

**PROCEEDINGS OF THE
SAWTOOTH SOFTWARE
CONFERENCE**

March 2000

Copyright 2000

All rights reserved. This electronic document may be copied or printed for personal use only. Copies or reprints may not be sold without permission in writing from Sawtooth Software, Inc.

FOREWORD

It is with pleasure that we present the Proceedings of the Eighth Sawtooth Software Conference, held in Hilton Head Island, South Carolina in March 2000. Nearly two-dozen papers were presented on various topics such as conjoint/choice analysis, Web interviewing, perceptual mapping, segmentation, and hierarchical Bayes estimation. Most of the papers focused on conjoint/choice estimation and predictions.

The “Most Valuable Presentation” award was voted on by conference attendees and announced at the end of the final session. Keith Chrzan (co-author Bryan Orme) received the award for his presentation entitled “An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis.” Keith reviewed the theory and practice of creating experimental designs for choice-based conjoint and compared different methods and software for generating those designs. A paper by Keith Sentis (co-author Lihua Li) entitled “HB Plugging and Chugging: How Much Is Enough?” received an honorable mention and represented a stunning amount of HB simulation work to provide guidelines regarding convergence of HB algorithms.

The authors (plus a few other researchers by invitation) also played the role of discussant to another paper presented at the conference. Discussants spoke for five minutes to express contrasting or complementary views. Many discussants have prepared written comments for this volume.

The papers and discussant comments are in the words of the authors, and we have performed very little copy editing. We wish to express our sincere thanks to the authors and discussants whose dedication and efforts made the 2000 Conference a success.

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

Sawtooth Software

September, 2000

CONTENTS

MOVING STUDIES TO THE WEB: A CASE STUDY	1
<i>Karlan J. Witt, Millward Brown IntelliQuest</i>	
Comment by <i>Keith Sentis</i>	19
TROUBLE WITH CONJOINT METHODOLOGY IN INTERNATIONAL INDUSTRIAL MARKETS	23
<i>Stefan Binner, bms GmbH</i>	
VALIDITY AND RELIABILITY OF ONLINE CONJOINT ANALYSIS.....	31
<i>Torsten Melles, Ralf Laumann, and Heinz Holling, Westfaelische Wilhelms-Universitaet Muenster</i>	
Comment by <i>Gary Baker</i>	41
BRAND/PRICE TRADE-OFF VIA CBC AND CI3	43
<i>Karen Buross, The Analytic Helpline, Inc.</i>	
CHOICE-ADAPTED PREFERENCE MODELLING	59
<i>Roger Brice, Phil Mellor, and Stephen Kay, Adelphi Group Ltd</i>	
CUTOFF-CONSTRAINED DISCRETE CHOICE MODELS.....	71
<i>Curtis Frazier, Millward Brown IntelliQuest and Michael Patterson, Compaq Computer Corporation</i>	
Comment by <i>Torsten Melles</i>	77
CALIBRATING PRICE IN ACA: THE ACA PRICE EFFECT AND HOW TO MANAGE IT.....	81
<i>Peter Williams and Denis Kilroy, The KBA Consulting Group Pty Ltd</i>	
Comment by <i>Dick McCullough</i>	97
USING EVOKED SET CONJOINT DESIGNS TO ENHANCE CHOICE DATA.....	101
<i>Sue York and Geoff Hall, IQ Branding (Australia)</i>	
Comment by <i>Bryan Orme</i>	111

PRACTICAL ISSUES CONCERNING THE NUMBER-OF-LEVELS EFFECT.....	113
<i>Marco Hoogerbrugge, SKIM Analytical</i>	
AN EXAMINATION OF THE COMPONENTS OF THE NOL EFFECT IN FULL-PROFILE CONJOINT MODELS.....	125
<i>Dick McCullough, MACRO Consulting, Inc.</i>	
CREATING TEST DATA TO OBJECTIVELY ASSESS CONJOINT AND CHOICE ALGORITHMS.....	141
<i>Ray Poynter, Millward Brown IntelliQuest</i>	
CLASSIFYING ELEMENTS WITH ALL AVAILABLE INFORMATION.....	153
<i>Luiz Sá Lucas, IDS-Interactive Data Systems, DataWise-Data Mining and Knowledge Discovery</i>	
AN OVERVIEW AND COMPARISON OF DESIGN STRATEGIES FOR CHOICE-BASED CONJOINT ANALYSIS	161
<i>Keith Chrzan, Maritz Marketing Research, and Bryan Orme, Sawtooth Software, Inc.</i>	
CUSTOMIZED CHOICE DESIGNS: INCORPORATING PRIOR KNOWLEDGE AND UTILITY BALANCE IN CHOICE EXPERIMENTS.....	179
<i>Jon Pinnell, MarketVision Research, Inc.</i>	
UNDERSTANDING HB: AN INTUITIVE APPROACH	195
<i>Richard M. Johnson, Sawtooth Software, Inc.</i>	
HB PLUGGING AND CHUGGING: HOW MUCH IS ENOUGH?	207
<i>Keith Sentis and Lihua Li, Pathfinder Strategies</i>	
PREDICTIVE VALIDATION OF CONJOINT ANALYSIS	221
<i>Dick Wittink, Yale School of Management</i>	
COMPARING HIERARCHICAL BAYES DRAWS AND RANDOMIZED FIRST CHOICE FOR CONJOINT SIMULATIONS	239
<i>Bryan Orme and Gary Baker, Sawtooth Software, Inc</i>	

SUMMARY OF FINDINGS

Nearly two-dozen presentations were given at our most recent Sawtooth Software conference in Hilton Head. We've summarized some of the high points below. Since we cannot possibly convey the full worth of the papers in a few paragraphs, the authors have submitted complete written papers for the 2000 Sawtooth Software Conference Proceedings.

Moving Studies to the Web: A Case Study (Karlan J. Witt): Karlan described the process her company has gone through in moving a tracking study (media readership for publications) from paper-and-pencil data collection to the Web. Her firm designed a study to compare the paper-and-pencil approach to three different administration formats over the Web. Respondents were recruited and pre-screened in the same way, and randomly assigned to one of the four questionnaire version cells. The Web data collection cells had lower response rates than the paper-based technique. The demographics of the respondents did not differ by gender, age, education or computer expertise between cells. But, there were some significant differences. Individuals with faster microprocessors were slightly more likely to complete the Web based survey, as well as those who owned other high-tech products. Karlan also reported some significant differences in the readership estimates for certain publications, depending on the data collection modality.

She concluded that generally similar results can be obtained over the Web as with paper. The look and feel of the graphical user interface of a Web survey, she maintained, can strongly affect both the response rate and the final sample composition. She suggested that the survey layout over the Web can affect the results as much as the choice of survey collection modality. Rather than try to replicate the look and functionality of past non-Web questionnaires over the Web, Karlan suggested designing the study to take advantage of what the Web can offer.

Trouble with Conjoint Methodology in International Industrial Markets (Stefan Binner): Stefan reported on a small survey his firm conducted among users of conjoint analysis (research professionals) in industrial or business to business markets. They achieved 37 responses. Stefan reported general satisfaction among these researchers for conjoint methods applied within b-to-b markets. The study also revealed some problems and challenges for conjoint researchers. The biggest hurdles in selling conjoint analysis projects are that clients don't always understand the technique and the costs of conjoint analysis studies. Other concerns voiced regarding conjoint analysis are the limited number of attributes that can be studied, the perception by some that the conjoint questions are not realistic, and the length of conjoint questionnaires. Despite the challenges, 70% of respondents reported being either satisfied or very satisfied with the use of conjoint analysis for industrial b-to-b markets. Furthermore, 63% of the respondents planned to increase the number of conjoint projects in the future.

Validity and Reliability of Online Conjoint Analysis (Torsten Melles, Ralf Laumann and Heinz Holling): Torsten presented evidence that useful conjoint analysis data can be collected over the Internet, although its reliability may be lower than for other data collection modalities. He cautioned that the suitability of conjoint analysis over the Internet depends on a number of aspects: the number of attributes in the design, the characteristics of the respondent, and the researcher's ability to identify unreliable respondents. Torsten suggested using multiple criteria

for determining the reliability of respondents using individual-level parameters. He also cautioned that IP-addresses and personal data should be carefully compared to guard against double-counting respondents.

Brand/Price Trade-Off via CBC and Ci3 (Karen Buros): Brand/Price Trade-Off (BPTO) is a technique that has been in use since the 1970s. Karen reviewed the complaints against BPTO (too transparent, encourages patterned behavior). She also proposed a hybrid technique that combines aspects of BPTO with discrete choice analysis. Her approach used an interactive computer-administered survey (programmed using Ci3) that starts with all brands at their mid-prices. Prior self-explicated rankings for brands are collected, and that information is used in determining which non-selected brands should receive lower prices in future tasks. She argued that her procedure results in an interview that is not as transparent as the traditional BPTO exercise. Furthermore, the data from the experiment can be used within traditional logit estimation of utilities.

Choice-Adapted Preference Modeling (Roger Brice, Phil Mellor & Stephen Kay): The high cost of interviewing physicians makes it especially desirable to maximize the amount of information obtained from each interview. Although choice-based conjoint studies have many desirable properties, traditional conjoint procedures involving ranking or rating of concepts can provide more information per unit of interview time. One of the features of choice-based conjoint is the availability of the “None” option. The authors used an extension of traditional conjoint which asked respondents not only to rank concepts, but also to indicate which of them were below the threshold of the respondent’s willingness to accept. They reported that this question extension allowed them to extend the None option into the simulation module on an individual-level basis and then to more realistically simulate the ability of new products to capture share from their current competition.

Cutoff-Constrained Discrete Choice Models (Curtis Frazier & Michael Patterson): With traditional discrete choice data, analysis is performed at the aggregate by pooling data. This assumes respondent homogeneity, and ignores differences between individuals or segments. The authors tested two recent techniques for incorporating heterogeneity in discrete choice models: Hierarchical Bayes (HB) and “soft penalty” cutoff models.

The “soft penalty” models involve asking respondents to state what levels of attributes they would never consider purchasing. “Soft” penalties are estimated through logit analysis, by creating a variable for each attribute level that reflects whether (for categorical variables) or by how much (for quantitative variables) a choice alternative violates a respondent’s cutoff.

A research study with 450 respondents was used to test the predictive validity of HB estimation versus the “soft penalty” approach. HB performed better, both in terms of hit rates and share predictions for holdout choice tasks. They concluded that penalty models do not always produce better predictive validity, often have odd coefficients, and can make the model explanation more complex than with HB estimation.

Calibrating Price in ACA: The ACA Price Effect and How to Manage It (Peter Williams & Denis Kilroy): Even though ACA is one of the most popular conjoint analysis techniques, it has been shown often to understate the importance of price. Peter summarized hypotheses about why the effect happens, and what to do about it. He suggested that the ACA price effect is due to

a) inadequate framing during importance ratings, b) lack of attribute independence, c) equal focus on all attributes, and d) restrictions on unacceptable levels.

To overcome the effect, Peter explained, the first step is to quantify it. Other researchers have proposed a “dual conjoint” approach: employing either full-profile traditional conjoint or CBC along with ACA in the same study. Peter proposed a potentially simpler technique, that of using Choice-Based holdout tasks (partial profile). The holdout tasks typically include a stripped down product at a cheap price, a mid-range product at a medium price, and a feature-rich product at a premium price. He suggested counting the number of times respondents choose higher-priced alternatives (this requires multiple choice tasks), and segmenting respondents based on that scored variable. Then, Peter developed a separate weighting factor for the ACA price utilities for each segment. He demonstrated the technique with an ACA study involving 969 respondents. He found that no adjustment was necessary for the group that preferred high-priced, high-quality offers; a scaling factor of 2 for price was required for the mid-price segment; and a scaling factor of slightly more than 4 was required for the price sensitive segment.

Using Evoked Set Conjoint Designs to Enhance Choice Data (Sue York & Geoff Hall): Some conjoint designs require very large numbers of brands or other attribute levels. One way around this problem is to ask each respondent about a customized “evoked set” of brands. Sue tested this approach using a study with slightly more than a dozen brands and a dozen package types, along with price, which needed to be modeled using CBC. Respondents were either shown a traditional CBC interview including all attribute levels, or were given a customized CBC interview (programmed using Ci3), based on an evoked set of brands and package sizes. For the customized design cell, respondents indicated which brands and package types they would be most likely to buy, and which they would not buy under any conditions (“unacceptables.”) Holdout tasks were used to test the predictive validity of the two design approaches, and to compare the performance of aggregate logit versus individual-level HB modeling.

HB estimation generally was superior to a main-effects specified logit model. The customized “Evoked Set” design resulted in higher hit rates, but lower share prediction accuracy relative to the standard CBC approach. Sue reported that the “unacceptables” judgements had low reliability. Nine percent of respondents chose a brand in the holdout choice tasks that they previously declared unacceptable, and a full 29% chose a package size previously marked unacceptable. This inconsistency made it difficult to use that information to improve the predictability of the model. She concluded that customized “evoked set” CBC designs can be useful, especially if many levels of an attribute or attributes need to be studied and predicting respondents’ first choices is the main goal.

Practical Issues Concerning the Number-of-Levels Effect (Marco Hoogerbrugge): Marco reviewed the “Number of Levels Effect” (NOL) and examined its magnitude in commercial data sets. The NOL effect is a widely-experienced tendency for attributes to be given more importance when they have larger numbers of intermediate levels, even though the range of values expressed by the extreme levels remains constant. Marco concluded “ the NOL effect clearly exists and is large when comparing 2-level attributes with more-level attributes, but the effect is questionable when you only have attributes with at least 3 levels.

An Examination of the Components of the NOL Effect in Full-Profile Conjoint Models (Dick McCullough): There has been much discussion about whether the “Number of Levels Effect” (NOL) is “algorithmic,” e.g., due to mathematical or statistical artifacts, or

“psychological,” e.g., due to respondents’ behavioral reactions to perceiving that some attributes have more levels than others. Dick McCullough carried out an ingenious and elaborate experiment designed to measure the extent to which the NOL might be caused by each factor. Data were collected on the Web, and the variable whose number of levels was manipulated was categorical, rather than a continuous one. Preliminary self-explicated questions were used to determine each respondent’s best-and-least liked categories, and then two conjoint studies were carried out for each individual, one with just those extreme levels and the other with all levels. There were four cells in the design, in which those two treatments were administered in different orders, and with different occurring activities between them. Dick found some evidence for a positive algorithmic effect and a negative psychological effect, but the results may have been compromised by difficulties with respondent reliability. This is a promising method that we hope will be tried again, manipulating the number of levels of a continuous rather than a categorical variable.

Creating Test Data to Objectively Assess Conjoint and Choice Algorithms (Ray Poynter): Ray characterized the success of conjoint techniques in terms of two factors: the ability of the technique to process responses (efficiency), and the likelihood that the technique elicits valid responses. In his research, he has sometimes observed respondents who don’t seem to respond in the expected way in conjoint studies. He suggested that respondents may not be able to understand the questions, visualize the attributes, or the respondent’s own preferences may not be stable. He even suggested that respondents may not know the answer, or that asking the questions may change the responses given.

Ray described a procedure for using computer-generated data sets based on the part worths contained in real data sets to test different conjoint designs and different respondent response patterns. He found no evidence of a number-of-levels effect from simulated data. He suggested that analysts can make better decisions regarding the type of conjoint design using simulated data.

Classifying Elements with All Available Information (Luiz Sá Lucas): Luiz described a neural network procedure that combines different classification solutions obtained using different methods. He showed how results from different approaches such as Discriminant Analysis and Automatic Interaction Detection can be combined to produce final classifications better than those of any single method. This can have practical value classifying new individuals into pre-existing groups, such as segments expected to respond differently to various marketing approaches.

*** An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis** (Keith Chrzan & Bryan Orme): Keith reviewed the basics of designing conjoint experiments, including the use of design catalogues for traditional card-sort designs. He described the additional complexities of designing Choice-Based Conjoint experiments, along with the potential benefits that accompany them (i.e. ability to model cross-effects, alternative-specific effects, inclusion of “none”). Keith compared four methods of generating CBC tasks: catalog-based approaches, recipe-based designs for partial-profile experiments, computer optimized designs (SAS OPTEX) and random designs using Sawtooth Software’s CBC program. Different model specifications were explored: main-effects, interaction effects, cross- and alternative-specific effects.

Using computer-generated data sets, Keith compared the relative design efficiency of the different approaches to CBC designs. The main conclusions were as follows: minimal level overlap within choice tasks is optimal for main effects estimation, but some level overlap is desirable for estimating interactions. SAS OPTEX software can create optimal or near-optimal designs for most every design strategy and model specification. Sawtooth Software's CBC system can create optimal or near optimal designs in most every case as well, with the exception of a design in which many more first-order interaction terms are to be estimated than main effects (in that case, its best design was 12% less efficient than SAS OPTEX). Keith concluded that no one design strategy is best for every situation. Keith suggested that analysts create synthetic data sets to test the efficiency of alternative design generation methods for their particular studies.

(* Best Presentation Award, based on attendee ballots.)

Customized Choice Designs: Incorporating Prior Knowledge and Utility Balance in Choice Experiments (Jon Pinnell): Jon investigated whether customizing the degree of utility balance within CBC designs can offer significant improvements in predictive validity. He reviewed past findings that have pointed toward increased design efficiency if the alternatives in each choice set have relatively balanced choice probabilities. He reported results for two experimental and three commercial CBC studies. Jon found that those choice tasks with a higher degree of utility balance resulted in slightly more accurate predictive models than tasks with a lower degree of utility balance. He also observed that HB estimation can significantly improve hit rates at the individual-level relative to the simple aggregate logit model.

Jon reported results from a computer-administered CBC survey that dynamically customized the choice design for each respondent using the criterion of utility balance. Again, he demonstrated small positive gains for utility balance. One interesting finding was that HB estimation applied to a partial-profile CBC design resulted in inferior predictions (hit rates) relative to aggregate logit for one data set. Jon was puzzled by this finding, and used this experience to suggest that we not blindly apply HB without some cross-checks. Overall, he concluded that the gains from HB generally outweighed the gains from utility balance for the data sets he analyzed.

Understanding HB: An Intuitive Approach (Richard M. Johnson): Many speakers at this Sawtooth Software conference spoke about Hierarchical Bayes estimation or presented HB results. HB can be difficult to understand, and Rich's presentation helped unravel the mystery using an intuitive example. In Rich's example, an airplane pilot named Jones often had rough landings. With classical statistics, Rich stated, we generally think of quantifying the probability of Jones having a rough landing. With Bayesian reasoning, we turn the thinking around: given that the landing was rough, we estimate the probability that the pilot was Jones. Rich presented Bayes' rule within that context, and introduced the concepts of Priors, Likelihoods and Posterior Probabilities.

Rich outlined some of the benefits of HB analysis. He argued that it can produce better estimates from shorter conjoint questionnaires. Analysts can get individual estimates from CBC, customer satisfaction models or scanner data instead of just aggregate results. Rich showed results from real data sets where HB improved the performance of CBC, ACA and traditional ratings-based conjoint data. He also provided some run time estimates ranging between 1 and 14

hours for 300 respondents and models featuring between 10 to 80 parameters. He stated the opinion that analysts should use HB whenever the project scheduling permits it.

*** HB Plugging and Chugging: How Much Is Enough?** (Keith Sentis & Lihua Li): Hierarchical Bayes (HB) analysis is gaining momentum in the marketing research industry. It can estimate individual-level models for problems too difficult for previous algorithms. But, the added power and accuracy of the models come with the cost of run times often measured in multiple hours or even days. HB is an iterative process, and it has been argued that tens of thousands of iterations may be necessary before the algorithm converges on relatively stable estimates.

Keith and Lihua investigated whether HB run times could be shortened without sacrificing predictive accuracy (as measured by hit rates for hold out tasks). Based on 22 real choice-based conjoint data sets, thousands of separate HB runs and tens of thousands of computing hours, Keith and Lihua concluded that 1,000 initial iterations and 1,000 saved draws per respondent are enough to maximize predictive accuracy of hit rates. The authors presented evidence that their findings with respect to choice-based conjoint also generalize to regression-based HB and HB for Adaptive Conjoint Analysis.

(* Honorable Mention, based on attendee ballots.)

Predictive Validation of Conjoint Analysis (Dick Wittink): Dick enumerated the conditions under which conjoint analysis predictions can be accurate reflections of actual market choices: respondents are a probability sample of the target market; respondents are decision makers; respondents are motivated to process conjoint tasks as they would in the marketplace; all relevant attributes are included; respondents understand the product category and attributes; respondents can be weighted by purchase volume; individual-level brand awareness and availability data are available.

