatively unimportant feature can get a large positive or negative part worth. Imagine a feature “Intel Inside® sticker on the case at a cost of $5”. It seems doubtful that many would pay $5 for such a sticker, so with BYO this feature might achieve a large negative partworth. But in CBC it would probably have little impact, compared to other attributes, and achieve a partworth of nearly zero.

PREDICTION OF BYO CHOICES

So far we have considered prediction of holdout choice tasks, rather than prediction of BYO product configurations chosen by respondents. We turn now to predicting BYO choices. There were 20,736 possible combinations of product features that could be selected. Considering all these combinations would have resulted in data too thin to provide adequate check of prediction accuracy. We narrowed our focus by choosing a subset of five attributes, thus drastically reducing the possible number of feature combinations, and by combining data for each of the three price treatments. Prices represent the middle of the three levels considered:

Attribute 1
1. 14-inch screen, 8 pounds, $499
2. 17-inch screen, 5 pounds, $999

Attribute 2
1. Windows XP Media Center Edition, +$0
2. Windows XP Home Edition, +$0
3. Windows XP Professional, +$119

Attribute 3
1. 512 MB Memory, +$0
2. 1 GB Memory, +$160
3. 2 GB Memory, +$475

Attribute 4

1. Integrated video, shares computer memory, +$0
2. 128 MB Video card, adequate for most use, +$49
3. 256 MB Video card for high-speed gaming, +$299

Attribute 5

1. 3 Hour battery, +$0
2. 4 Hour battery, +$99
3. 6 Hour battery, +$299

These attributes and levels permitted only 2x3x3x3x3 = 162 possible combinations. The data show that among 604 respondents, 129 of the possible 162 products were actually chosen. The data are shown in Figure 3, sorted by frequency of configuration.

Figure 3
Configuration Data, n=604

Config	Freq	Config	Freq	Config	Freq	Config	Freq
23333	28	21212	5	22332	3	13123	1
12221	25	21222	5	23211	3	13133	1
12111	24	22322	5	23231	3	13213	1
12211	21	22333	5	23233	3	13231	1
13222	19	23213	5	23311	3	13313	1
22222	16	11112	4	23321	3	13331	1
12222	14	11223	4	11121	2	21122	1
22221	14	11231	4	11131	2	21133	1
13221	13	11232	4	11213	2	21231	1
13212	12	12121	4	11312	2	21232	1
11222	11	13232	4	11313	2	21323	1
13211	11	13333	4	11333	2	22112	1
22212	11	21111	4	12113	2	22122	1
23222	11	21221	4	12231	2	22123	1
23323	11	21321	4	12332	2	22131	1
11221	10	21333	4	13111	2	22132	1
13323	10	22111	4	13113	2	22231	1
12212	9	22213	4	13121	2	22233	1
23322	9	22232	4	13122	2	22312	1
12322	8	22313	4	13233	2	22331	1
13223	8	23212	4	13311	2	23113	1
13322	7	23223	4	13321	2	23133	1
22121	7	23313	4	21121	2	23312	1
23332	7	11122	3	21313	2	23331	1
11111	6	11212	3	21331	2		
11211	6	11233	3	22311	2		
12213	6	11311	3	22323	2		
12311	6	12223	3	11322	1		
12321	6	12233	3	11323	1		
22211	6	12313	3	11331	1		
22223	6	13332	3	12123	1		
23221	6	21211	3	12131	1		
12112	5	21223	3	12331	1		
12312	5	21322	3	12333	1		
12323	5	22321	3	13112	1		

The most frequently configured product was the most feature-filled (most expensive) combination, 23333, which was chosen by 28 of 604 (5%) of respondents. Our initial attempts did not have great success in predicting this outcome. The base product, 11111, was also chosen often. Those two products (11111, 23333) reflect extreme situations in which respondents refuse to trade up, or trade up on all features. Some of these respondents may have tried to simplify their task by answering the questions in an easy way, and some may have legitimately desired extremely simple or complex products.

We decided to omit respondents choosing both extreme combinations, attempting to build a simulation model to predict the other 160 outcomes in which respondents had evidently made at least some tradeoff with price. Therefore, the remaining analysis considers only the remaining 571 respondents.

We randomly split 570 cases (we randomly dropped one case) into two groups of 285 respondents: Sample 1 and Sample 2. This was done five times, with different random splits each time, and several measures were computed for each split.