Dick stressed that if the conditions are right, conjoint analysis can provide accurate forecasts of market behavior. However, most conjoint analysis projects do not have the benefit of actual purchase data, collected some time after the conjoint study. Rather, holdout tasks are added to the study. Dick argued that holdout tasks are easy to collect, but they are in many respects artificial and reflective of the same thought processes used in the conjoint judgements. Holdout task predictability, he maintained, is therefore biased upward relative to how well the conjoint model should predict actual behavior. Nonetheless, holdout tasks can be useful for comparing the suitability of different methods, modes and models. Dick advised that holdout tasks should be carefully designed to resemble marketplace alternatives, and so that no one option dominates.

Dick also shared some guiding principles gleaned from years of experience: if using aggregate measures of performance, such as the error in predicting shares, complex models (e.g. flexible functional forms, interaction effects between attributes) generally outperform simple models; on individual measures of performance, such as the percent of holdout choices predicted, simple model specifications often outperform complex models; constraints on part worths, as currently practiced, generally improve individual-level hit rates but usually damage aggregate share prediction accuracy; motivating respondents improves forecast accuracy.

Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations (Bryan Orme & Gary Baker): A number of recent advances can improve part worth estimation and the accuracy of market simulators. The authors stated their opinion that HB

estimation is the most valuable recent addition to the conjoint analyst's toolbox. Another advancement has been the introduction of Randomized First Choice (RFC) as a market simulation method.

Conjoint simulators have traditionally used part worths as point estimates of preference. HB results in multiple estimates of each respondent's preferences called "draws" which might also be used in simulations. Both HB and Randomized First Choice reflect uncertainty (error distributions) about the part worths. RFC makes simplifying assumptions. The authors showed that HB draws, though theoretically more complete, have some unexpected properties, and the data files can become enormous (on the order of 100 Meg or easily more).

Bryan and Gary provided theoretical and practical justifications for RFC. They demonstrated that RFC using point estimates of preference performed slightly better than first choice or share of preference (logit) simulations on either point estimates or the draws files for one particular data set. Their findings suggest that draws files are not needed for market simulations, which is good news for in-the-trenches researchers. Using RFC with the much smaller and more manageable file of point estimates seems to work just as well or even better.

MOVING STUDIES TO THE WEB: A CASE STUDY

Karlan J. Witt
Millward Brown IntelliQuest

BACKGROUND

Millward Brown IntelliQuest is a research supplier focused on the clients in the technology industry. As the technology industry has matured, several common information needs have converged. These needs include market sizing, brand share, buyer profiles, customer satisfaction, and so on. Often where these common needs exist, IntelliQuest provides multi-client studies, allowing clients who share the same information needs to share the same data at a reduced cost.

One such study that has run consistently for the last seven years is a study that performs two key functions: to determine the size and profile of both the home and the business technology purchase influencer universe, and then to examine extensive media readership information among those purchase influencers. The name of the study is the Technology Industry Media Study.

Volumes have been published discussing the collection of media readership data using different data collection methods. Based on that collective research, IntelliQuest uses a hybrid data collection methodology, with a telephone screener and a paper-by-mail survey (the methodology is detailed in the Research Design section below).

The Internet has grown as a means of collecting a variety of data for technology clients (Thompson, 1999; Gates and Heloton, 1998). Consequently, we formed an advisory council of clients subscribing to this to evaluate the potential of migrating this tracking study from telephone and paper to telephone and web. This paper discusses the results of that test.

The advisory council considered both strategic and practical issues related to moving the study to the web. These included:

- Strategic issues:
 - Greater depth of information required
 - Shorter turn-around time
 - Maximize cost efficiency
 - Maximize accuracy of data
- Practical issues:
 - Paper survey methodology becoming more cumbersome to field
 - Acceptable levels of respondent cooperation harder to achieve
 - Reduction of respondent burden

RESEARCH DESIGN

To determine the viability of migrating this study to the web, our objectives were to evaluate the following issues:

1. overall response rates,
2. demographic composition of respondents by modality,
3. responses to behavioral and attitudinal questions, and
4. readership levels as measured by each method.

In order to test the differing modalities, we decided to create four test cells, one utilizing the existing paper modality and three looking at variants of the Web option. We began with one general sample source, screened all respondents using the same screener, and then randomly assigned respondents to one of four test modality cells. In this section we describe each phase, as well as define the variations in the test cells.

Phase One

Since the goal of this test was to compare results using paper and web modalities, we needed to ensure that the survey was going to people similar to those we survey in our on-going tracking study. To obtain a large number of technology purchase influencers to populate the four test cells needed, we drew the sample from a database of those who had registered a computer or computer-related product in the past year. IQ2.net, a database marketing firm that until recently was owned by IntelliQuest¹, provided a list of 27,848 records selected at random from their computer software and hardware registration database.

Telephone Pre-Screen

During the screening interview, respondents were qualified based on:

- their involvement in the purchase process of technology products for their home in the last 12 months, or their intention to in the next 12 months,
- being 18 years of age or older,
- living in the U.S.,
- having access to the web, and
- their willingness to supply both mailing and email addresses.

Approximately 2760 phone interviews were completed. Once IntelliQuest obtained the respondents' email and post office addresses, we sent them a "thank you" letter by regular mail and a two-dollar participation incentive.

¹ IQ2.net now operates independently of IntelliQuest as part of Naviant Technology Solutions.

Phase Two

Once qualified, respondents were allocated on a random basis to one of the following four test cells for Phase Two of the interview:

- paper-by-mail (sent with the thank you letter and incentive)
- web survey using horizontal layout
- web survey using modified horizontal layout
- web survey using vertical layout

Non-responders received up to two additional mail or email follow-ups as appropriate. Regardless of whether the questionnaire was web-based or paper-based, the respondent had to supply an email address. When attempting to contact the web respondents electronically, 18% of their email addresses turned out to be invalid. In order to ensure comparable samples, all recipients of the paper questionnaire were also sent an email to verify their email address. Those who failed such validation were removed from the respondent base so that all four groups were known to have valid email addresses.

Phase Two Survey Instruments

The content of all four surveys was identical. In the Comparison of Results section of this paper, we will show the comparison of overall response rates, and differing distributions of responses across survey questions. For simple questions such as product ownership, job title, and income, question format was virtually identical across the four executions. For one area of the questionnaire, however, we had the opportunity to examine the impact of *format within modality* on the web. The layout of that question series for each version is shown below. This section was for measuring the media readership information.

Paper-by-mail


Research to-date on the validity of readership information has demonstrated tremendous title confusion with publications of similar titles when asking respondents verbally based on the name of the publication. That is, when asked “have you read ‘*PC Week*’?” studies have found that respondents confuse similar publications such as *PC Magazine* or *InfoWeek* when reporting their answer. For this reason, the most valid method of gathering that information is considered to be by visually showing the logo or cover of the magazine as a stimulus. Below is an example of how this is shown in our paper-by-mail study.

Web survey using horizontal layout

As mentioned, in addition to testing the raw response rate between paper-by-mail and web, this test evaluates the effect of differing stimuli within the same modality. The first execution was made to look as similar as possible to the existing paper questionnaire, to pick up differences in answers based on the survey mode rather than the stimulus. Below is a sample of this format.

This section is about some of the publications that you, yourself, may read or look into. When answering whether you have read or looked into each of the publications, please include any issues you may have read at work, at home, at school or elsewhere, as well as those you happen to glance through. Beginning with the first publication in the list, please fill out the entire row of questions from left to right before moving to the next publication in the list.

If you do read a publication, answer "YES" to the first question. Please note, that if you do not answer every question on the screen, you will be returned to this screen with an error message. If you do not read a publication, answer "NO" to the first question and leave the rest of the questions as "NA".

Publications	Frequency	Have you read or looked into any issue of this publication in the past six months?	How many issues (of the publications listed below) do you usually read or look into out of every four that are published? [Click on the appropriate answer]	How closely do you read or examine the advertising for computer hardware, software, and communications related products and services in this publication? [Click on the appropriate answer: 1=not at all closely 2=Somewhat closely 3=very closely 4=extremely closely.]	What percentage of the pages do you usually look at or read in the course of your reading a typical issue? [Click on the appropriate answer: 1=just a few 2=about 25% 3=about 50% 4=about 75% 5=all or most.]
	Monthly	<input type="radio"/> Yes <input type="radio"/> No	<input type="radio"/> Less than one <input type="radio"/> One <input type="radio"/> Two <input type="radio"/> Three <input type="radio"/> Four <input checked="" type="radio"/> NA	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> NA	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input checked="" type="radio"/> NA

Respondents received the screening and three follow-up questions for each publication before going on to the same sequence for the next publication until all 94 publications were completed. As with the paper-based format, all publications were represented by a black-and-white logo with publication frequency shown prior to the screening question. Respondents had to provide a “yes” or “no” to the six-month screening question, and were expected to provide reading frequency and two qualitative evaluations (referred to collectively as “follow-up questions”) for each publication screened-in.

Due to design considerations inherent to the web modality, the horizontal web version differed from the paper questionnaire in a few ways. First, only seven publications appeared on screen at a time compared with the 21 shown on each page of the paper version. This formatting constraint stemmed from a limitation in some versions of AOL’s web browser.

Second, the horizontal web version did not allow inconsistent or incomplete responses as did the paper version. The Quancept Web surveying program used does not currently allow

respondents to continue to the next screen if a question remains unanswered on the previous one. This means that the follow-up questions could not be left blank, even in the instance where a respondent claimed not to have read the publication in the past six months.


To reduce the respondent’s burden but still provide complete responses, the horizontal Web version included an “NA” (not applicable) category at the end of each follow-up question. The “NA” button was the default selection for each question, meaning that the respondent did not have to actively select that response for all non-screened publications. The programming did not permit the respondent to go to the next set of publications unless all questions were answered completely and consistently for the ones previous.

Web survey using modified horizontal layout

This was identical to the horizontal version except that it assumed an answer of “no” for the six-month screen. This allowed respondents to move past publications they had not read in the last six months, similar to the way respondents typically fill out paper questionnaires. As in the web-based horizontal version, only seven publications appeared on each screen.

This section is about some of the publications that you, yourself, may read or look into. When answering whether you have read or looked into each of the publications, please include any issues you may have read at work, at home, at school or elsewhere, as well as those you happen to glance through. Beginning with the first publication in the list, please fill out the entire row of questions from left to right before moving to the next publication in the list.

If you do read a publication, answer "YES" to the first question. Please note, that if you do not answer every question on the screen, you will be returned to this screen with an error message. If you do not read a publication, don't answer the first question and leave the rest of the questions as "NA".

Publications	Frequency	Have you read or looked into any issue of this publication in the past six months?	How many issues (of the publications listed below) do you usually read or look into out of every four that are published? [Click on the appropriate answer]	How closely do you read or examine the advertising for computer hardware, software, and communications related products and services in this publication? [Click on the appropriate answer: 1=not at all closely 2=Somewhat closely 3=very closely 4=extremely closely.]	What percentage of the pages do you usually look at or read in the course of your reading a typical issue? [Click on the appropriate answer: 1=just a few 2=about 25% 3=about 50% 4=about 75% 5=all or most.]
	Monthly	<input type="radio"/> Yes <input checked="" type="radio"/> NA	<input type="radio"/> Less than one <input type="radio"/> One <input type="radio"/> Two <input type="radio"/> Three <input type="radio"/> Four <input checked="" type="radio"/> NA	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> NA	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input checked="" type="radio"/> NA

Web survey using vertical layout

This version differed most from the original paper-based horizontal format and maximized our ability to enforce skip patterns on the web. Respondents were first subjected to a six-month screen using only black-and-white logo reproductions. After all 94 publications were screened, respondents received follow-up questions (frequency of reading and the two qualitative questions) for only those titles screened-in. It was assumed that hiding the presence of follow-up questions on the web would lead to higher average screen-ins similar to the findings of Appel and Pinnell (1995) using a disk-based format and Bain, et al (1995) and (1997) using Computer-Assisted Self-Interviewing.

For the vertical version, each respondent was asked to indicate those publications they had read in the past six months. To indicate non-readership of all seven publications on a screen in a manner consistent with Quancept programming requirements, a “none of the above” option must have been selected. This appeared as a checkbox following the seventh publication, displayed in a horizontal format to facilitate maximum exposure on the average computer screen. The respondent must have selected at least one publication or the “none of the above” box to continue. An example of that format is shown below.

This section is about some of the publications that you, yourself, may read or look into. When answering whether you have read or looked into each of the publications, please include any issues you may have read at work, at home, at school or elsewhere, as well as those you happen to glance through.

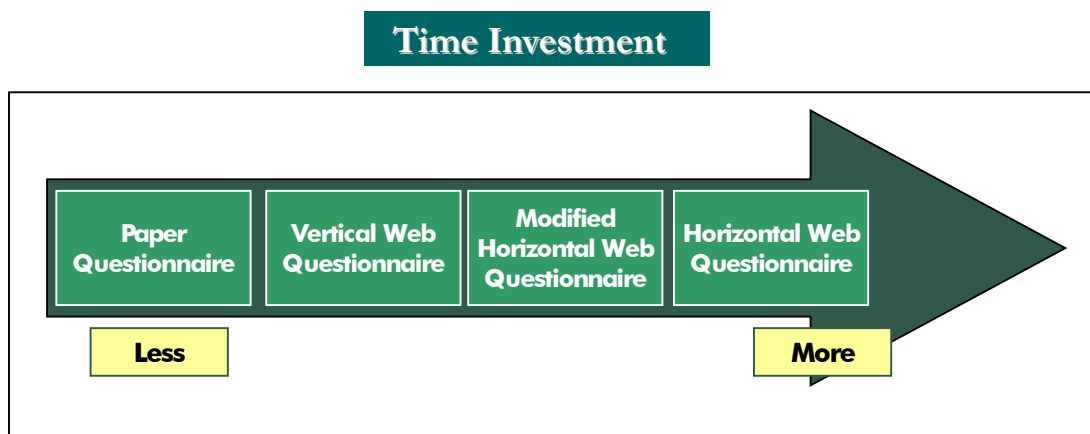
Have you read or looked into any issue of these publications in the past six months?
[Check each publication read]

<input type="checkbox"/> COMPUTER SHOPPER	<input type="checkbox"/> PC MAGAZINE	<input type="checkbox"/> Smart Computing
<input type="checkbox"/> HomeOffice Computing	<input type="checkbox"/> PCWEEK	<input type="checkbox"/> None of the above
<input type="checkbox"/> PCComputing	<input type="checkbox"/> PCWORLD	

Timing

- Telephone recruitment (June-July 1999)
- Survey mail-out / links (July-August 1999)
- Data processing (August-September 1999)

The graph below shows a comparison of the time required to complete each of the versions of the Phase Two questionnaire.



COMPARISON OF RESULTS

In this section we compare the results of the studies, as described under the research objectives:

- overall response rates,
- demographic composition of respondents by modality,
- responses to behavioral and attitudinal questions, and
- readership levels as measured by each method.

Overall Response Rates

The response rates to the email solicitations and paper questionnaires are shown in the table below, from which it can be seen that the paper-based questionnaire generated the highest response rate with 54.3%, or 376 respondents. The web vertical questionnaire generated the next highest response rate (47.8%) and the two horizontal versions were statistically indistinguishable from one another in third place (37.4% and 39.7%).

Main and Email Responses

Version	Invited	Completed	Percent
Paper, Horizontal	693	376	54.3%*
Web, Horizontal	690	258	37.4%**
Web, Modified Horizontal	688	273	39.7%**
Web,	689	329	47.8%*
Total (E)Mailed	2,760	1,236	44.8%

*Significantly different from all versions at the 95% confidence level
**Significantly different from vertical web and horizontal paper versions at the 95% confidence level

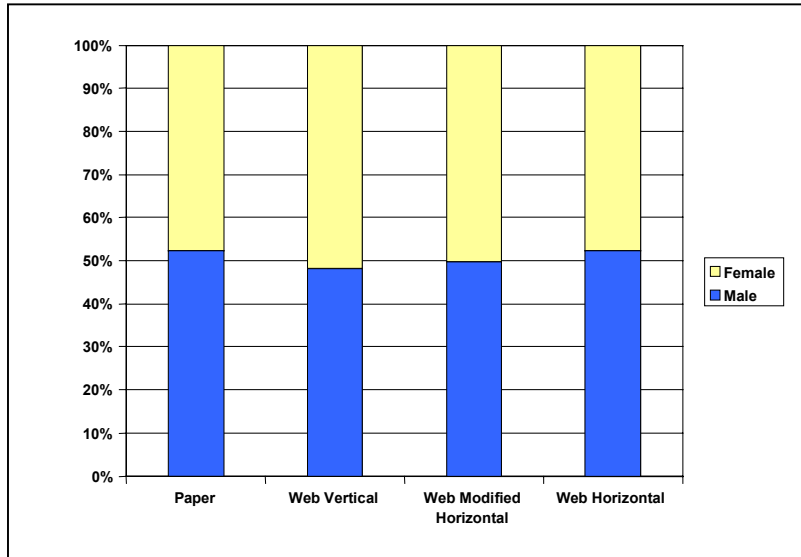
The fact that the vertical web questionnaire generated higher response rates than either of the horizontal web versions was not unexpected. Theoretically, the vertical web format is superior to the two horizontal formats because the respondent has no way of knowing that there is a penalty to be paid for screening-in. The penalty appears after the screen-in questions in the form of follow-up questions to be answered.

Although we were aware of the impact this might have on readership estimates, we could not be sure to what degree it would translate to a response rate increase. One of the benefits of performing surveys on the web is the ability to determine how many individuals started the questionnaire and where they dropped out. We now know that a vertical format has a significant advantage in terms of retaining respondents, although the vertical web format's response rate still falls short of that achieved by the horizontal paper format using the same incentive and number of contacts.

Demographic composition of respondents by modality

In this section we compare the demographic composition of each audience. First, we see the comparison of gender by survey modality.

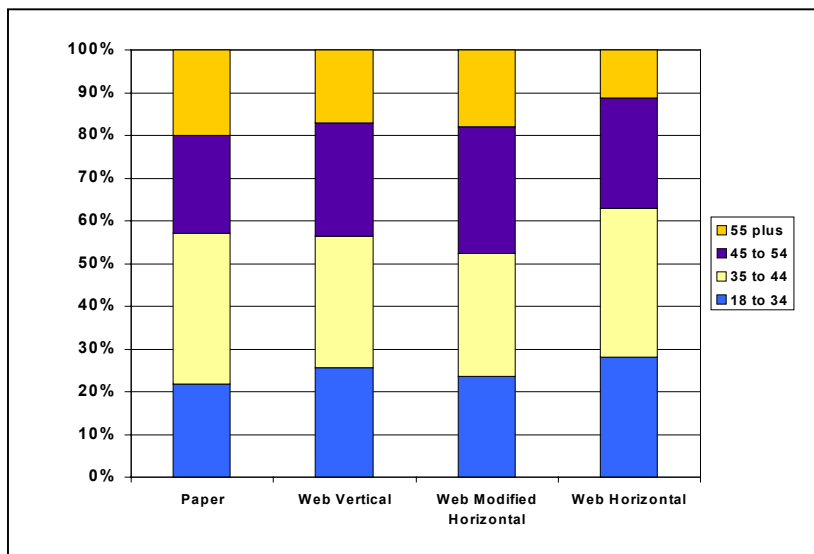
Respondent Gender



These slight differences are not statistically significant, showing that both survey modalities and all survey executions had similar response rates for male and female respondents.

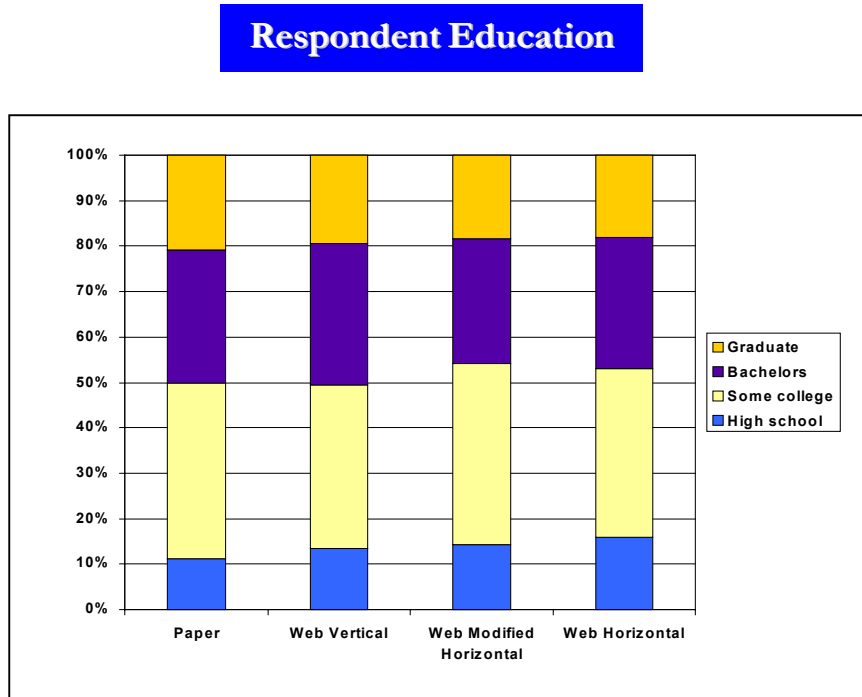
The next demographic compared was respondent age.

Respondent Age



Once again, the differences were not statistically significant.

We then looked at the distribution across level of education by mode and the results are shown below.

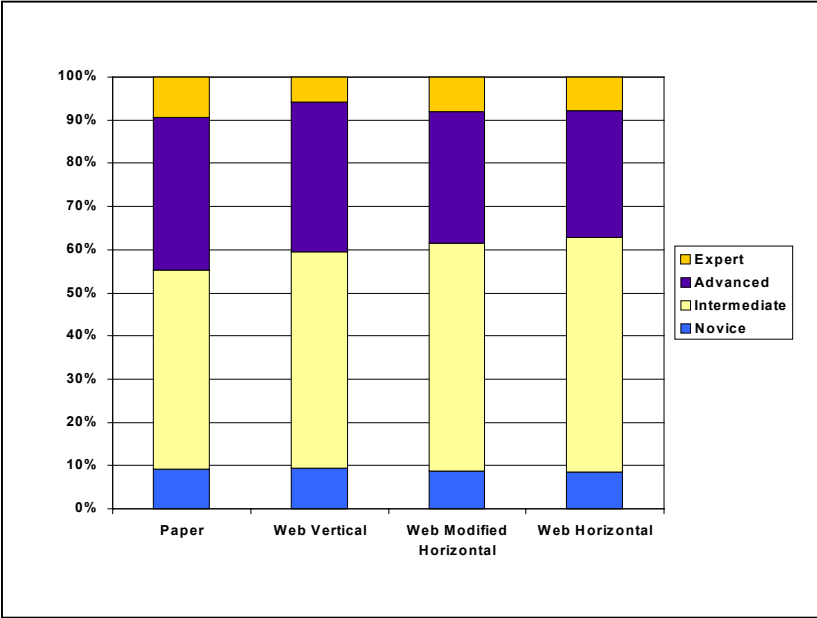


On this last demographic variable, we found no statistical differences as well, indicating that the various modes represented the total population similarly with respect to demographic composition.

Responses to behavioral and attitudinal questions

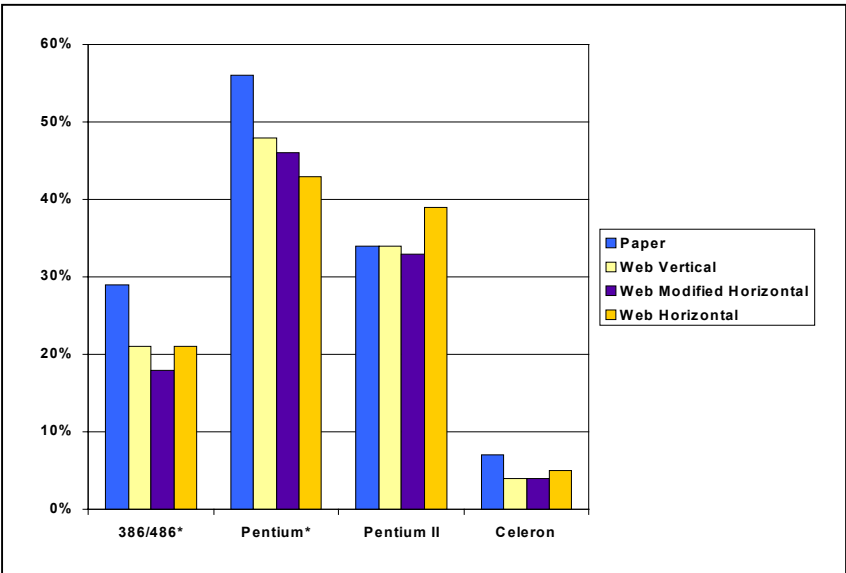
While the audiences were similar demographically, we wanted to evaluate answers to other types of questions, such as behavioral or attitudinal questions. The first one we looked at was a respondent's computer expertise.

Computer Expertise



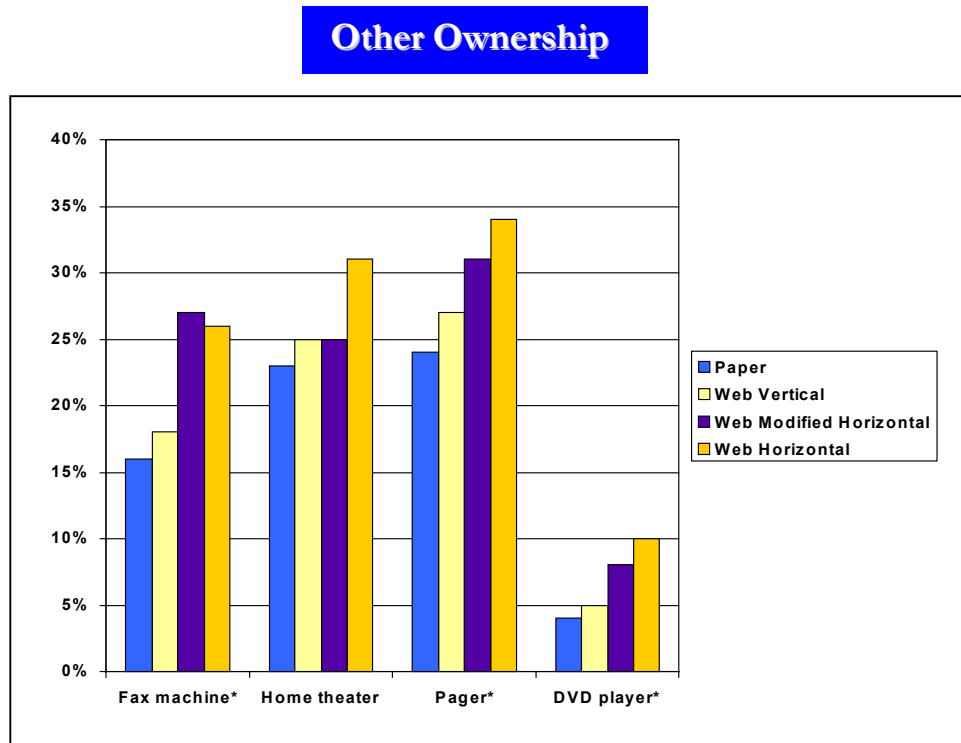
On this first variable we found no statistically significant differences in the responses across modes, so we then looked at PC ownership, and the results are shown below.

PC Ownership



In PC ownership we see the first true difference in the representation of the groups across modalities. Paper surveys have a higher representation among the lower (older) type PCs, while the web horizontal is markedly higher for top-end machines.

To explore other areas where this phenomenon might occur, we show below ownership of other household electronic devices.



Here the differences for three out of the four technologies is also found to be significantly different across survey methods. As with PC ownership, we see that the paper respondents represent the less tech-intensive segment of the population. Additionally, we see that between web methods, we have the most tech-intensive showing up in the group that was willing to complete the entire horizontal version, which was the most tedious and time-intensive to complete.

Readership levels as measured by each method

One of the primary disadvantages of a paper-and-pencil questionnaire is that respondents are allowed to provide incomplete or inconsistent information. This characteristic required a number of editing procedures be employed prior to tabulating the readership data. Given the greater control provided using online error prevention, none of these editing rules were applied to the three web versions.

The first editing rule eliminated respondents who provided exaggerated readership claims. Any respondent identified as reading all or nearly all 94 publications was considered not to have returned a valid questionnaire, and removed from the base. In most cases, this behavior seemed indicative of confused respondents who provided reading frequencies for all publications

regardless of whether or not they read them. This confusion only occurred with the paper format, as no web respondents exhibited exaggerated readership claims. Nine paper questionnaires (in addition to the 376 reported in-tab) were excluded for this reason.

For each remaining respondent, the editing rules for the paper questionnaire allowed for three mutually-exclusive contingencies:

1. Any title left completely blank (i.e., no screening or follow-up information was provided) was considered not to be read by that respondent in the past six months. The screen-in response was set to “no” for that publication.
2. Any title whose screening question was either answered “no” or left blank but had an answer for the frequency question was re-categorized as having screened-in.
3. Any title which screened-in but did not have an answer to the frequency question had the answer to the frequency question ascribed. The ascription was done in such a way that the ascribed frequency distribution was the same as was the distribution actually reported by those individuals who completed the question for each title.

The extent of such editing is presented in the table below, which may be read as follows: 376 respondents were supposed to answer questions about 94 publications resulting in a gross number of 35,344 questioning occasions ($94 * 376 = 35,344$). On 3,115 occasions (8.8% of all possible occurrences) the respondent left a title’s row on the questionnaire completely blank. In this instance, the respondent was assumed to be a non-reader with a reading frequency of zero out of four.

Necessary Paper Edits

	Gross Number of Titles Edited (Number)	(% of 35,344)*	Edits Per Respondent **
Blank rows set to non-screener	3,115	8.8%	8.3
Screener changed to “yes”	93	0.3%	0.2
Reading frequency ascribed	115	0.3%	0.3
Total Necessary Edits	3,323	9.4%	8.8

*(94 titles) (376 respondents) = 35,344
**Edits per respondent = total of edits / 376

The right most column of the table above takes the number of editing occasions and divides them by the number of respondents returning a questionnaire (in this case 376) to produce the average number of edits per respondent. The average questionnaire contained completely blank rows for 8.3 titles. The second and third rows of the table are interpreted the same way. When the rows are summed we see that 9.4% of the questioning occasions required some form of editing, or 8.8 edits per respondent on average.

Readership Comparisons

The screen-in data generated by all four versions are shown in the next table. These results indicate that, consistent with theory, the vertical web version produced the highest mean screen-ins per respondent (12.2) and the modified horizontal web version produced the lowest (7.9). The paper version and the modified horizontal web version produced mean screen-in levels that were statistically indistinguishable from one another (8.6 vs. 7.9).

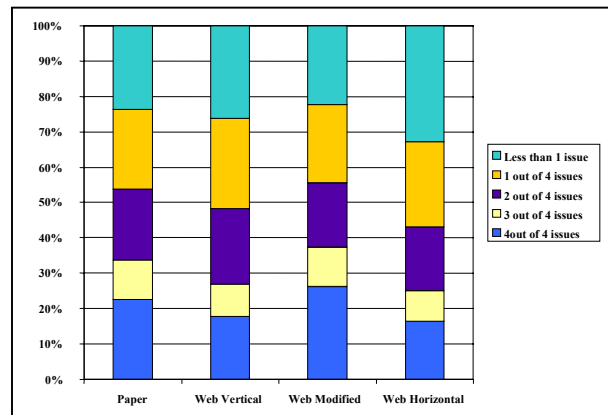
Average Number of Screen-Ins Per Respondent

	Paper Horizontal	Web Horizontal	Web Mod. Horizontal	Web Vertical
(Base)	(376)	(258)	(273)	(329)
Mean	8.6**	10.0*	7.9**	12.2*
Standard Deviation	6.8	7.5	6.5	7.4

*Significantly different from all other versions at the 95% confidence level
 **Significantly different from horizontal web and vertical web versions at the 95% confidence level

The horizontal paper and modified horizontal web questionnaires not only produced a similar number of screen-ins, but these readers look similar to one another in terms of reading frequency as well. The table below contains the distribution of reading frequency among all screen-ins. The base for these estimates is the total number of respondents multiplied by the average number of screen-ins per respondent (from previous table).

Distribution of Reading Frequency



The pattern of frequency responses follows an expected pattern. The vertical web version – the version with the highest number of average screen-ins – produced the screen-ins with the lowest reading frequency. Conversely, the most frequent readers are those from the modified horizontal web version, which also registered the fewest average screen-ins per respondent. The paper and modified horizontal web versions have comparable reading frequency distributions that are congruent with their similar screen-in levels.

In order to produce average issue audience data, each respondent was assigned a probability of reading the average issue of each title using the following probabilities associated with each frequency claim:

Frequency Claim	Probability
4 out of 4 issues	1.00
3 out of 4 issues	0.75
2 out of 4 issues	0.50
1 out of 4 issues	0.25
Less than one	0.10
Non-screeners	0.00

For each of the 94 publications, for each questionnaire version, these probabilities were summed and divided by the sample size to produce an estimated rating (or coverage percentage). The coverage percentages for each of the three web-based versions were then indexed using the coverage percentage for the paper-based version as the base.

The vertical web version – the version that produced the highest web version response rate – produced ratings which, on average, were 49% higher than were the paper-based ratings. The horizontal and modified horizontal web ratings were, on average, 38% and 22% higher than the horizontal paper version, respectively. Depending upon which web version is employed, the web estimates are between 22% and 49% larger than these provided by the horizontal paper version. These mean indices are shown here.

Mean Audience Ratings Indexed To Paper

	Paper Horizontal	Web Horizontal	Web Mod. Horizontal	Web Vertical
(Publication Base)	(94)	(94)	(94)	(94)
Mean Index Across Titles	100*	138	122	149
Standard Deviation	N/A**	185	132	112

*Index of 100 is based on an average coverage of 4.4% per magazine
 **All paper titles are indexed to 100, thus producing no variation in this estimate

If the 94 indices that comprise these means can be regarded as independent observations, they are all highly statistically significant (the t-ratios range from 12.8 to 7.2). The substantial elevation of indices across all three web versions demonstrates that while aggregate differences in screen-ins and reading frequency may have been marginal, the impact on individual publications was distributed unevenly. The table below restates the indices shown in the previous table within two divisions of publications grouped by audience coverage.

Mean Audience Ratings Indexed to Paper

	Paper Horizontal	Web Horizontal	Web Mod. Horizontal	Web Vertical
47 Larger Titles	100*	98	92	114
47 Smaller Titles	100**	179	152	186

*Index of 100 is based on an average coverage of 7.3% per magazine

**Index of 100 is based on an average coverage of 1.5% per magazine

Among the most widely read publications, the three web versions produced estimates that were most comparable to those produced by the paper version, ranging from an 8% decline to a 14% increase on average. Readership levels were substantially higher among smaller publications. The horizontal web version eventually produces the highest claimed readership among the smallest titles, explaining why its standard deviation is the highest of the three web versions.

Despite this higher claimed readership of smaller publications by the web-based questionnaires, there is a strong relationship among audience levels produced by these three web-based versions with the levels produced by the paper questionnaire now in use. This final table shows the product-moment correlation of the audience levels produced by the three web versions with the audience levels produced by the paper horizontal version. All of the correlation estimates are uniformly very high, ranging from +.95 to +.97.

Correlation of Audience Estimates

	Paper Horizontal	Web Horizontal	Web Mod. Horizontal	Web Vertical
Paper Horizontal	1.0			
Web Horizontal	.95	1.00		
Web Mod. Horizontal	.97	.97	1.00	
Web Vertical	.95	.95	.95	1.00

CONCLUSION

Our overall objective was to evaluate the impact of migrating an existing tracking study to the web. This is a multi-faceted issue, and this section discusses our conclusions with regard to each area.

General Conclusions

Within the group of web-enabled respondents we surveyed, we found that we can successfully obtain similar response rates and media readership levels using the web as compared to paper. Finding all test cells to be demographically similar is encouraging. However, the findings with respect to technology ownership indicate that nuances in the look and feel of the graphical user interface (GUI) or screen layout can impact not only response rate, but also skews the sample of completed interviews to a more or less tech-savvy crowd depending on the specific execution. This is consistent with the findings of other published work in this area (Findlater and Kottler, 1998).

Additionally, web administration of a media questionnaire offers numerous benefits, including interview standardization, controlled skip patterns, and consistent data entry. Since nearly 10% of all paper responses require some form of editing after the fact, the web offers a substantial reduction in back-end processing and a potential increase in response accuracy.

Measuring magazine audiences over the web, however, has two additional consequences. First, web surveys result in a material reduction in response rates, both from the loss of respondents from invalid email addresses and an increased failure to complete the self-administered questionnaire. Further training of phone interviewers and greater exposure to email formats and standards among the general public will ameliorate the screening problem over time, but it is doubtful that it will ever be fully eliminated. As for responses to the web and paper questionnaires, further experimentation with contact methods and incentives will need to be done to determine the intransigence of the disparity between the two modalities.

The second consequence concerns a material increase in average issue audience levels. In the present study the increase ranged from 49% for the vertical web version – the most theoretically defensible format which also had the highest response rate of the three web versions studied – to 22% for the modified horizontal web version which permitted respondents to screen-in comparably to the paper version. These increases were largely exhibited among the publications with the lowest audience coverage estimates in all three web versions.

Since all three web versions shared the characteristic of only displaying 7 titles per screen rather than the 21 shown on each page of the paper questionnaire, perhaps asking respondents to focus on fewer titles increases audience estimates more so than any difference between vertical or horizontal approaches. To the extent that this is driven by space limitations of computer presentation (at least until 21” monitors are the norm), it seems unlikely that any web-based approaches to readership estimation will completely overcome this. Further analysis into this issue should reveal more insight into the impact of web formatting.

From this study, we have learned a great deal about moving a paper questionnaire to the web. One of our most consistent and important findings was that it is virtually impossible to replicate a survey from one format to the other. Differences between paper and web questionnaires can never truly be isolated to the modality itself, since numerous divergences dictated by each

approach appear along the way to confound the issue. Since response patterns are affected by uncontrollable factors such as by respondent familiarity with the web medium (Jeavons, 1999), we must face the fact that web and paper surveys will almost always produce different results, no matter how much care is taken to make the two appear similar. Thus the requirements to adopting the web as a survey modality are thoroughly understanding all the forces that drive these differences and creating comfort and understanding among the industries that produce and use such data.

Thus, our finding is that it doesn't make sense to simply attempt to replicate previous surveys on the web. This brave new web-based world requires zero-based design to minimize respondent burden, and to take full advantage of what the web has to offer. We believe the web has a rich environment in which to collect information, but to truly optimize it, survey designs will need to be adapted to the native web format (Witt, 1997) with an openness to a break with traditional paper survey design.

REFERENCES

- Appel, V. and Pinnell, J. (1995). How Computerized Interviewing Eliminates the Screen-In Bias of Follow-Up Questions. *Proceedings of the Worldwide Readership Research Symposium 7*, p. 177.
- Bain, J., Arpin, D. and Appel, V. (1995). Using CASI-Audio (Computer-Assisted Self Interview with Audio) in Readership Measurements. *Proceedings of the Worldwide Readership Research Symposium 8*, p. 21.
- Bain, J., Frankel, M. and Arpin, D. (1997). Audio Visual Computer Assisted Self Interviewing (AV-CASI): A Progress Report. *Proceedings of the Worldwide Readership Research Symposium 8*, p. 437.
- Findlater, A. and Kottler, R. (1998). Web Interviewing: Validating the Application of Web Interviewing Using a Comparative Study on the Telephone. *Proceedings of the ESOMAR Worldwide Internet Seminar and Exhibition*.
- Gates, R. and Heloton, A. (1998). The Newest Mousetrap: What Does It Catch? *Proceedings of the ESOMAR Worldwide Internet Seminar and Exhibition*.
- Jeavons, A. (1999). Ethology and the Web: Observing Respondent Behaviour in Web Surveys. *Proceedings of Net Effects: The ESOMAR Worldwide Internet Conference*, p. 201.
- Thompson, M. (August 23, 1999). When Marketing Research Turns Into Marketing. *The Industry Standard*.
- Witt, K. (1997) Best Practices in Interviewing Via the Internet. *Sawtooth Software Conference Proceedings*, p. 15.

COMMENT ON WITT

Keith Sentis
Pathfinder Strategies

I enjoyed reading Karlan's paper for several reasons. First, because it addresses such an important question regarding the shift from paper and pencil to on-line data collection methods. I am sure that anyone who either has considered moving to on-line methods or has already done so, would have grappled with some of the issues that Karlan addresses in this study.

The second reason that I enjoyed the paper is that the findings were quite intriguing. In particular, the differential response rates caught my attention. Since response rates are highly relevant to everyone who might be moving studies to the Web, I will focus my comments on this set of results.