A naïve simulation model might assume that each of the 160 product combinations was chosen with the same frequency. If we use this to predict Sample 1's frequencies, we achieve an average RMS error of 2.29, as shown in the table below. This reflects the low-water mark, and any prediction model should be expected to perform much better.

RMS Errors

	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Average
Null Simulation Model	2.21	2.32	2.33	2.33	2.26	2.29
Split Sample Reliability	1.97	1.76	1.85	1.76	2.00	1.87
Tuned Split-Sample	1.66	1.64	1.72	1.64	1.74	1.68
ME Simulation Model	1.64	1.73	1.65	1.76	1.70	1.70
Joint Effects (2-way) Model	1.57	1.57	1.56	1.53	1.44	1.53

Next, we compared the raw frequencies for the 160 combinations of products between the two samples. Using the frequencies for Sample 2 to predict those for Sample 1 leads to an average split-sample reliability RMS error of 1.87.

When predicting one set of frequencies from another, the best predictions are usually made by "tuning" the predictors, shrinking them a little toward their means. We have done that and the results are shown in the third row of the table. The RMS error is reduced somewhat by this operation.

Using a main-effects-only model (again, tuned to obtain best fit), we get an average RMS error of 1.70, slightly larger than the error one would get using one sample's frequencies to predict the other. However, the joint-effects (two-way) simulator does better, with RMS error of 1.53. These error values are standard deviations so if we square them to reflect variances, the joint-effects model provides a 17% reduction in error variance as compared to the between-sample variance. We find it interesting that fitting the data with even such a simple model provides better external predictions than when using the raw data themselves.

We should comment on the ease with which such simulations can be done. For the main effects model we estimated logit partworths by just taking logs of the selection percentages (or "counts") for each attribute level and then zero-centering those values within each attribute. BYO choice shares were obtained by adding the partworths characterizing each product, exponentiating the partworth sums, and percentaging.

For the joint-effects model the procedure is just a little more complicated. Main effects partworths are handled as before. Two-way interactions are handled by computing the logs of counts in each two-way table and then double-centering each table. Then, when simulating, we add not only the main-effect partworths for each attribute level, but also the partworth for each two-way combination of levels of different attributes.

One of the things that is clear from this research is that large sample sizes are required for successfully modeling BYO choices. This requirement makes itself felt in two ways. First, the number of possible product configurations can be so enormous that extraordinarily large sample sizes might be required just to observe their relative frequencies (for purposes of predictive validity testing) without being overwhelmed by random error. We reduced this problem by reducing the number of attributes for our prediction attempts.

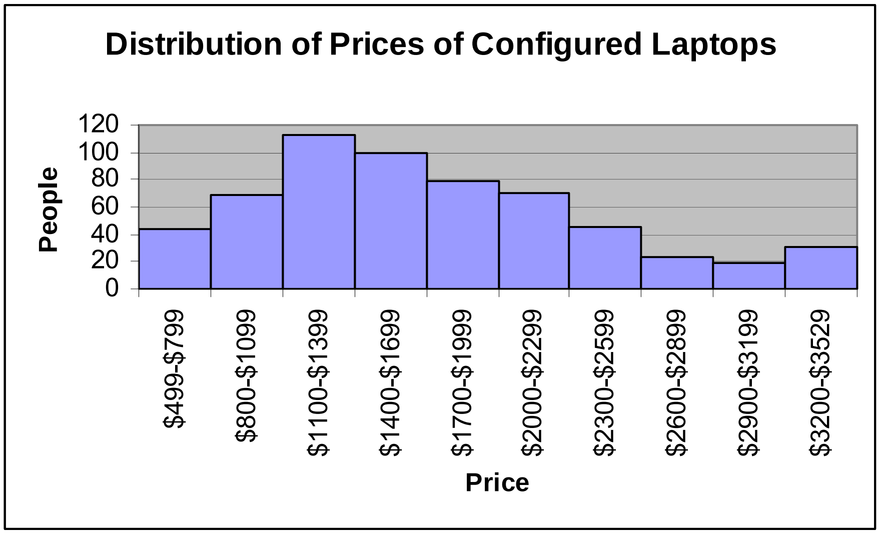
Second, the joint-effects model has greater requirements than the main effects model. With approximately 600 respondents, if attributes have three levels, an average of 200 respondents will choose each level. But for the two-way model, the interaction tables have nine cells rather than three, so there will be an average of only 67 respondents per cell, and some cells will have many fewer. The partworths depend directly on how many respondents end up in each cell, and the variability of such small frequencies can lead to instability.