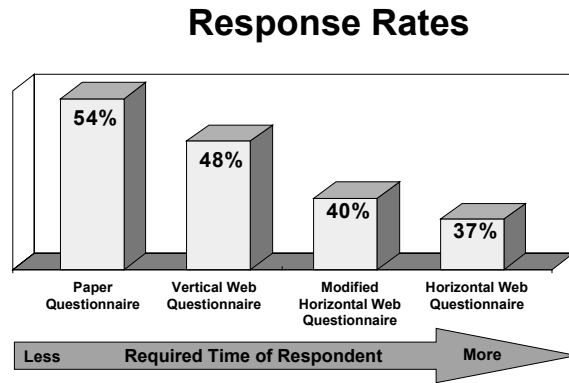
Recall that the response rates in the four cells looked like this:

- paper horizontal 54%
- Web vertical 48%
- Web mod horizontal 40%
- Web horizontal 37%

In the paper, the four cells are described in terms of the Time Investment that is required of respondents to complete the questionnaire. I do not know how this Time Investment measure was obtained. I would guess that the timing data were collected during the on-line session in the Web cells and perhaps some pilot work was done to measure the time to complete the paper version. In any event, Karlan reports that the four cells are ordered according to the time required as follows:

- the paper cell required the least time
- the vertical Web required more time
- the modified horizontal Web required still more time
- the horizontal Web required the most time

In the chart below, the responses rates plotted against the time required to complete the interview.



As illustrated in this chart, the results for response rates in the four cells can be restated as:

- with the same incentive, response rates go down as more of respondents' time is required to complete the interview

These findings are quite consistent with what we know about respondent incentives. The incentive is an integral part of an implied “contract” between the researcher and the respondent. This contract governs the exchange of value between the respondent and the researcher. The respondent gives up his or her time to provide valuable data to the researcher. In return, the researcher gives something that is ostensibly of value to the respondent.

This notion of a contract between the researcher and the respondent led me to wonder if Karlan has examined the response rate for the three Web cells as a function of length of interview. The contract notion would predict that the response rates would be roughly equivalent across the three Web cells for any given interview length. It is just that respondents drop out when their sense of having fulfilled the contract has been reached. Thus, any differences in the overall response rates between Web cells would be due to the longer time required to complete the horizontal versions.

The contract with respondents also prompted me to consider, in particular, the higher response rate for the paper modality. Karlan found that the paper cell had a substantially higher response rate than the Web cells. I have already offered one explanation for this result, namely, that the paper cell required the least amount of respondents' time. However, this paper cell also differed from the Web cells in another way that possibly could have influenced the response rates. Recall that the paper respondents received the questionnaire at the same time as they received the incentive. In the paper cell, the incentive was juxtaposed with the task. That is, the questionnaire was in the respondents' hands at the same time as they received the payment. This context may have served to reinforce the implied contract:

- “they are paying me these two dollars to complete this questionnaire”

This reinforcing context was not as strong for the Web cells because the payment was disconnected in time from the respondent's task.

One way of testing this idea would be to examine the response rates by modality after each of the three contacts with respondents. If this "enhanced contract" hypothesis is correct, then the paper cell should have exhibited a higher response rate to the initial contact. If this turns out to be the case, then we might consider ways of "enhancing the contract" for Web-based surveys. I would expect that Karlan has already looked at this and I would be curious about the findings.

Another intriguing finding was that the low-tech respondents were more likely to complete the paper questionnaire than their high-tech counterparts. I wonder how much of this was due to the overall quality of the "Web experiences" of the low-tech folks. If respondents with older processors and less technology also have slower modems and less broadband access to the Web, then their Web experience would have been less favourable than that of their high-tech fellow respondents. The low-tech respondents' reaction to the implied contract might have been something like this:

- "they are paying me these two dollars to have another bad Web experience"

I wonder if, as part of the inventory of respondents' equipment, the survey got down to details like modem speed. If so, that would be interesting to check this explanation of the low-tech bias towards the paper cell.

Related to this, perhaps the low-tech folks made an attempt to do the Web surveys but bailed out because of the frustration with the experience. I wonder if information about the respondents' Web access equipment had been collected during the telephone screening of respondents. If so, it might be possible to determine if the low-techies actually did start the Web-based interviews but dropped out at a higher rate than those respondents with faster Web access.

So in summary, Karlan has made an important contribution to our understanding of some key issues regarding Web-based surveys. I thank her for sharing this learning with us and look forward to further work on these issues.

TROUBLE WITH CONJOINT ANALYSIS IN INTERNATIONAL INDUSTRIAL MARKETS

Stefan Binner
bms GmbH

INTRODUCTION

bms is a full service market research company based in Munich, Germany and Zug, Switzerland. Our focus is international industrial market research. Among our clients are multinational Blue Chip corporations as well as middle size high-tech companies.

International industrial marketing research is a comparatively small sector compared to consumer goods or opinion research. Real industrial research represents approximately 5% of budgets spent for marketing research in Europe.

Many industrial research firms are specialised on a specific sector e.g. chemical or semiconductor, some others, like *bms*, offer their services to a variety of markets.

Compared to consumer goods, the market researcher finds different circumstances in international industrial markets. Here some examples:

- small and fragmented market structures
- different marketing channels in different countries
- niche markets
- commodity markets as well as fast developing high tech industries
- many technical driven companies (inside out)
- often marketing function comparatively new and in competition with sales

Budgets for marketing research are usually very limited and this, in turn, limits the possibilities of research designs. The critical issue is always whether the client understands the value added by marketing research or not.

Typical studies in international industrial markets are

- Customer satisfaction
- New product opportunity
- Competitive intelligence
- Market description

In the past strategically important issues such as new product development and pricing were dealt with mainly internally. Management would not share them with external parties (e.g.

market researchers). The decisions would be based on own market understanding and gut feeling.

With increasing international competition and increasing importance of marketing the industrial companies have become more open to outside specialists such as market researchers and their services.

Conjoint Analysis is one of those comparatively new methods which were hitherto often unknown among marketing management of industrial companies.

Also in industrial markets Conjoint Analysis is mainly used for

- Research & Development
- New product development
- Service
- Pricing

bms - Conjoint Survey:

In the last months of 1999 *bms* conducted a small survey among users of Conjoint Analysis in industrial or business to business markets. The author will refer to this survey which includes the answers of 37 research professionals all over the world in the course this paper.

- 50% of the research professionals which participated in our survey are not always able to convince their clients to use Conjoint Analysis.
- Client's understanding the method and the costs of conjoint studies are the strongest hurdles for the usage of conjoint.

What limits the use of Conjoint Analysis in industrial marketing research?

SELLING CONJOINT ANALYSIS

Quite often the first challenge is to explain (teach) to the client the theory and principle of conjoint. This is very critical. Sometimes the client claims to know all about it and in the course of the project one recognises that she or he has wrong expectations.

Sometimes we co-operate with clients which have no market research experience and knowledge. There have been a lot of proposals on how to explain Conjoint Analysis within a few minutes. However, the question still remains as to whether the research user really has understood the method or not.

Additionally the vendor has to prove the value added by investing into a conjoint analysis study and trying to persuade the client to reserve enough resources (time and money) for it.

THE COMPLEXITY OF THE MARKET

The information one needs to establish an attribute value system is often not available without further investigation. In contrast to many consumer markets in which almost every individual has experiences with certain products (toothpaste, cars, beer) the attributes of industrial products are not always intuitive and are often very complex.

A systematic collection of market information is often not available or, even worse, such information is seen as secret or sensitive and the client does not like to share it with the research agency.

In industrial markets we find quite often circumstances which make the understanding of these markets difficult. We have to deal with complex decision making, distribution channels, hidden agendas, corporate strategies etc. Two companies can be competitors in one market and could have a joint venture in another market.

Another example to illustrate the complexity in these markets is price. Which price are we talking about? Purchase wholesale price? Resale wholesale price? Purchase retail? Are value added or sales taxes included? Are discounts and margins included? Which volumes are we talking about in the interview?

The international aspect of these industrial markets adds even more complication such as cultural differences (also in corporate culture), different laws, market mechanisms etc. Gaining support in the client's regional companies can be a critical success factor. Often the local management sees the research commissioned by the headquarters as a threat rather than an opportunity.

SPECIFIC PROBLEMS

Once we have convinced and taught the client about CA as necessary, and have got the market understanding and support from local management there are still further issues which can sometimes eliminate Conjoint Analysis from the list of possible methods.

Heterogeneity:

Industrial markets are often very fragmented. Companies come from completely different industries, use the product in a different way and have a different value system.

Sample Sizes:

Some markets are very specialised. As there are not many users of specific products or services (e.g. oil drilling platforms, desalination plants) the sample size of studies are usually small. This may even be the case if one includes all relevant persons in each target customer company. It is very important to sample a high proportion of the potential customers.

Decision making process:

One would think that industrial purchasers are more rational based and know their own value system. This is not always true. In many industrial markets we find emotional aspects as well as rational aspects. In addition, influences may include purchase guidelines or decisions from senior management. To understand the 'Decision Map' in the companies is a real challenge. More than

once we decided to investigate this subject and postponed the planned conjoint study as we found it impossible to understand the decision making process.

Limitation to a certain number of attributes:

Industrial products are often more complex than consumer goods. The list of relevant attributes is often very long. Sometimes we have values which would need a different set of attributes for different sets of customers.

bms - Conjoint Survey:

A majority of respondents experienced problems and limitations of Conjoint Analysis in their work. The problems are:

- Limited number of attributes (33,3%)
- Unrealistic (18,5%)
- Length of questionnaire (14,8%)
- Client understanding (14,8%)
- Accuracy of predictions (11,1%)
- Paper & pencil approach (7,4%)
- Questionable data quality (7,4%)
- Just rational attributes (7,4%)
- Dependent attributes (3,7%)
- Fixed attribute sets (3,7%)
-

THE ANALYSIS PROCESS

“What are UTILITY VALUES?” - Every researcher who has heard his client utter this question during the research process knows that she or he is facing a challenge. Did the client not understand the explanation at the beginning? It is the critical part of the analysis. If they do not understand the logic behind this they might not understand and believe the rest.

Once one gets to the point of establishing market models and scenarios, one needs information on competitors’ products such as specifications according the attribute / level system or prices. Especially in industrial markets this kind of competitive intelligence is not always available. Sometimes the researcher has to go back to the market and investigate. In the author’s experience it is wise to ask this information in the course of the conjoint interview. More than once we were able to prove wrong understanding of the client’s competitive perception.

Commissioning a conjoint study means the client has to team up with the researcher, not only for the set-up and development but also for the analysis and modelling. The researcher needs the input from the client e.g. what should we simulate?

Industrial companies are often very lean on marketing functions. Marketing or product management is often not prepared to input ideas and time into the simulation. The more they do the higher the value of result they get from this exercise.

bms - Conjoint Survey:

“...explaining the results to clients is a major challenge”

“It demands clear thinking and a great deal of involvement and communication with the client if the results are to be accepted and used well”

“CJ is also seen as a black box by clients who have no statistical background. They may not trust the results because they do not understand them”

THE THREE STEPS

A successful conjoint project in an international industrial market always consists of three steps:

- I. Preparatory
- II. Field Work
- III. Analysis

When asked about the effort and cost of a conjoint project the client is often surprised that these will be almost equally distributed among the three steps. (This is our experience and, of course, depends on the sample size.)

Step 1 Preparatory

Teaching the client is a challenge. The research user's knowledge about the method is a key to success. Whenever possible we present case studies to explain the method and, equally importantly, to prove the value added and power of the results.

The logical flow of such a presentation is

- Understanding Conjoint
- Examples of questionnaires and case studies
- Proposal of Conjoint Analysis for the client's specific research problem

Analysing the market is a must unless there is knowledge from prior studies or real comprehensive and systematic market information on the user side. In industrial and international markets this is often not the case. Investment in a pre-study internal (e.g. sales, local management) and external (customers, trade, insiders) pays back in the quality and added value of the conjoint project.

It is usually very difficult to raise a budget for Step I of the research. The common opinion is that there is enough in-house knowledge and market expertise to create the conjoint model.

Our experience proves that both lack of conjoint understanding and market analysis could lead to a complete failure of the method. This happened e.g. when we used too technical attributes such as combinations of several physical values. The respondents did not understand the attribute and ignored it.

Another example was in a market where the interviewed target group determined the specification but did not pay for the product and, to make things worse, received a bonus at the end of the year based on the purchase/specification volume. The result was negative price elasticity.

Developing the conjoint model. i.e. definition of the attributes and their levels requires agency's and client's efforts in order to get it right. Pre-testing will show whether the model is realistic. Internal "Guinea Pig" interviews are helpful but cannot replace real pre-test interviews. An engineer at our client's company trying to simulate the behaviour of a critical customer is not very productive in the process. In one case we had the problem that these internal interviews always ran over the time limit we had set for the interview. Later in the conjoint field work, the interview time was half that experienced in the internal simulation.

Step 2: Field work

Field work in industrial markets is different from consumer markets.

Respondents are often not computer literate, they may be spread all over the country or region and difficult to motivate. One needs stable and long lasting notebook computers doing interviews in plants or on construction sites. Interviews often gets interrupted by the ongoing business of the respondent. It is important to lay out a scenario i.e. explaining the situation we are trying to simulate, the attributes and their dimensions.

It can be strongly recommended to add conventional questions to support the explanation of the results. Also information about competitive product information should be included.

International industrial interviews require high level interviewers. However, we recommend that the first interviews should be conducted by the project manager in order to feel if the interview gets to the point.

Step 3: Analysis

If it is possible to create a team of researchers and users, the analysis is of course the most interesting and rewarding part. This team provides shared knowledge and information. At this stage the researcher will learn if the client knows his market and competitors, has understood the method and was open with the objectives of the project.

In a Conjoint Analysis we did with a client in the area of building control we were able to co-operate with the client to an extent that we shared knowledge of production costs, cross company prices etc. This information taken into the model allowed the calculation of the sales price producing the highest Pan-European total profit. This price was retained for a product they later introduced into the market and which became a big commercial success.

THE FUTURE

bms - Conjoint Survey:

Satisfaction with Conjoint Analysis

“B-to-b conjoint studies are generally more difficult than consumer conjoint studies but for us are generally just as productive”

“I think it is actually better for industrial/b-t-b applications than consumer goods”

“By itself it would be low. In conjunction with other methods: 8/10”

“mathematically results are stunning, psychologically often disappointing”

Although the title of this paper is “trouble with conjoint ...” we predict that Conjoint Analysis will become more important in the future of international industrial market research.

One reason will be that the marketing management of industrial companies will be more and more familiar with this technique, will have the necessary education and skills to fulfil their part of the co-operation.

Successful Conjoint Analysis will have a track record. Multimedia and to an extent the internet will increase the number of possible applications.

Of course, we expect also developments from the suppliers of Conjoint Analysis packages.

bms - Conjoint Survey:

Future Conjoint Analysis packages must be:

- easy to use
- have an internet integration
- be Windows based

Further expectations are better simulation and multimedia capabilities.

Although the many challenges will remain as they are given by the nature of the market Conjoint Analysis will become a standard tool in international industrial markets.

VALIDITY AND RELIABILITY OF ONLINE CONJOINT ANALYSIS

*Torsten Melles, Ralf Laumann, and Heinz Holling
Westfaelische Wilhelms-Universitaet Muenster*

ABSTRACT

Using conjoint analysis in online surveys is gaining growing interest in market research. Unfortunately there are only few studies that are dealing with the implementing of conjoint analysis in the World Wide Web (WWW). Little is known about specific problems, validity, and reliability using online measures for conjoint analysis.

We conducted an online conjoint analysis using a fixed design of thirty paired comparisons. A traditional computerized conjoint analysis was conducted in the same way. Several criteria were used to assess reliability and validity of both data collection methods.

The results show that data drawn from an Internet conjoint analysis seem to be somewhat lower in reliability (internal consistency) compared to traditional computerized conjoint analysis. Nevertheless, the reliability seems to be sufficient even in the case of its online form. Regarding predictive validity, both data collection methods lead to comparable results. There is no evidence that the number of thirty paired comparisons might be too high in the case of Internet conjoint analysis. More paired comparisons seem to be favorable taking the moderate internal consistency of responses into account and the additional possibilities of reliability testing.

BACKGROUND AND INTRODUCTION

After three decades using conjoint analysis there is still a growing interest for choosing this method to analyse preferences and predict choices in marketing research and related fields (Cattin and Wittink, 1982; Wittink and Cattin, 1989; Wittink, Vriens and Burhenne, 1994; Melles and Holling, 1998; Voeth, 1999). The contributions of many researchers have led to a diversification of methods for stimulus-construction, scaling, part-worth estimation, data aggregation and collecting judgments from the subjects. The following paper will focus on techniques for collecting judgments that can be used in conjoint analysis and on an empirical examination of the reliability and validity of the newly introduced method of online conjoint analysis, conducted over the World Wide Web (WWW). Little is known about the quality of data generated by an online conjoint analysis. Is the WWW an appropriate place for collecting complex judgments? Is the quality of this method comparable to that of other collection techniques?

Several techniques using different media have been proposed to collect multiattribute judgments. Up to the mid 80s conjoint analysis was nearly exclusively done by paper-and-pencil-tasks in the laboratory or by traditional mail surveys. The introduction of ACA has led to a radical change. Today, the method most often used for conjoint is the computeraided personal interview (CAPI).

The methods used in conjoint analysis can be categorized according to three dimensions (Table 1). This categorization is a simple attempt of aligning the methods. It is neither

comprehensive nor are the categories sharply distinct. Collection methods using configural stimuli are not listed as well as mixtures of different procedures like the telephone-mail-telephone (TMT) technique. Each one of the methods displays a specific situation for making judgments. Therefore, it cannot be expected that these methods are equivalent. In a traditional mail survey, for example, the questionnaire is done by paper and pencil without an interviewer present to control or help the subject. In the case of a computeraided interview the stimuli are shown on a screen and an interviewer is present. This facilitates a higher level of control and help. Problems can arise in cases where interviewer biases have to be expected.

Table 1: Methods of collecting multiattributive judgments in conjoint analysis. Visual methods use verbal descriptions or pictures.

		computeraided	non-computeraided
personal	Visual	computeraided personal interview (CAPI)	personal paper-and-pencil-task
	Acoustic	(personal interview)	
non-personal	Visual	disk-by-mail (DBM), online-interview	traditional mail survey
	Acoustic	computeraided telephone-interview (CATI)	telephone-interview

Comparisons have been made between traditional mail surveys, telephone interviews, personal paper-and-pencil tasks (full-profile-conjoint) and ACA, the most used computeraided personal interview method (e.g. Akaah, 1991; Chrzan and Grisaffe, 1992; Finkbeiner and Platz, 1986; Huber, Wittink, Fiedler, and Miller, 1993). It is very difficult to draw conclusions from these comparisons because of many factors confounded, favouring one method against the other. For instance, ACA uses a specific adaptive design and a specific scaling of judgments and estimation procedure. So differences to part-worths gained from mail-surveys can arise from each of these characteristics or their interaction as well as from the specific data collection method. Apart from this limitation, personal paper-and-pencil task and ACA can be viewed as nearly equivalent in reliability and validity. Traditional mail surveys and telephone interviews can lead to the same level of accuracy. However, this depends on several characteristics of the target population and it is only suitable with a low number of parameters (six attributes or fewer).

Using the Internet for conjoint analysis receives growing interest, especially in marketing research (Saltzman and MacElroy, 1999). Nevertheless, little is known about problems arising from the application of conjoint analysis over the Internet and the quality of this data. Only few studies are dealing with online conjoint analysis. These exceptions are studies published by Dahan and Srinivasan (1998), Foytik (1999), Gordon and De Lima-Turner (1997), Johnson, Leone, and Fiedler (1999), Meyer (1998), Orme and King (1998).

Meyer (1998) observed that the predictive validity of his online conjoint analysis (using a full-profile rating task) was much better than random generated estimations. But there is no way to compare it with data gained from other collection methods. Orme and King (1998) tested the

quality of their data using a holdout task (first choice). They found single concept judgments to perform as well as graded paired comparisons. The stimuli were full-profiles consisting of four attributes. Orme and King (1998) emphasize on the common features of Internet surveys and traditional computerized surveys. Only Foytik (1999) compared the Internet to other data collection methods in conjoint analysis. Drawing from several studies he reports higher internal consistency measured by Cronbach's Alpha and the Guttman Split-Half test of the Internet responses compared to traditional mail responses as well as more accurately predicted holdout choices.

At this point there are some unresolved questions regarding the quality and specific features of Internet conjoint analysis. Some of these questions are:

- How many judgments should / can be made?
- Is the predictive validity and reliability comparable to traditional computerized surveys?
- Which ways can be effective in handling problems of respondents' drop-out during the interview and "bad data"?

METHOD

The research questions were tested by conducting a conjoint analysis of call-by-call-preferences. Resulting from a liberalization of the German telephone market there are several suppliers that offer one single telephone call without binding the consumer. This call-by-call use is possible through dialing a five-digit supplier-specific number before the regular number. A choice is made each time before a call to be made. Call-by-call services vary between weekdays and weekends, in the time of day, and the target of the call. There is no supplier dominating the others in general.

The selection of attributes and levels based on results of earlier studies, expert interviews and a pilot study. The attributes should have been relevant at the moment the decision between different suppliers is made and the levels should have been realistic. Having this criteria in mind, four attributes (price per minute, possibility to get through, interval of price cumulation, extras) were chosen. Two of them had three levels, the other two had two.

Subjects were users of call-by-call-services who visited the internet site <http://www.billiger-telefonieren.de> and decided to participate in the study. This was done by 9226 respondents during the two week period the conjoint survey was available on the website. In order to elicit their true preferences that are related to choice, subjects were asked to evaluate services that were relevant to them and that were adapted to their telephoning habits. If one is calling mainly weekdays between 7 and 9 pm to a distant target in Germany, he was asked to judge the services in front of this situation. So it is possible to distinguish different groups of users, and it is assured that the subjects are able to fulfill the task.

Judgments were made by a graded paired comparison task. As in ACA no full-profiles were used. Due to cognitive constraints the number of attributes was limited to three (e.g. Agarwal, 1989; Huber and Hansen, 1986; Reiners, Jütting, Melles, and Holling, 1996).

Each subject has been given 30 paired comparisons. This number provides a sufficiently accurate estimation of part-worths given a design with fewer than six attributes and three levels

on each attribute. Reiners (1996) demonstrated for computeraided personal interviews that even more than 30 paired comparisons can lead to slightly more reliable and valid part-worths. Additionally, results from other studies show that one can include more questions in personal interviews than in non-personal interviews (e.g. Auty, 1995). Due to these differences of online testing to personal interviews - that may cause a lower level of control and a lower level of respondent motivation - and regarding the length of the whole questionnaire, 30 paired comparisons seemed to be an appropriate number. We chose an approximately efficient design by using a random procedure that selected from various designs that one with minimal determinant of the covariance matrix (Det-criterion). The sequence of paired comparisons was randomized as well as the screenside of the concepts and the position of the different attributes because of possible sequence- and position-effects.

Different precautions were taken to prevent “bad data” caused by a high drop-out rate of respondents:

- functional, simple web-design in order to maximize the speed of the data transfer
- providing an attractive incentive after finishing the questionnaire (participation in a lottery)
- explicitly emphasizing on the fact that the whole interview takes 20 minutes to perform, before the respondent finally decided to participate
- emphasizing on the importance of completely filled in questionnaires.

IP-addresses and responses to personal questions were checked in order to prevent double-counting of respondents. Datasets with identical IP-addresses together with the same responses to personal questions were excluded as well as datasets with identical IP-addresses and missing data to personal questions.

The quality of responses and conjoint-data was measured by multiple criteria:

- Estimating part-worths using an OLS-regression provided with R^2 a measure of internal consistency (goodness of fit). This gives a first indication to the reliability of judgments. But it is necessary to emphasize that the interpretation of this measure can be misleading and must be made carefully. Beside several important problems of this measure there are two specific ones that are related to the distribution of responses. A high R^2 can result from “bad data” (e.g. due to response patterns without any variance) and a low R^2 can result from using only the extremes of the graded scale. For the special case of dichotomous responses and a situation where proportions of success are bounded by $[.2, .8]$ Cox and Wermuth (1992) have shown that the maximum possible value of R^2 is .36.
- Stability of the part-worth estimations on the group level has been measured by intercorrelations between part-worths using each single paired comparison as an input. This means that the responses to the first paired comparison have been taken and aggregated in an estimation on the group level. The same has been done with the second paired comparison and so on. This aggregate estimation was possible as a fixed design has been used and the position of each paired comparison across a high number of respondents has been randomized. Between each pair of estimated part-worth-sets

Pearson r has been calculated and plotted in an intercorrelation matrix. Assuming homogeneity of preference structures the mean correlation of each row, respectively of each paired comparison, is a measure of stability of estimated part-worths. Due to a warm-up-effect¹ and descending motivation together with cognitive strain while performing the task, an inverted u-function is expected.

- A Split-Half test has been performed to test the reliability of responses and part-worths on the individual level. Part-worth-estimations using the first fifteen paired comparisons have been correlated with part-worths derived from the last fifteen paired comparisons. To provide a reliability measure for the whole task the correlation coefficient has been corrected by the Spearman-Brown-Formula. This reliability measure must be interpreted carefully and can only be taken as a heuristic because the reduced design is not efficient taking Det-criterion into account.
- We used a holdout task as a further measure of reliability, respectively internal validity. Due to the difference between this criterion and the task in this case it seems to be more a measure of internal validity than a measure of reliability (see Bateson, Reibstein, and Boulding, 1987, for a discussion on these concepts). Estimated part-worths were used to predict rankings of holdout concepts that were drawn from actual call-by-call-offers made by the suppliers. The name of the supplier was not visible to the subjects. The rankings were drawn from the first choice, second choice and so on between the concepts. The maximum number of concepts that needed to be selected was five. The ranking was correlated using Spearman Rho with the predicted rank order of the same concepts for each individual subject.
- A choice task that asked the respondents to select between different suppliers (concepts not visible) was used to measure external validity. This task was in analogy with the holdout task.

We conducted a computeraided personal interview that was similar to the Internet interview in order to compare the reliability and validity of both conjoint analyses. A student sample (N=32) was asked to indicate their preferences regarding suppliers offering a long distant telephone call at 7 pm weekdays.

RESULTS

The percentage of dropped out respondents over the interview gives a first impression of the quality of measurement. This number is encouragingly low (<15%). The percentage of missing data raised from the first paired comparison task to the last at about 7%.

Stability of aggregated part-worths was tested for each paired comparison. Assuming homogeneity of respondents' preferences, the mean correlation between one set of part-worths with the other sets can be accepted as a measure of stability of the first set. In general, the intercorrelations are nearly perfect indicating a high degree of stability of aggregate preference estimations. As we expected, there is a minimal indication of a warm-up-effect. Stability is rising during the first five paired comparisons. After five paired comparisons its degree persists at a

¹ Several studies have shown that respondents need some trials to adapt to the task.

high level even after thirty paired comparisons. So there is no evident effect of motivational or cognitive constraints that would reduce the stability of aggregated part-worths during a paired comparison task with thirty pairs.

In order to test Split-Half reliability, part-worth-estimations using the first fifteen paired comparisons were correlated with part-worths derived from the last fifteen paired comparisons. The Spearman-Brown-corrected mean correlation (after Fisher-Z-transforming the correlations) was .94 indicating reliable part-worth estimations on the individual level. In contrast, the median R^2 after thirty paired comparisons was insufficiently low ($Md_{R^2} = .31$), indicating a low level of internal consistency. This was also true for the computeraided personal interview, though the median R^2 was slightly higher ($Md_{R^2} = .44$). Split-half reliability was also higher in the case of the personal interviews (.97). There is much room for speculation when looking for a reason R^2 being low since part-worth estimations seem to be reliable. A first step to bring light to the dark is to take a look at the distribution of responses. Doing so there are two striking features: The first is that respondents tend to use extreme categories on the graded scale when judging call-by-call services. There is no difference between the computeraided personal interview and Internet interview. This response behavior might be due to a missing trade-off between the attributes that is typical for decisions without a high involvement (e.g. buying decisions that take little cognitive effort). Strictly speaking this decision behavior is not compatible with the additive rule that is used in conjoint analysis. Since there is no compensation between advantages and disadvantages on different attributes part-worths provide only ordinal information. This has to be kept in mind when predicting choices as in running market simulations. The second feature is a response bias in the Internet interview. The distribution was biased to the right side of the scale. The question if these two features have led to a low R^2 can only be answered finally with running Monte Carlo simulations but it was not possible in the scope of this study.

There is no difference between the methods (online and CAPI) regarding predictive validity measured by the holdout task (Table 2). Both provide nearly accurate predictions. This does not seem to be the case, when predictive validity was measured by a choice between different suppliers. The personal conjoint analysis does better in predicting the ranks than the Internet conjoint analysis. But this result is misleading. Taking into account the differences between the Internet sample and the personal interview sample leads to a conclusion that the methods are equivalent. If only subjects that are younger than 30 years, have a high school diploma (Abitur), and were asked to indicate their preferences for suppliers offering a long distant telephone call at 7 pm weekdays were selected, the coefficient was slightly higher in the case of the Internet conjoint analysis. This small difference might be due to the higher interest of the subjects participating in the Internet study of telephone services.

Table 2: Validity of conjoint analysis conducted through a computeraided personal interview (CAPI) and the Internet.

	CAPI	Internet-Interview	
internal validity (holdout task)	.968 (N=32)	.970 (N=7813)	.977 (N=941)
external validity (choice task)	.539 (N=30)	.412 (N=5663)	.552 (N=691)

Remark: The coefficients in the second column are based on all respondents that participated in the Internet survey. The third column is based on a sample of that survey that is equivalent to the computeraided personal interview.

CONCLUSIONS AND DISCUSSION

The overall conclusion that can be drawn from the results is the general suitability of Internet conjoint analysis to measure preferences in a reliable and valid manner. This statement has to be qualified in several categories. There is a lot of “bad data” resulting from double-counted respondents and response patterns caused by respondents that decide to take a look at the interview but do not participate seriously. Though taking much effort cleaning the data, the reliability of individual level estimates seems to be somewhat lower than in personal interviews. This may be due to the anonymous situation and a lower level of cognitive efforts spent on the task. As in the cases of using conjoint analysis in traditional mail surveys and telephone interviews, the suitability of the Internet conjoint analysis depends on characteristics of the respondents and on the design of the task respectively the number of parameters that must be estimated. The higher the number of parameters, the higher the number of responses that are required for a detailed analysis at the individual level. This again might decrease the motivation to fulfill the task and is a further critical factor for receiving reliable results from an Internet conjoint analysis. Nevertheless, in this particular case reliability was still high even after 30 paired comparisons. Apart from design characteristics reliability and validity of conjoint analysis vary within the characteristics of respondents (Tscheulin and Blaimont, 1993). Up to now, there is no evidence whether this effect might be moderated or not by the data collection method. This could be a further limitation to a broad application of specific data collection methods like telephone or the Internet. Assuming the Internet to be an appropriate medium regarding respondents and design, the following precautions should be taken to assure a high degree of reliability and validity:

- use as many criteria as possible to test the reliability and validity of the conjoint analysis
- use incentives to motivate respondents in giving reliable responses (e.g. giving a feedback of goodness-of-fit)
- encourage respondents to give a feedback and use as much feedback as possible
- IP-addresses and personal data should be controlled for double-counting whenever the participation is based on a self selection of respondents
- analysis on the individual level should precede aggregate analysis to identify “bad data”

Beside the problems of reliability and validity the choice of a data collection method for conjoint analysis is still a practical one. The suitability of Internet conjoint analysis depends on constraints regarding time, money and experience. Preparing an Internet survey needs more time and money than preparing a comparable personal interview. On the other hand, it serves as an advantage in collecting data. There are no interviewers and no notebooks needed to interview the subjects in the case of Internet surveys. Instead of, they will mostly be recruited through the WWW, which can sometimes be easy but as well be expensive and time consuming, if it fails. The opportunities and dangers of the most common options for contacting Internet users are shortly discussed by Foytik (1999). Preparing data for analysis is a much more complicated job in Internet conjoint analysis than in computeraided personal interviews. There is many more “bad data” and identifying it is very time consuming and requires some theoretically guided reflection.

Regardless of the data collection method, we recommend to take a look at the distribution of responses across graded paired comparisons and across each respondent. This is helpful to identify response patterns and simplifying tactics. If, for example, respondents adopt a lexicographic rule, it is not appropriate to assume a compensatory additive one. Ordinary Least Square (OLS) is very robust against such violations but predicting decisions could be better done by a “lexicographic choice model” than BTL or First Choice (both models assume that the choice is made by reflecting all attributes of the objects). Assuming a lexicographic choice means that the object with the maximum utility of the most important attribute will be chosen. If two or more objects have the same utility, the decision is made by taking the next important attribute into account and so on. Moreover, such a rule is more able of covering psychological processes involved in buying decisions that take little cognitive effort. This is often being neglected by researchers using conjoint analysis and running market simulations. Trade-offs are assumed in buying yoghurts, dog food, biscuits, jam, cheese, shampoos, coffee, and so on. Since some early articles from Acito and his coworkers reflections on decision processes when applying conjoint analysis seem to be lost. Predicting and understanding decision making behavior at the market place as well as judgments in conjoint tasks requires some return to the basics. Data collection methods that enhance simplifying tactics cannot be valid in predicting choices that are made through a complex trade-off. Otherwise they might be useful in predicting choices that rely on the same simplifying rules and serve no disadvantage against alternative methods. We found respondents using simplifying tactics (extreme scale categories) in the Internet survey as well as in computeraided personal interviews. The question whether these tactics are more often used in Internet interviews and limiting the suitability of this medium is an issue for further research.

REFERENCES

- Agarwal, M.K. (1989). How many pairs should we use in adaptive conjoint analysis? An empirical analysis. In American Marketing Association (Ed.), *AMA Winter Educators' Conference Proceedings* (pp. 7-11). Chicago: American Marketing Association.
- Akaah, I.P. (1991). Predictive performance of self-explicated, traditional conjoint, and hybrid conjoint models under alternative data collection modes. *Journal of the Academy of Marketing Science*, 19 (4), 309-314.
- Auty, S. (1995). Using conjoint analysis in industrial marketing. The role of judgement. *Industrial Marketing Management*, 24 (3), 191-206.
- Bateson, J.E.G., Reibstein, D.J., and Boulding, W. (1987). Conjoint analysis reliability and validity: A framework for future research. In M.J. Houston (Ed.), *Review of Marketing* (pp. 451-481). Chicago: American Marketing Association.
- Cattin, P. and Wittink, D.R. (1982). Commercial use of conjoint analysis: A survey. *Journal of Marketing*, 46 (Summer), 44-53.
- Chrzan, K. and Grisaffe, D.B. (1992). A comparison of telephone conjoint analysis with full-profile conjoint analysis and adaptive conjoint analysis. In M. Metegrano (Ed.), *1992 Sawtooth Software Conference Proceedings* (pp. 225-242). Sun Valley, ID: Sawtooth Software.
- Cox, D.R. and Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *The American Statistician*, 46 (1), 1-4.
- Dahan, E. and Srinivasan, V. (1998). *The predictive power of Internet-based product concept testing using visual depiction and animation*. Working paper, Stanford University, CA.
- Finkbeiner, C.T. and Platz, P.J. (1986, October). *Computerized versus paper and pencil methods: A comparison study*. Paper presented at the Association for Consumer Research Conference. Toronto.
- Foytik, M. (1999). Conjoint on the web - lessons learned. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 9-22). Sequim, WA: Sawtooth Software.
- Gordon, M.E. and De Lima-Turner, K. (1997). Consumer attitudes toward Internet advertising: A social contract perspective. *International Marketing Review*, 14 (5), 362-375.
- Huber, J. and Hansen, D. (1986). Testing the impact of dimensional complexity and affective differences of paired concepts in adaptive conjoint analysis. In M. Wallendorf and P. Anderson (Eds.), *Advances in consumer research* (No. 14, pp. 159-163). Provo, UT: Association for Consumer Research.
- Huber, J., Wittink, D.R., Fiedler, J.A., and Miller, R. (1993). The effectiveness of alternative preference elicitation procedures in predicting choice. *Journal of Marketing Research*, 30, 105-114.
- Johnson, J.S., Leone, T., and Fiedler, J. (1999). Conjoint analysis on the Internet. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 145-148). Sequim, WA: Sawtooth Software.

- Melles, T. und Holling, H. (1998). *Einsatz der Conjoint-Analyse in Deutschland. Eine Befragung von Anwendern*. Unveröffentlichtes Manuskript, Westfälische Wilhelms-Universität Münster.
- Meyer, L. (1998). *Predictive accuracy of conjoint analysis by means of World Wide Web survey* [Online]. Available: <http://www.lucameyer.com/kul/menu.htm>.
- Orme, B.K. and King, W.C. (1998). *Conducting full-profile conjoint analysis over the Internet*. Working paper, Sawtooth Software.
- Reiners, W. (1996). *Multiattributive Präferenzstrukturmodellierung durch die Conjoint-Analyse: Diskussion der Verfahrensmöglichkeiten und Optimierung von Paarvergleichsaufgaben bei der adaptiven Conjoint Analyse*. Münster: Lit.
- Reiners, W., Jütting, A., Melles, T. und Holling, H. (1996). *Optimierung von Paarvergleichsaufgaben der adaptiven Conjoint-Analyse*. Forschungsreferat zum 40. Kongreß der Deutschen Gesellschaft für Psychologie.
- Saltzman, A. and MacElroy, W.H. (1999). *Disk-based mail surveys: A longitudinal study of practices and results*. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 43-53). Sequim, WA: Sawtooth Software.
- Tscheulin, D.K. and Blaimont, C. (1993). Die Abhängigkeit der Prognosegüte von Conjoint-Studien von demographischen Probanden-Charakteristika. *Zeitschrift für Betriebswirtschaft*, 63 (8), 839-847.
- Voeth, M. (1999). 25 Jahre conjointanalytische Forschung in Deutschland. *Zeitschrift für Betriebswirtschaft – Ergänzungsheft*, 2, 153-176.
- Wittink, D.R. and Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing*, 53, 91-96.
- Wittink, D.R., Vriens, M., and Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, 11, 41-52.

COMMENT ON MELLES, LAUMANN, AND HOLLING

Gary Baker

Sawtooth Software, Inc.

A market researcher is concerned with both respondent reliability: are the data consistent, *and* validity: do the data correctly model behavior? Proper measures to test data reliability and validity should be a concern in every study, not just those (like this one) comparing different methods of data collection.

The paper concludes that although the online study appears to have more noise, nevertheless the results appear very acceptable. I think the researchers did an excellent job before and after data collection ensuring that respondent dropout and bad data were minimal. A functional, quick-loading page design is crucial in keeping respondents engaged in the interview, especially with online channel-surf mentality. The respondents were also warned up-front on the amount of time they would commit to (20 minutes), another good practice to lower the dropout rate. One interesting note in this regard can be drawn from two 1992 Sawtooth Software Conference papers. Witt/Bernstein and Salzman cited studies where disk-by-mail surveys stating a 20-minute completion time slightly lowered the response rate compared to not stating a completion time (also the response rate was higher for shorter stated completion times). More definitive research could be done comparing how completion times affect the response rate and the dropout rate in online data gathering techniques versus traditional techniques. In any case, we should be honest with respondents and honor the time a respondent gives us.

Perhaps additional methods could be added to test for clean data and to encourage respondents to give clean data. In the 1999 Sawtooth Software Conference, Huber *et al.* reported including a test/retest measure repeating a holdout task during the course of the interview. This gives an idea of internal validity. Another method used in that study was providing extra incentive for the respondents to give consistent answers. Respondents were told that the computer would track answers and the respondent's reward would be increased for consistent answers. Those who took enough time and passed a test/retest reliability measure were rewarded extra for their efforts. This practice might potentially bias the conjoint utilities, and perhaps more research should be done. If it does prove helpful, this technique could be adapted to web surveys.

Achieving cleaner data also rests upon how respondents are solicited for online research. One should use the same careful selection techniques for online respondents as for any other data collection method. It is a mistake to suppose that if web-based reliability is approximately equal to that of disk-by-mail, simply posting a click-through banner ad on the Web will produce valid data. Although the total respondent count is often very high in such a case, overall response rate can be extremely low, possibly resulting in a high degree of self-selection bias. In this study it may not be critical, but nevertheless should be remembered.

To conclude, the authors project favorable results for online conjoint studies, but warn that this assumes the Internet is an appropriate data collection method for the study. We are not guaranteed reliable and valid data from Internet conjoint. The researchers warn in their conclusions that study preparation for an online interview is more intensive than comparable personal interviews. They also note that although data collection via the Internet takes less time

and money, preparing that data for analysis is more complicated than traditional survey methods in spite of the extra precautions taken in study preparation.

REFERENCES

- Witt, Karlan J., and Bernstein, Steve (1992) "Best Practices in Disk-by-Mail Surveys," Sawtooth Software Conference Proceedings, pp. 5-6.
- Saltzman, Arthur (1992) "Improving Response Rates in Disk-by-Mail Surveys," Sawtooth Software Conference Proceedings, pp. 31-32.
- Huber, Joel, Orme, Bryan K., and Miller, Richard (1999) "Dealing with Product Similarity in Conjoint Simulations," Sawtooth Software Conference Proceedings, p. 259.