THE EFFECT OF TOTAL PRICE ON CHOICE IN BYO

Even though the joint-effects BYO simulation model performed better than the actual choices themselves in predicting choices for new individuals, we noted a shortcoming. We built a joint-effects simulation model that included all 9 attributes and plotted the residuals of prediction vs. the total price for each configured product. We found an inverted U-shape pattern of residuals, suggesting that the model consistently under-predicts product combinations of middle prices and over-predicts product combinations of extreme prices. Total price for configured laptops ranged from a low of $499 to a high of $3,529.

The consistent under-prediction for mid-priced configurations may have something to do with the fact that respondents tended to configure laptops of mid-range rather than extreme prices (see Figure 3).

Figure 3



A model utilizing just lower-order effects (such as main effects plus two-way effects) might have difficulty accounting for the complexities involved in configuring a 9-attribute laptop subject to the propensity of respondents to trade up from bare-bones configurations on the low end and to avoid expensive configurations due to budgetary constraints on the high end.

We categorized total price into the ten categories as shown in Figure 3, computed counts for these categories, and applied the zero-centered logs of these counts (as main effects only) within the previously described joint-effects simulator. We found that adding total price as an additional effect to the simulator reduced or eliminated the U-shape pattern in the residuals. However, this solution seems rather *ad hoc* to us, rough (counts based on discrete price categories gives a rough stair-step function) and is not theoretically appealing. We suppose we might have dealt with this differently by fitting a quadratic function to the residuals based on total price. Whether this additional step is useful or necessary for other data sets remains to be seen, and we need to apply these models elsewhere before we can draw conclusions regarding a separate effect for total price in BYO simulators. If new data sets are considerably larger, we could try modeling even higher-order (such as three-way) effects.

SUMMARY

Our experiment confirmed earlier findings, and provided new information about differences between BYO and CBC. Among our findings are that:

- BYO tasks are quick and apparently easy for respondents. The average respondent completed the BYO task in slightly over one minute.
- BYO data can be analyzed with multinomial logit analysis, and two ways of conceptualizing the analysis produce identical results. Equivalent partworths can also be obtained without doing logit analysis at all.
- Aggregate simulations of shares for choice tasks can be done easily with BYO data, using methods customary in conjoint analysis. However, these predictions reflect the likelihood that respondents would configure a given product (via BYO) rather than choose a product from a set of pre-configured alternatives.
- BYO partworths contain less random error than CBC partworths, and for optimal prediction of choice shares from CBC tasks they need to be scaled down.
- Overall, BYO data were about as good as CBC data at predicting holdout choice shares, though neither method did well.
- CBC's uncharacteristically poor performance suggests that it is a bad idea to show alternatives with feature-specific prices. Partworths estimated from such data do not predict holdout shares well.
- Neither method measured price sensitivity well. When price sensitivity is handled with between-respondent comparisons, sample sizes much larger than 150 per treatment are required for stability.
- It is easy to model BYO choices using partworths estimated from simple "counts" frequencies. A two-way effects model had 17% smaller error variance than would be

obtained from predicting one half of the sample's choices from those of the other half of the sample.

Perhaps most interesting, we found notable differences between BYO and CBC partworths. Context effects are apparent in BYO data. Respondents tend to trade up one level from the (lowest) default level, but to avoid the top levels. The BYO exercise may be a good way to induce buyers to choose feature-rich products, but it may not be a good way to predict non-BYO choices. Also, unimportant attributes can achieve large partworths, which depend on the prices at which they are offered, rather than on tradeoffs with other attributes.

The fact that BYO partworths are able to predict share of choices as well as CBC (in this experiment), despite the presence of artifacts not found in the holdout data and weaker measurement of price sensitivity, indicates that it must have compensating strengths. Chief among these may be the fact that respondents can specify exactly what they prefer, rather than having to choose among sub-optimal alternatives.

BYO seems to provide a different kind of information than CBC, but it is information with less random error than CBC. One is reminded of the frequent statistical trade-off where one willingly accepts a small amount of bias to achieve a large reduction in random error. That may be the tradeoff that the researcher makes in choosing BYO over CBC when attempting to predict results from choice tasks. It may be the best approach for modeling purchase processes that are similar to, or make use of, BYO tasks.