BRAND/PRICE TRADE-OFF VIA CBC AND Ci3

Karen Buros

The Analytic Helpline, Inc.

INTRODUCTION

This paper will discuss an adaptation of Brand/Price Trade-off (BPTO) to computer-based interviewing using Discrete Choice analysis. The paper first presents a brief overview of the traditional approaches to pricing a product or service within a competitive environment – Brand/Price Trade-off and Discrete Choice Analysis – and the blending of these approaches into Brand/Price Trade-off via Discrete Choice. This modification, which will be called BPTO Discrete Choice, incorporates the “trading up” interviewing approach of BPTO programmed through Ci3, with Discrete Choice analysis using CBC. Ray Poynter at the 1997 Sawtooth Conference presented an adaptation of BPTO for computer-assisted interviewing.¹ This approach differs most notably in the use of Sawtooth Software’s CBC in the analysis of the data.

A comparison of BPTO, Discrete Choice and BPTO Discrete Choice, using actual studies will be presented. The comparison will point out how the approaches differ in their data gathering techniques and the differing response patterns used by respondents to complete the task.

The author gratefully acknowledges the contributions of data for this analysis from Goldfarb Consultants on behalf of Ford Motor Company and Data Development Corporation.

PAPER AND PENCIL BRAND/PRICE TRADE-OFF

Interviewer-administered Brand/Price Trade-off was introduced in the 1970’s, first into the US in a paper published by Market Facts² and into Europe³. BPTO continues to be a popular approach for studying price within a competitive environment and has been extensively described in published articles.

While the approach can have many variations, in essence it involves a range of brands (generally four to eight brands) and a range of prices (either an array of six to nine “steps” or a series of prices keyed to each brand).

The respondent is shown a display of all brands, each with its lowest price, and is asked to choose the brand he/she “would buy”. The price of the chosen brand is raised to its next higher “step”, keeping all other brands at “lowest price”, and the respondent selects from the new array. Each time a brand is chosen its price is raised. When a brand has been chosen at its highest “price step” it is removed from the set. The process continues through all possible choices (providing a complete response to the brand/price matrix) or through a set number of choices (at least one brand through its highest price and a set number of choices).

¹ Ray Poynter (August 1997), “An Alternative Approach to Brand Price Trade-Off”, Sawtooth Software Conference Proceedings.

² Richard Johnson (1972), “A New Procedure for Studying Price-Demand Relationships,” Chicago, Market Facts, Inc.

³ Chris Blamires (April 1987), “Trade-Off Pricing Research: A Discussion of Historical and Innovative Applications,” Journal of the Market Research Society.

Data from the BPTO study can be analyzed in a number of ways.

- “First choice” for each brand can be calculated by “counting” the number of times each brand is chosen versus other brands for any given price scenario across respondents.
- Alternatively, marginal values for each brand and price level can be calculated by viewing the data as a respondent-level “matrix” of rank order choices under a conjoint approach (such as Sawtooth Software’s CVA program). Under the conjoint approach, the “part worths” for brand and price can be used to determine which brand a respondent is likely to buy in a competitive price scenario.

As pointed out in a 1996 paper by Johnson and Olberts⁴, the “counting” method allows for brand-price interactions, while “part-worths” do not, for individual-level data.

Despite its popularity for pricing studies in consumer goods, durables and services, practitioners have expressed concern about the approach. Anecdotal evidence indicates that respondents quickly “catch on” to the pattern of price increases and may view the interview as a “game”. The implication is that price would be over-emphasized, as respondent’s may choose the lowest price alternative, or under-emphasized as respondents “stay with” a favorite brand despite pricing they would never accept in the marketplace.

BPTO is easy to administer in the field and can be tailored to local market pricing conditions.

DISCRETE CHOICE PRICING

Discrete Choice was introduced in the 1980’s⁵ (Ben-Akiva and Lerman (1985) as a viable alternative to BPTO for studying price within a competitive environment. Like BPTO it can take many forms, both in the interviewing approach (“paper and pencil” or computer-administered) and in the analysis (aggregate analysis – “total market”, latent class analysis, or individual-level analysis – Sawtooth Software’s ICE or, more recently, Hierarchical Bayes⁶). This paper will discuss an often-used approach – computer-administered CBC using Latent Class and ICE from Sawtooth Software.

Under the Discrete Choice approach, arrays of brands and prices (a single array or prices keyed to each brand) are developed. In a CAPI interview, the respondent is presented with a scenario of brands and prices, is asked to choose the brand he/she “would buy” followed by a new array. The display of prices is not “ordered” (dependent on the prior selection). The respondent may be presented with 10 to 20 choices, depending on the design of the particular study. The respondent is permitted to “opt out” – indicate that all choices are unacceptable.

In the analysis, “utilities” (part worths) are computed for each brand and price using Multinomial Logit or an equivalent approach. Price can be treated as individual “levels” or as a linear price function. Brand/price interaction terms can be included. The “utilities” can be used to simulate a brand’s share under any array of competitive prices.

⁴ Rich Johnson and Kathleen Olberts (1996), “Using Conjoint Analysis in Pricing Studies, Is One Price Variable Enough?”, Sawtooth Software Conference Proceedings.

⁵ Moshe Ben-Akiva and Steven Lerman (1985), “Discrete Choice Analysis: Theory and Application to Travel Demand”, MIT Press Series in Transportation Studies.

⁶ Greg M. Allenby (1996), “Recent Advances in Dissaggregate Analysis: A Primer on the Gibbs Sampler”, presented at the American Marketing Association Advanced Research Techniques Forum.

While the approach is easy to administer, with computers available for interviewing, and frequently used, some practitioners have expressed concern that the approach may still place too much emphasis on price. Alternatives, such as adding a “non-important” disguising variable, have been suggested.

BPTO VIA DISCRETE CHOICE

This approach can be thought of as a melding of the two approaches – BPTO and Discrete Choice. The approach retains, but modifies the ordered presentation of prices in the interview and treats the data as though it had been collected in a Discrete Choice, CBC interview. The approach is adaptive. The respondent is asked unpriced brand preference, prior to the pricing exercise, which is used to set the pattern of price changes.

Interview Flow

An array of brands (four to eight) and prices is developed. Price is calculated as a percentage increment from each brand’s “market” or expected price. That is, a brand with a “market price” of \$3.00 might have a price array of \$2.10, \$2.40, \$2.70, \$3.00, \$3.30, \$3.60, \$3.90 and \$4.00, or 10% changes from the “market price”.

The interview is programmed in Sawtooth Software’s Ci3. Before the pricing exercise, the respondent is asked rank order unpriced brand preference.

The BPTO Discrete approach differs from BPTO in ways designed to imitate marketplace activity and to take advantage of computer-assisted interviewing.

The pricing exercise starts with a display of brands and “market prices” rather than “lowest prices” so that the respondent sees prices that are familiar.

The respondent is asked to select the brand he/she “would buy” or “none” from the array of brands at “market price”. Like BPTO, the price of the chosen brand is raised one step, while others are held constant and the respondent is asked to choose again.

After the initial choices, the pattern is altered. The price of the highest ranked (unpriced preference) non-chosen brand is lowered. In subsequent choices, the prices of other, non-chosen brands are lowered while the price of the chosen brand is raised. The concept of lowering prices is comparable to a brand “going on sale” and seems to be accepted by the respondent as realistic.

When a brand reaches its “highest price”, it remains in the display while prices of other brands are lowered. If the pattern becomes unstable, too many brands at too high or low prices, the pattern is reset to begin again at a new starting price scenario.

In total, the respondent makes 14 to 20 choices depending on the number of brands and prices in the study.

Data Set and Analysis

While the tasks are diverse across sample, they do not conform to any “design” criteria nor are they generated randomly. These scenarios and responses are exported into a file that can be read by the CBC Latent Class, ICE and Hierarchical Bayes programs.

Although a negative price utility can be enforced using “constraints” through CBC, the approach includes additional information about price and brand preference in the data file. For obvious reasons, respondents are not asked, in a task, if they would prefer to pay a higher or lower price for the same brand. Nonetheless, it seems sensible to include price preference information. However, this inclusion could affect the importance of price relative to brand. To create a “balance”, a limited number of “price preference” scenarios and “brand preference” scenarios (from the brand ranking) are added to the data set for each respondent. In experience, this does not completely eliminate the need for “constraints” in studies, but does markedly reduce the incidence of reverse price utilities.

From this point, the study is analyzed as though the data were collected via CBC.

EVALUATING THE APPROACHES

The ideal evaluation of pricing approaches would administer equivalent interviews using the three approaches, compare the strategies recommended by each, implement the strategies in a controlled situation and determine whether the brand in question performed as predicted. In practical terms, this is most difficult.

Rather, this paper will attempt to address the “concerns” practitioners raise with each approach and compare the approaches on that basis. Specifically, each approach will be evaluated on:

- The nature of the data collected
- Simplification strategies or “games playing” by the respondent

To do this, two to three studies using each approach are analyzed. The studies range from high-ticket vehicles and services to food products. The studies use a range of designs. Two studies are analyzed under the BPTO approach, the first involving vehicles (4 products at 12 price points (where each respondent completed the full price matrix) and an FMCG (14 products at 8 price points) using a partial completion of the matrix by each respondent. The Discrete Choice studies involved a service (10 brands at 8 prices) and an FMCG (7 brands at 9 price points). The BPTO via CBC studies (3) involved two vehicle studies of 4 brands at 8 prices and 6 brands at 8 prices, and a study of high-ticket services.

Despite these differences, this analysis suggests that respondents react differently to the alternative ways of presenting price and provides some basis for comparing the approaches.

THE NATURE OF THE DATA COLLECTED

In approaching the examination of the data, we want to keep in mind the goal of the analysis. The company conducting the study wants to determine the “worth” of the brand – that is, whether their brand can command a price premium in the marketplace and, if so, how much. Under “ideal” circumstances, we would simply ask the respondent how price sensitive they are, how brand loyal they are and how much of a premium they would pay for a particular brand. But, those questions are too hard for the respondent to answer in a way that we can translate to “market reaction”.

Rather than ask the question directly, we try to present the respondent with “pricing scenarios” and ask which product they would buy under the scenarios. Then across respondents we estimate how “our brand” will fare. While the three approaches are similar in the question to the respondent, the data are quite different.

BPTO

While BPTO may be faulted for turning the price question into a game, it has some strengths that should not be ignored. We can assume that the majority of people would prefer to pay a lower price for a product than a higher price for the same product. (Recognizing that price can be indicative of quality and ignoring that issue for now.) BPTO is an orderly presentation of price. By starting at the lowest price and raising price, the respondent can never choose the same product at a higher over a lower price.

Since the exercise is often ended when a “sufficient” number of choices have been made, the majority of information is collected at the lower and mid-range price points. One of the studies analyzed here involved completion of the entire price grid, which might look like this in terms of the ranks chosen:

	<i>BRAND A</i>	<i>BRAND B</i>	<i>BRAND C</i>	<i>BRAND D</i>
PRICE 1	1	2	5	9
PRICE 2	3	4	8	13
PRICE 3	6	7	12	18
PRICE 4	10	14	15	23
PRICE 5	11	17	19	24
PRICE 6	16	21	25	29
PRICE 7	20	26	28	31
PRICE 8	22	27	30	32

Or like this, when the exercise is ended when a brand reaches its highest price and a set number of choices is attained.

	<i>BRAND A</i>	<i>BRAND B</i>	<i>BRAND C</i>	<i>BRAND D</i>
PRICE 1	1	2	5	9
PRICE 2	3	4	8	13
PRICE 3	6	7	12	18
PRICE 4	10	14	15	
PRICE 5	11	17	19	
PRICE 6	16	21		
PRICE 7	20			
PRICE 8	22			

While each respondent may not be entirely consistent in their responses, this example respondent would clearly pay a premium for Brand A over the other brands. On the other hand, if the brand’s “market price” is at Price 4 or 5, we are collecting the least amount of information above the “market price”, precisely the question we are trying to answer, unless the respondent completes the entire price grid.

Discrete Choice

Discrete Choice takes a very different approach. Price is not constrained by the exercise. Each choice is independent of the earlier choices. Across all choice tasks and all respondents, the full price grid is likely covered. But, the respondent must also re-evaluate each scenario on its merits. For example, the respondent may see:

	BRAND A	BRAND B	BRAND C	BRAND D
PRICE	Price 7	Price 2	Price 4	Price 5

Choosing Brand B at Price 2. Then see:

	BRAND A	BRAND B	BRAND C	BRAND D
PRICE	Price 3	Price 8	Price 1	Price 4

Choosing Brand C in a response consistent with the earlier example. This can be more difficult for the respondent and can result in “price reversals” (higher prices preferred to lower prices for the same brand) due to inconsistency of response.

Discrete BPTO

Discrete BPTO takes a third approach. Assuming that the mid price points are of the most interest (“market price”) and that the respondent will answer more consistently with an orderly presentation of prices, this example respondent would first see:

	BRAND A	BRAND B	BRAND C	BRAND D
PRICE	Price 4	Price 4	Price 4	Price 4

Choosing Brand A at Price 4. Then see:

	BRAND A	BRAND B	BRAND C	BRAND D
PRICE	Price 5	Price 4	Price 4	Price 4

Continuing to Choose brand A, this respondent would see:

	BRAND A	BRAND B	BRAND C	BRAND D
PRICE 1	Price 6	Price 3	Price 4	Price 4

And the respondent should choose Brand B. In other words, this approach does not enforce price ordering but attempts to aid the respondent by making the choices more obvious. Far fewer preferences are expressed at the lower price points than is the case with BPTO, since only less preferred products are seen at the lowest price points. And, unlike Discrete Choice, the choices expressed at the higher price points are more often for preferred products, since a product is not seen at the higher price points unless it is chosen at lower price points. While Discrete BPTO is obtaining more information about the “gaps” in price for preferred products, it lacks information about the general ordering of brand and price for the respondent.

We know that, logically, respondents would prefer a lower price to a higher price. This information is currently ‘added’ to the choice task set by psuedo-tasks indicating the same

product should be preferred at a lower price rather than at a higher price. To balance the price ordering, similar information is added about basic product preference at the ‘same price’.

Johnson and Olberts suggest a different approach in their 1996 paper. --- that if Brand A is preferred to Brand B at the same price point, it will also be preferred to all other brands (and Brand B) when Brand A is at lower price points. While this approach has not yet been implemented, it offers a fruitful area of investigation.

SIMPLIFICATION STRATEGIES OR “GAMES PLAYING”

An often voiced concern about any approach to pricing, has been that respondents will view the exercise as a “game”, offering responses that are not related to actual decision-making. The theory is that a respondent, faced with a task they may find tedious, unrealistic or uninvolved, will use a method to simplify the task. These studies are examined to determine the extent to which different strategies are being used. The findings are applicable to these studies and may not apply to studies of different products or different markets.

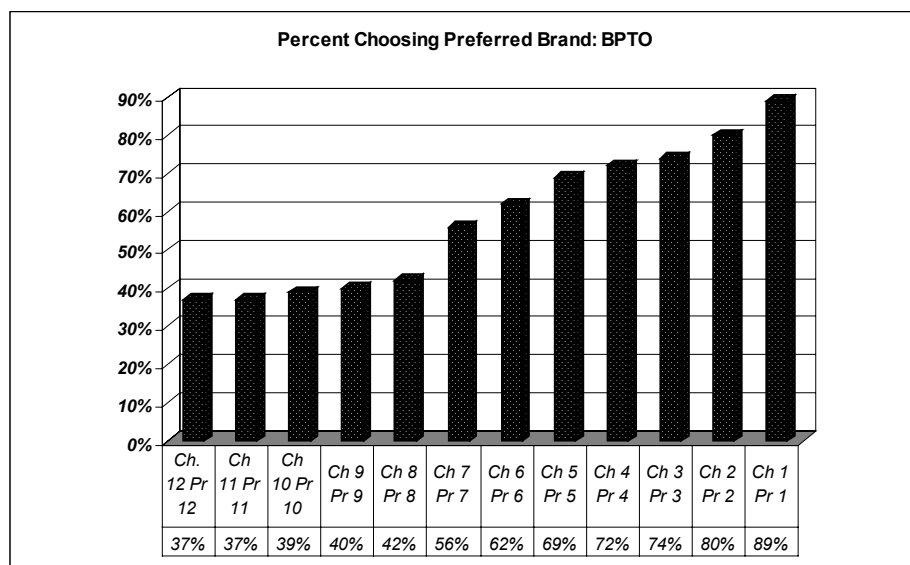
BPTO Simplification Strategies

One of the goals of pricing studies is to determine how much of a “premium” in price a brand can command over its competitors. So, to a greater or lesser degree, we would expect some respondents to “stay with” a brand despite its premium price relative to competition.

It is a fine line between a brand commanding a premium price and the “game” of simply selecting a brand regardless its price.

Favored Brand Strategy in BPTO

The first BPTO study has 12 price points and the respondent completes the full matrix. In this study, 37% of respondents selected a brand in the first task and continued to select the brand consecutively through 12 tasks, or all 12-price points. Only after the brand was removed did they switch to another brand. An additional 19%, for a total of 56% of respondents followed the same brand from price point 1 task 1 to price point 7 task 7. Thus, more than half the respondents “patterned” a brand over the first 7 price points.



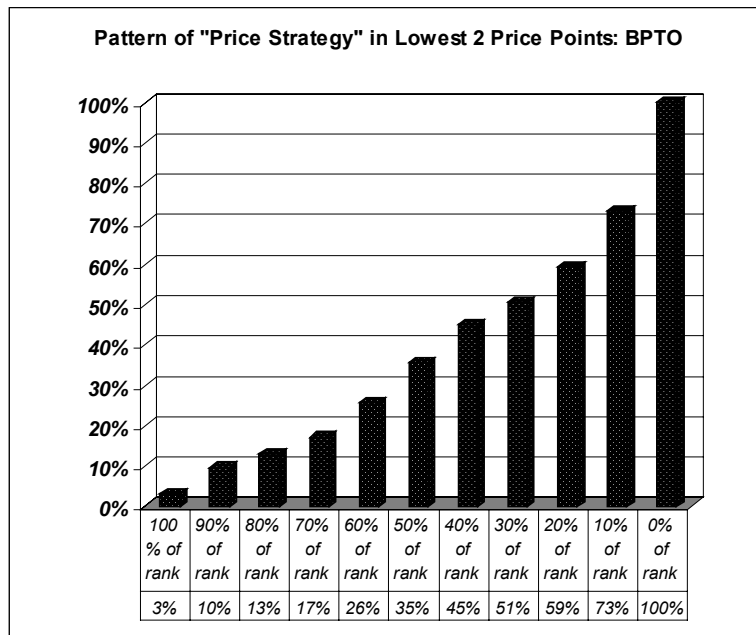
The second study shows a similar pattern. More than half of the respondents (52%) chose a favored brand in all 7 of its possible price points, thus removing it from the display of choices.

Minimum Price Strategy

Another possible strategy is to always select the lowest price brand available. Again, this is a fine line. The respondent may genuinely prefer the lowest price among the choices available.

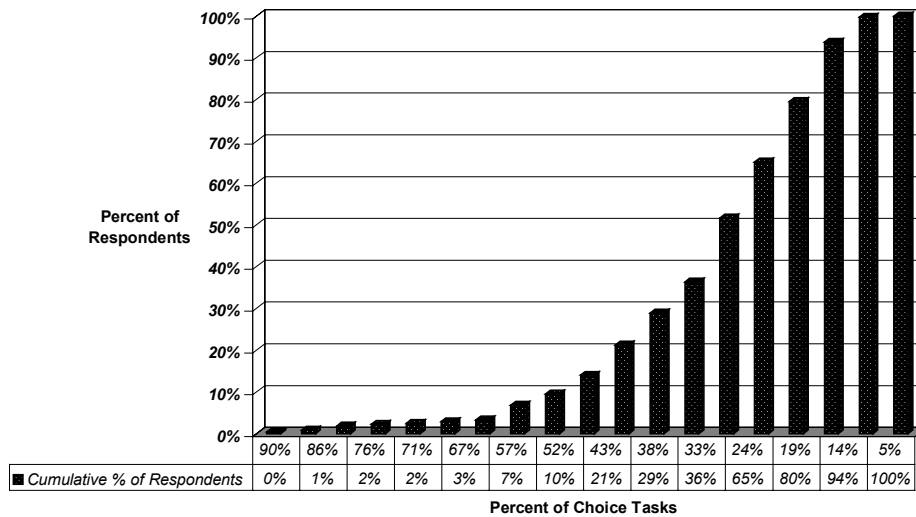
To obtain a sense of the use of this strategy in the first BPTO study, “rank orders” given across the brands in the first two price points were examined. While each individual matrix could be examined, the sum of the ranks in the first two price points was used as an indicator of a “minimum price” strategy.

There was evidence of a “minimum price” strategy among these respondents. Although only one respondent had all eight possible ranks within the first 2 price points (4 brands within 2 price points); more respondents had 6 or 7 out of the 8 possible ranks within the first 2 price points. Half of the respondents had 1/3 of their ranks in the first 2 price points.



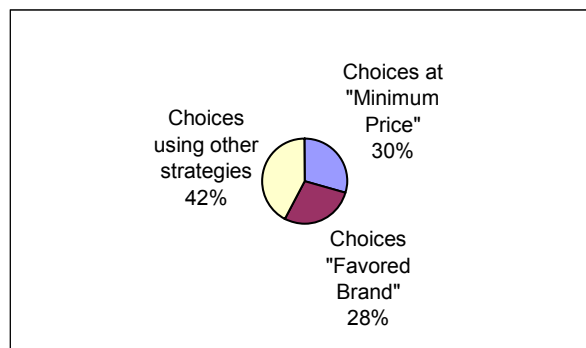
For the second BPTO study, we can view the data on a task basis. A “minimum price” strategy in this case is defined as a brand choice (not the favored brand) when no other brand has a lower price. Using this definition, the second BPTO study also showed evidence of a “minimum price” strategy. While only 10% of respondents chose the minimum price alternative in half or more of their tasks, about half of the respondents chose the “minimum price” alternative in 1/3 or more of their tasks.

Cumulative % of Respondents Using a Minimum Price Strategy



In both of these BPTO studies, respondents did not have the option of choosing “none”.

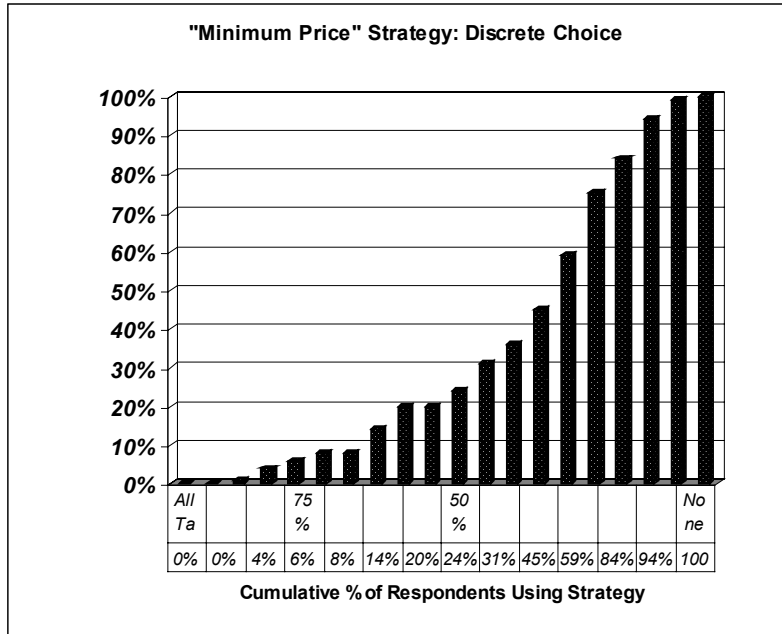
Examination of the total number of tasks in the second BPTO study shows about an equal split between the “Favored Brand” and “Minimum Price” choices



Discrete Choice Simplification Strategies

The most obvious simplification strategy in Discrete Choice is to choose “none of these”. In both Discrete Choice studies, virtually no one chose “none” all the way through the exercise. However in the first study, 20% of the respondents did choose “none” in at least half their tasks. In the second study, only 6% chose “none” in at least half of their tasks. So it is not evident from these studies that respondents were using “none” to simplify the task. In fact, there were more brands in the study than could be shown on the screen at one time. So not all brands were available in all tasks. Thus, “none of these” could be a legitimate response if the favored brand(s) are not shown.

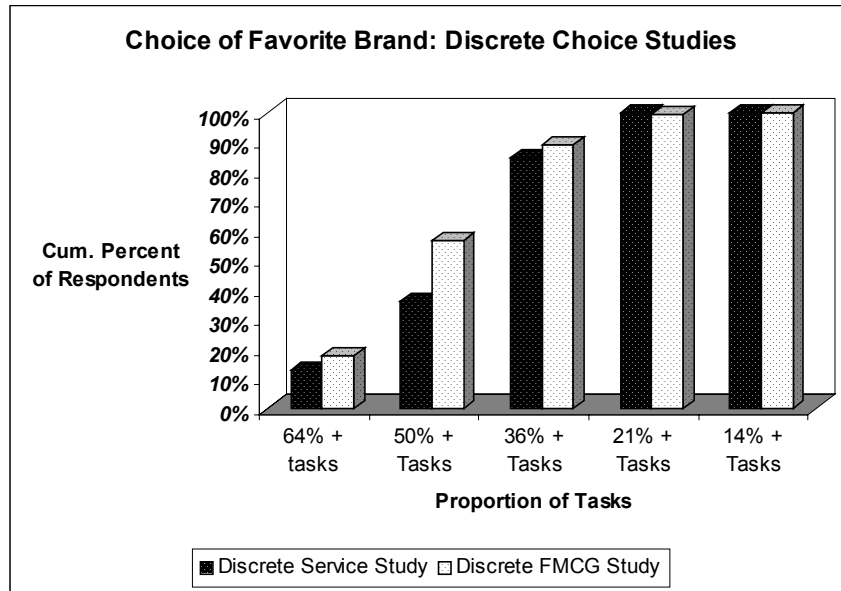
The “minimum price” approach is more prevalent. While no one always chose the “minimum price” alternative, in the first study, about 1 in 4 respondents chose the lowest price option in half of their tasks.



In the second study, the “minimum price” option was less often used. Overall 17% of respondents chose the minimum price option in half or more of their tasks.

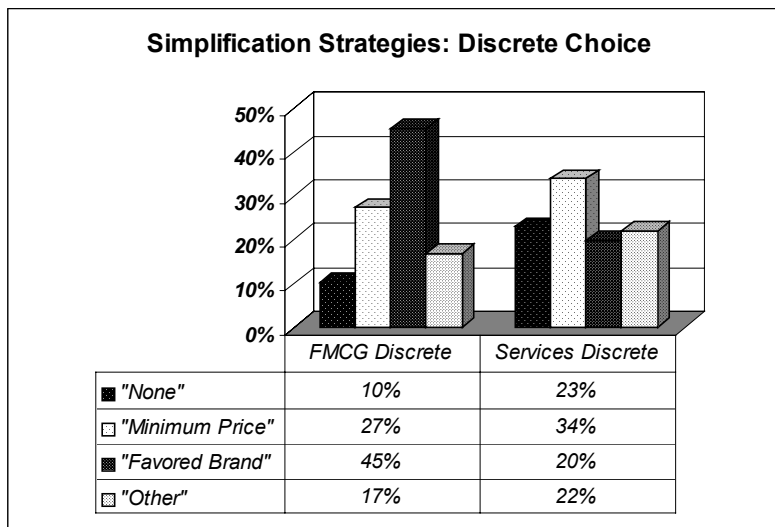
The first study included more brands (10) than were shown in each task. So we would expect to see more brand “switching”. As expected, respondents to this exercise tended not to stay with a favorite brand throughout the tasks. There was nonetheless a fair amount of brand loyalty in the study. One in three, 34%, of the respondents stayed with a favored brand in half of the tasks. Even accounting for the fact that brands rotated through the tasks, brand patterning seems less prevalent in this study than in the BPTO approach.

The second study used a similar design, in that not all brands were shown in all tasks. In the second study, “favored brand” choice was more prevalent. 56% of the respondents chose a “favored brand” in half or more of their choices.



Examination of the data on a task, rather than respondent base, shows dispersion of strategies employed under the Discrete Choice approach. In the first study, the most prevalent choice was to choose the minimum price alternative, in direct contrast to BPTO. Only 20% of the choices involved a “favored brand”. One in 4 choices involved “none” and the remaining tasks were completed under some other type of approach.

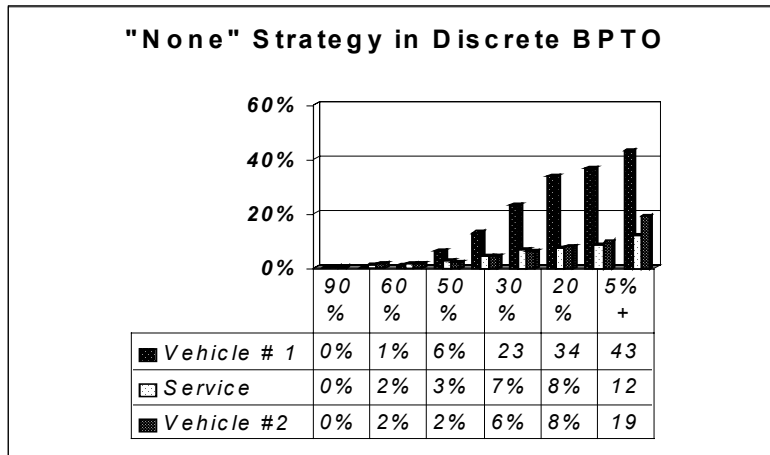
In the second study, the distribution of responses is different. The “Favored Brand” strategy is used more often and the “none” option is used less often. Use of an “Other” strategy is now just under 20% of tasks.



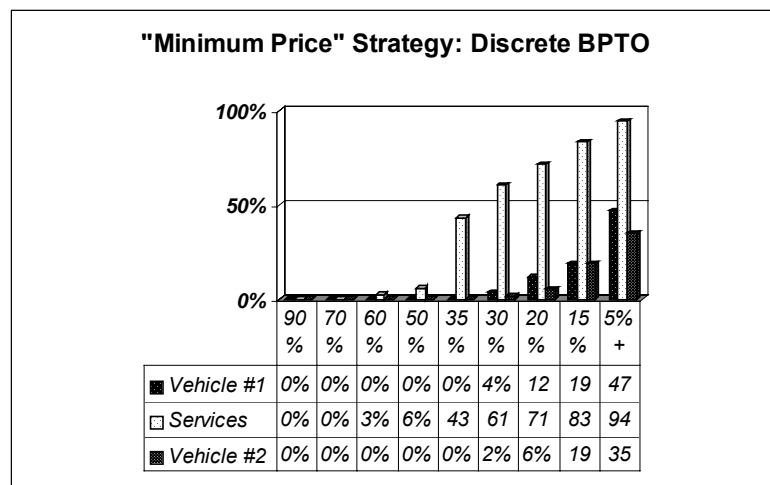
Discrete BPTO Strategies

In these studies, there was little use of the “none” option as a simplification strategy. For example, in the first vehicle study, only 6% of the respondents chose “none” in half or more of their tasks.

In the second vehicle study, the “none” option was never used by 80% of the respondents. This low use of the “none” option, in contrast to the discrete choice studies, is at least partially due to the fact that all brands were shown in all tasks in both studies.

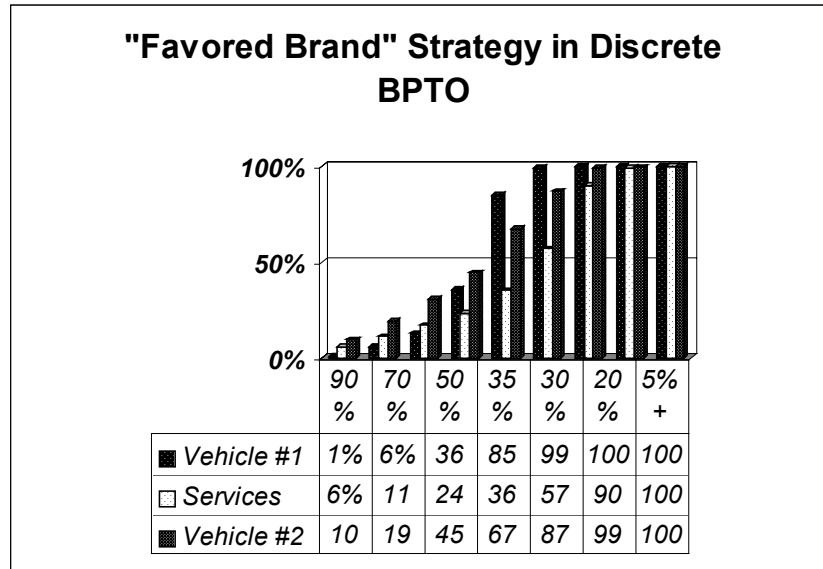


Likewise, there was little use of the “Minimum Price” strategy in the vehicle studies. About half the respondents never chose the minimum price option. Keep in mind, however, that in many tasks there was no single “minimum”. In explanation, in the first task, all prices are at the “mid-point”. In the second and third tasks, at least two prices are “mid-point” and prices do not drop until the fourth task. So, in at least 3 of the 14 tasks, depending on responses, have no possibility of being counted a “minimum price” choice. Nonetheless, the incidence of “minimum price” is low.



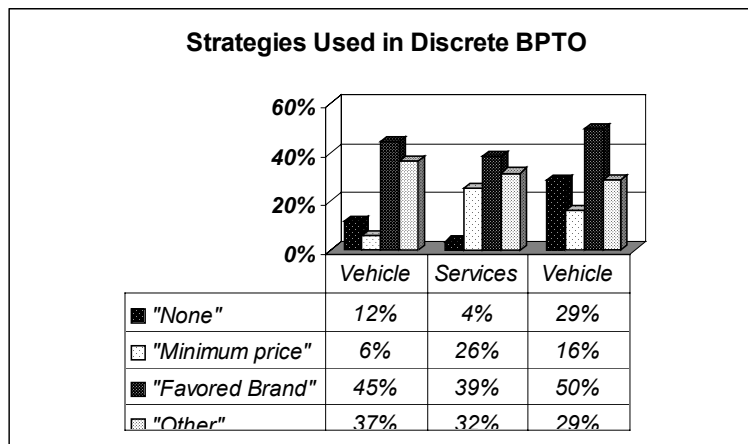
In the Services study, use of “Minimum Price” is greater, suggesting that the category or specific offerings, rather than the approach, determines the extent to which “Minimum Price” is employed.

In all the Discrete BPTO studies, there is a greater incidence of choosing a favored brand. In fact, like BPTO, this was the most prevalent strategy. An average of one in three respondents chose a favored brand in half of their choice tasks – more than in the other approaches studied.



The greater use of a favored brand strategy is evident also when the data are examined on a task basis. While in the vehicle studies, most tasks involved the selection of a “favored brand”; more than one-third of the tasks involved some other choice pattern. This is a very different pattern than the Discrete Choice study that involved fewer “other” choice patterns, but a more even dispersion across the types of patterns.

In the Services study, “Favored Brand” is the prevalent strategy, followed by “Other” and “Minimum Price”, in other words a greater mix of strategies than in the vehicle studies.



CONCLUSIONS

This paper has described an approach, Discrete Choice BPTO, which combines elements of the interviewing approach of BPTO including enhancements to make the choice tasks more realistic, with the analytic approach of Discrete Choice.

Two to three studies were examined for each of the three methodologies. While the studies were not completely comparable, they are used to illustrate the use of patterns in response to the choice tasks.

The three approaches employ different techniques to enforce “price ordering” – the preference for a lower price over a higher price.

- Under the BPTO approach, price is ordered during the interview – A respondent cannot prefer a higher price to a lower price for the same product. The option is never seen.
- Under Discrete Choice, price ordering is not enforced during the interview, but can be incorporated into the analysis by constraining price.
- Under the Discrete BPTO approach, the presentation of prices is more orderly than under randomized Discrete Choice. Price ordering is added in the analytic stage by adding “logical tasks” (lower prices preferred to higher prices) prior to the calculation stage for each respondent.

The approach of “filling out” the choice grid, as suggested by Johnson and Olberts, appears to be a fruitful one, warranting further study.

Likewise, the approaches differ in the type of data collected. Discrete BPTO collects data primarily in the mid and upper price ranges, BPTO concentrates in the lower to mid price points while Discrete Choice collects data across all price ranges.

Across the 7 studies examined, a variety of strategies were used – “Favored Brand”, “Minimum Price” and “Other” strategies. Both BPTO and Discrete Choice approaches evidenced a good mixture of strategies in the non-vehicle studies. In the Discrete BPTO approach, the Services study showed a better mixed use of strategies than did the Vehicle studies. It appears from these studies that “games playing”, if it is used by respondents to simplify the task, is at least used in a wide variety of ways.

Importantly, the three approaches employ different techniques to enforce “price ordering” – the preference for a lower price over a higher price.

- Under the BPTO approach, price is ordered during the interview – A respondent cannot prefer a higher price to a lower price for the same product. The option is never seen.
- Under Discrete Choice, price ordering is not enforced during the interview, but can be incorporated into the analysis by constraining price.
- Under the Discrete BPTO approach, the presentation of prices is more orderly than under randomized Discrete Choice. Price ordering is added in the analytic stage by adding “logical tasks” (lower prices preferred to higher prices) prior to the calculation stage for each respondent.

More recently, the approach of “filling out” the choice grid, as suggested by Johnson and Olberts, appears to be a fruitful one, warranting further study.

Importantly, computer-based interviewing using either Discrete Choice or Discrete BPTO, makes the task more complex, and in this author’s opinion, more interesting for the respondent. This should improve the quality of response

CHOICE-ADAPTED PREFERENCE MODELLING

*Roger Brice, Phil Mellor, and Stephen Kay
Adelphi Group Ltd*

BACKGROUND

The use of conjoint techniques in pharmaceutical market research presents its own set of issues affecting the choice of conjoint design. These are:

1. The high unit cost of interviews leading to generally smaller than ideal sample sizes.
2. Complex product offerings often requiring a larger number of product attributes than is allowed by some commercially available conjoint paradigms.
3. Problems in defining “the customer” leading to a need to include both physician and patient exercises, which then need to be integrated.
4. Complex patterns of product use with poly-pharmacy being a common feature of many markets – yet it is single product offerings that our clients are developing and we are asked to research.
5. A highly variable, non-symmetrically distributed assessment of the performances of current products by physicians.

The last mentioned issue, we believe, contributes as much as (if not more than) the question asked of respondents when faced with a conjoint task, to the well recognized over-estimation of new product share in market simulations¹. Consequently, it is an important focus of this paper.

Our hypothesis is that the predictive power and simulation realism of choice-based designs can be equaled (or even improved on) through full profile ranking (or rating) exercises with the addition of a ‘choice threshold’ question after the ranking exercise has been completed. This means that smaller sample sizes can be utilized than would be recommended for discrete choice paradigms such as Choice based Conjoint (CBC) but that the benefits of such paradigms can be included.

We have used the data so gathered to generate utility scores using standard multinomial logit (MNL) analysis, at the individual respondent level. Through then maintaining true individual choice thresholds (analogous to the ‘none’ option in CBC) throughout the simulation exercise, we are able to conduct market simulations that better reflect realistic market structures. They may, therefore, have validity advantages compared to discrete choice paradigms as, unlike most discrete choice simulation modules, these simulations preserve true individual level choice thresholds throughout the simulation. They certainly offer an advantage on a per unit cost basis.

This paper forms a progress report on our ongoing approach to testing the hypothesis above and to adding realism to market simulations. This has reached the stage where we now have a

¹ Ayland C & Brice R, ‘From Preference Share to Market Share – Some Experimental Results’, *Marketing and Research Today*, vol. 28, no.3, August 1999, pp89-98.

fully tested conjoint simulation module (CAPMOD) which has been used successfully for a number of clients.

We use data collected on behalf of one of these clients and also data generated from an extension to this funded by Adelphi exclusively for this paper.

THE EXPERIMENT

Background

Our client had a product in development for the treatment of a common chronic disorder in which acute episodes can occur. Market research was required to assist in future development (clinical trial) decisions. In particular, the relative importance of selected clinical drivers and the trade offs that physicians are prepared to make among these when making treatment decisions.

The data for this study were gathered in March 1999. The experiment described here included these data (made suitably anonymous to preserve confidentiality) plus data gathered during the Adelphi-funded study conducted in March 2000. The starting point (attribute and level grid) was common to both. They differed in terms of the conjoint methodology used thus enabling comparisons to be made.

The Conjoint Exercises

Both the 1999 and 2000 conjoint designs were based on a matrix with nine attributes each with two, three or four levels.

Attributes	Levels			
Sustained improvement in efficacy attribute 1	Less than market leader	Equal to market leader	Some improvement over market leader	Great improvement over market leader
Sustained improvement in efficacy attribute 2	No improvement		Some improvement	Great improvement
Sustained reduction in efficacy attribute 3	No reduction		Some reduction	Great reduction
Quality of life data	No evidence of improvement (and no data)		Evidence of clinically significant quality of life improvements	Evidence and approved claim/label for quality of life improvements
Maintenance dose regimen	4 times daily		3 times daily	2 times daily
Additional dosing above maintenance dose	Not indicated for this use		Indicated. Slight increase in side effects	Indicated. No increase in side effects
Prevention of accelerated decline in function	No evidence			Evidence
Side effect profile	Worse than existing class		Similar to existing class	Improvement over existing class
Reduction in number of acute attacks	No			Yes

Table 1: Attribute & level grid used for conjoint design

Also common to both designs was the conjoint question setting. Physicians were asked to complete the exercises in the context of the last patient seen suffering from the condition for which it was anticipated the new drug would be indicated. We selected this question setting, rather than presenting a patient pen picture or using “a typical patient” or “the last 20 patients” for example, as we wanted to ensure that:

1. The focus of the physician was on an actual patient and his/her treatment decision. This ensured that the respondent would know all the extra information about the patient necessary to make the conjoint exercise more accurate.
2. We wanted a representative sample of treatment decisions so that the source of any new business could be identified and profiled in our simulation process. Having either “the last 20 patients” or a sample of patients typical for each respondent, which is far from a representative (or typical) sample of patients presenting with the condition, would not have allowed this.

The two conjoint studies can be summarized as follows:

	March 1999	March 2000
Paradigm	Classic “Full Profile”	CBC (Sawtooth)
Sample Size	75	50
Stimulus Material	Full Profile, Cards	Full Profile, Cards
Conjoint Task	Ranking	Choice Sets of 3 Cards
Number of Tasks	28 cards	15 choice tasks
Hold Out	3 cards	1 choice task
Analysis Method	OLS	MNL
Utility Output Level	Individual	Aggregate
Simulation Methods	a) 1 st Past the Post(Conjoint Analyser - Bretton Clark) b) Probabilistic (Simgraf – Bretton Clark)	Share of Preference in CBC Simulation Module a) with no-buy option w/o no-buy option

Table 2: Summary of the two conjoint studies

Comparisons of the two Conjoint Studies

The two studies can be considered as alternative conjoint approaches to a common problem. We therefore made comparisons between them using three standard measures on which comparisons were possible: attribute relative importance based on overall attribute/level utility scores, holdout analysis and market simulation results.

1. Relative Importance of Attributes

Comparisons of results from the two MNL-based analyses (CBC with and without the ‘no-buy’ option and the OLS-based analysis (Conjoint Analyser) are shown in Table 3. Simgraf is a simulation module that uses output from Conjoint Analyser and, therefore, the concept of attribute relative importance is not applicable.

Attribute	Aggregate Level Probability Model (MNL) (CBC with no-buy)	Aggregate Level Probability Model (MNL) (CBC w/o no-buy)	Individual Preference Ranking (OLS) 1stPP (Conjoint Analyser)	Individual Preference Ranking (OLS) Probabilistic Model (Simgraf)
Attribute 1	25.1%	24.4%	27.3%	n.a.
Attribute 2	8.6%	9.2%	11.2%	n.a.
Attribute 3	9.4%	10.2%	9.3%	n.a.
Attribute 4	13.0%	13.3%	10.5%	n.a.
Attribute 5	7.4%	5.6%	3.4%	n.a.
Attribute 6	3.5%	3.3%	2.0%	n.a.
Attribute 7	14.4%	15.5%	15.2%	n.a.
Attribute 8	9.9%	9.3%	9.3%	n.a.
Attribute 9	8.7%	9.1%	11.7%	n.a.

Table 3: Relative importance of attribute ranges – based on overall utility scores

2. Hold Out Analysis

We compare (Table 4) the predicted results for the hold out cards with their actual shares (occasions each was chosen during the interviews) with those predicted by three of the simulation modules:

- (a) Individual level, OLS, 1st past the post (Conjoint Analyzer)
- (b) Individual level, OLS, probabilistic model (Simgraf)
- (c) Aggregate level, MNL, probability model (CBC) - only that without the ‘no-buy’ option can be included.

Holdout Set within CBC exercise	Actual Share (CBC) ¹	Aggregate Level Probability Model (MNL) (CBC w/o no-buy)	Actual Share (individual pref ranking) ²	Individual Preference Ranking (OLS) 1stPP (Conjoint Analyser)	Individual Preference Ranking (OLS) Probabilistic Model (Simgraf)
Holdout Card 1	58.0%	65.7%	81.3%	70.7%	45.4%
Holdout Card 2	2.0%	10.2%	0.0%	1.3%	20.7%
Holdout Card 3	40.0%	24.1%	18.7%	28.0%	33.9%

¹ % of occasions each card selected as the preferred within 3-card holdout set
² % of occasions each card was the preferred of the three within total set

Table 4: Hold out analysis

3. Market Simulations

The predicted ‘shares of preference’ for all four simulations are shown in Table 5.

Product	Aggregate Level Probability Model (MNL) (CBC with no-buy)	Aggregate Level Probability Model (MNL) (CBC w/o no-buy)	Individual Preference Ranking (OLS) 1 st PP (Conjoint Analyser)	Individual Preference Ranking (OLS) Probabilistic Model (Simgraf)
Prod 1	12.5%	17.0%	18.7%	20.0%
Prod 2	2.1%	5.1%	1.3%	4.5%
Prod 3	2.8%	5.6%	1.3%	1.8%
Prod 4	14.6%	19.5%	13.3%	20.3%
Prod 5	47.2%	53.3%	66.7%	53.4%
‘stay as is’ (no buy)	20.9%	n.a.	n.a.	n.a.

Table 5: Market simulation results (share of preference)

It is impossible to draw an overall conclusion on which is the best predictive model relying solely on these results. We can observe that, based on the usual relative importance measure, they give broadly similar results. We also have no basis for an evaluative assessment of any differences. The hold out analyses favour the aggregate level MNL (CBC) and 1st past the post OLS models in that, in this case at least, they are superior to the probabilistic OLS model. The market simulations for the aggregate MNL (CBC) and probabilistic OLS models are similar. In this comparison, the 1st past the post OLS model differs from the other two.

It is tempting to conclude that as the MNL model (CBC) has appeared to be superior on two measures and the two OLS models each on only one, that the MNL model is the best. However the data are limited and we would like to see more direct comparisons being made. (It is also our belief that the notion of a universally best conjoint paradigm is a contradiction in terms). It could also be argued that what really matters performance in simulations of current markets and, most importantly, in future market predictions.

PREDICTED VS. ACTUAL MARKET SHARES FROM CONJOINT METHODS

Observations

The 1999 experiment described above was designed solely to establish priorities and trade offs among a specific list of product attributes. The list was known to be inappropriate for actual marketplace simulations. The elapsed time is also too short for any new product entry to have occurred. We therefore turn to previously published data to illustrate the point (well accepted among pharmaceutical researchers) that ‘off the shelf’ conjoint paradigms severely over-estimate in pharmaceutical markets.^{2 3}

² Ayland 89-98.

³ Brice R, ‘Conjoint analysis: a review of conjoint paradigms & discussion of outstanding design issues’, Marketing and Research Today, vol. 23, no.4, November 1997, pp260-6.

Data from an Adelphi paper published in 1997⁴ are reproduced in Table 6 and Table 7.

Product/ Product Class	Actual Market Share (current at time of study)	Simulated Share OLS probabilistic (Simgraf)
Established products	58%	56.5%
Recent Entrant 'A'	21%	21.1%
Recent Entrant 'B'	13%	13.0%
Recent Entrant 'C'	8%	9.4%

Table 6: Example of current market simulation

Product/ Product Class	Actual Market Share (2 years after study)	Predicted Share OLS probabilistic (Simgraf)
Established products	63.5%	50.0%
Recent Entrant 'A'	16.3%	15.1%
Recent Entrant 'B'	12.2%	8.9%
Recent Entrant 'C'	5.0%	6.4%
New Product #1	1.4%	4.2%
New product #3	0.9%	7.0%
New Product #3	0.7%	8.4%

Table 7: Example of future market prediction

Conclusions

In our experience, these findings are not uncommon. It is often possible to replicate the current market - and probabilistic methods may improve the likelihood of this. Future predictions for new product entries are very frequently over-stated. The reasons for this include brand effect, the degree of innovation associated with the new product, number in therapy class launch sequence and time to market as well as volume and quality of promotional spend.

Most efforts among pharmaceutical researchers to correct this have focused on correction (discount) factors.⁵ Our belief is that the realism of the simulation, and the implications of this for earlier steps in the paradigm, should first be examined more closely.

ISSUES AND SOLUTIONS IN MARKET SIMULATIONS

In order to make the product profiles we enter into market simulations more realistic, we obtain profiles of current products on the attribute and level grid on which the conjoint stimulus material is based. This also has the added benefit of making respondents very familiar with the attributes and their ranges that they will see in the conjoint cards.

Our experience is that there is a wide range in the perceptions that physicians have for the same products. The distributions of these perceptions are also often skewed with both positive and negative skews being found. It is also a feature of some pharmaceutical markets that products with very similar images can have quite different market shares.

⁴ Brice 260-6.

⁵ Summary Report, EphMRA 1999 Annual Conference - Building a Competitive Culture through Shared Responsibility. (Presentation by McAuliffe, Burrows & Jack) – www.ephmra.org/4_203.html

Currently available conjoint simulators use perceptions of products entered in the simulation that are common to each individual respondent – even those that use individual level utility scores. Because of the image distribution problem referred to above, this could lead to very misleading results. In methods that rely on aggregate level data throughout, the problem is of particular concern. This has always been a major concern to us with the use of discrete choice methods. The recent introduction of Hierarchical Bayes analysis with CBC (CBC/HB) is therefore a particularly exciting development.

These issues indicate the need for analysis **and** simulation capability at the individual level throughout. It also seems appropriate to allow current products to be “as is” in terms of their images for individual physicians. In this way, we maintain the reality of individual preferences and choices throughout the entire exercise. This is very different from how the ‘no buy’ option is treated within discrete choice paradigms. In CBC, for example, the ‘no buy’ option is, in effect, treated as a null profile. ‘None’ becomes an attribute with its own utility. This means that, if the number of products in a simulation differs from the number of alternatives in the choice tasks presented to the respondent, the estimates produced by the simulation will not be correct.⁶ In CBC, the use of the ‘no buy’ option is recommended for considerations of question realism; it is not recommended for inclusion in simulations⁷, thus often making simulating a whole pharmaceutical market extremely difficult.

CHOICE ADAPTED PREDICTIVE MODELING (CAPMOD)

Introduction

CAPMOD is a simulation module that can be applied to any conjoint analysis output (utility scores) provided that a choice threshold question (or question set) has been asked of all respondents. This concept can be applied to either preference- or choice-based data collection and analysis methods. With choice-based methods we prefer analysis be done at the individual respondent level such that individual respondents’ utility sets can be used throughout the simulation exercises, although this is not essential.

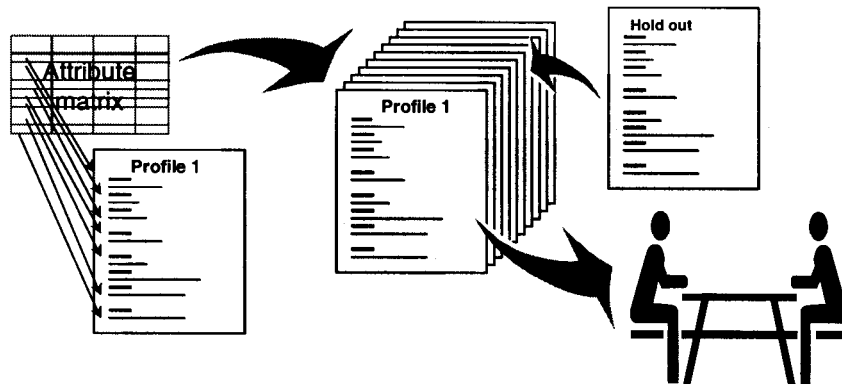
An Example of CAPMOD in Practice

The CAPMOD choice threshold question was used in the 1999 conjoint study described above. The question sequence is illustrated in Figure 1 and Figure 2. This shows how a choice question is applied to a “classic” full profile, OLS conjoint design. The objective is to obtain a threshold, the product profiles above which would have been prescribed had they been available and, importantly, below which none would have been prescribed even though preferences among them can be established. The logic here is that new product ‘A’ might be preferred to new product ‘B’ (and this information is utilized for utility score calculations) but neither would be purchased – a first essential step towards market share. By ranking all alternatives, including those below the choice threshold, the amount of comparative information per profile per respondent is maximized. This in turn leads to more robust calculations of utility scores than are possible with conventional discrete choice designs.

⁶ CBC User Manual V2.0, Sawtooth Software Inc., Sequim, WA, USA, Appendix E pp5.6.

⁷ CBC User Manual 5.6.

Stage 1: "Classical" full profile preference exercise



- ◆ Full profile derived from attribute matrix
- ◆ 28 profiles generated (3 holdout cards)
- ◆ Physician asked to rank the 28 profiles in order of their preferences for treatment of a specific patient

Figure 1: The conjoint interview: Stage 1

Stage 2: Choice threshold question

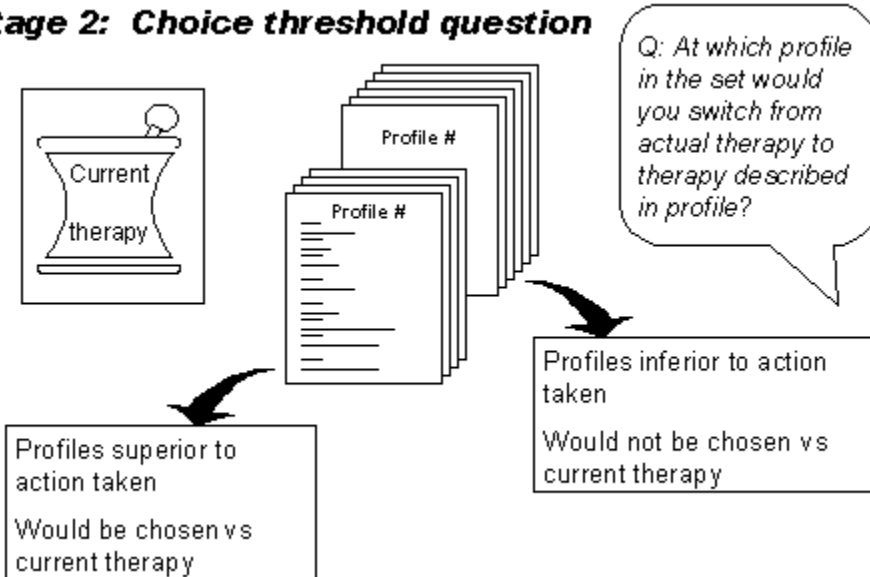


Figure 2: The conjoint interview: Stage 2

Using the choice question also allows market expansion through new product introductions to be calculated. The question shown can be worded to include whether the products in the set above the threshold would be prescribed in addition to current therapy or to replace it. In both cases non-drug therapy can be included.

Results of CAPMOD simulations are included in Table 8. We have taken a series of alternative possible outcomes for a new product and introduced each into the five-product market simulated above (Table 5). We have added the new product to the market simulation obtained from six different models.

1. preference ranking, individual level OLS, 1st past the post market model
2. preference ranking, individual level OLS, probability model
3. discrete choice, aggregate level MNL model (CBC w/o the ‘no-buy’ option)
4. discrete choice, aggregate level MNL CBC w/o ‘no-buy’ option) modified to recognize individuals’ differing images of current products
5. CAPMOD (preference ranking + choice threshold, OLS, 1st past the post)
6. CAPMOD (preference ranking + choice threshold, OLS, probabilistic)

Method	Share Prediction for New Products		
	Product #1	Product #2	Product #3
Individual level OLS (pref ranking, 1 st PP)	63.5%	34.0%	4.0%
Individual level OLS (pref ranking, probabilistic)	51.2%	25.5%	4.8%
Aggregate level MNL (CBC w/o no-buy in simulation)	56.5%	28.8%	9.8%
Aggregate level utilities MNL/CBC) + individual profiles of current product choice	49.5%	24.1%	7.9%
CAPMOD (pref ranking +choice threshold, OLS,1stPP)	40.5%	28.0%	6.0%
CAPMOD (pref ranking + choice threshold, OLS, probabilistic)	37.6%	21.3%	4.8%

Table 8: Comparisons of alternative new product share predictions

These results highlight a common problem with first past the post models (e.g CVA, ACA from Sawtooth and Conjoint Analyser, Simgraf from Bretton-Clark). They tend to over-state share for superior products and under-state share for inferior products when used to simulate current markets. There is every reason to be concerned with this when new products are being simulated. This is illustrated in Figure 3. In this case we are using our 1999 data to simulate the situation of a company having a new product in development and being confident about the eventual profile other than on one dimension. This happens to be an attribute that the conjoint study has shown to be important. The final outcome on this attribute will has a major effect on uptake. The traditional OLS based, first past the post model (Conjoint Analyzer) under-estimates at the low (inferior) outcome but, more importantly severely over-estimates relative to other simulation models at the high (superior) outcome. Amending the model to be probabilistic, rather than 1st past the post, reduces this low/high difference. However, we believe that, in this market, both share estimates are unrealistic. Maintaining the individual level focus throughout (from data collection, through data analysis to simulation) produces share estimates that appear to have greater basis of reality credibility, and are, intuitively more believable.

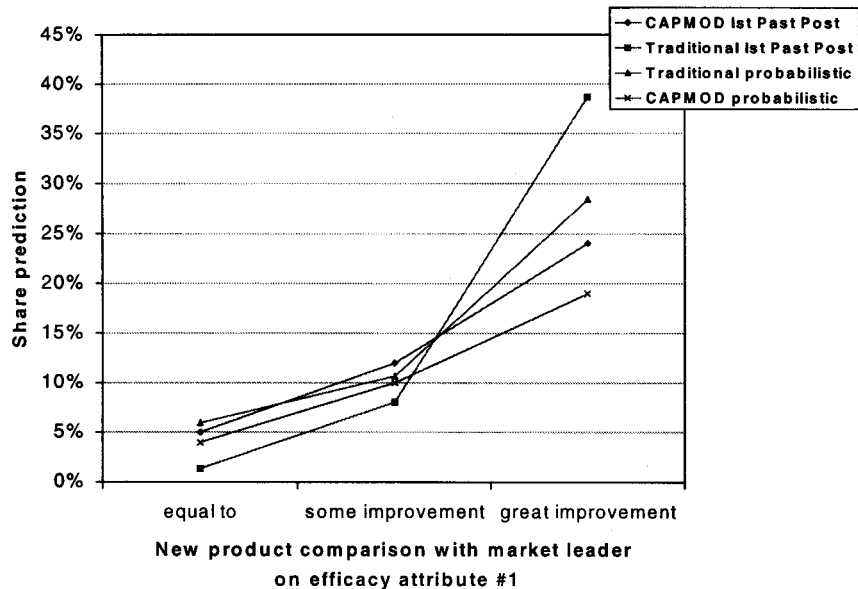


Figure 3: Comparisons of simulation results given alternative outcomes in new product profile

The CAPMOD simulation module, in this example, maintains individual level data throughout. All analysis and simulations are conducted at the individual level. This is not the case with pseudo thresholds such as the ‘no-buy’ option in discrete choice models such as CBC. Consequently, not only is the Independence of Irrelevant Attributes (IIA) problem not an issue but also that the source of new product share – in terms of current competitors, individual physician and patient types - can be identified. This principle can be applied to any conjoint design provided the choice threshold question has been appropriately worded.

CONCLUSIONS

Our early work comparing alternative simulation methods within conjoint models led to two main conclusions:

1. We cannot identify a clear “winner” from the standard preference/OLS and choice/MNL model comparisons. Even if all the limitations of and caveats that must be applied to our experiment were to be removed, we doubt whether this would ever change. So much must also depend on the situation.
2. New product predictions in pharmaceutical markets lack credibility due to an over-estimation apparent to the experienced marketer on presentation and confirmed post launch.

Through applying choice thresholds at the individual level we can:

1. Allow perceptions of current products to be “as is” for each respondent
2. Eliminate (or substantially reduce) reliance on external adjustment factors which, whatever the experience brought to their estimation, must have a substantial arbitrary element and, therefore, be questionable.

3. Remove the need to profile other than new products in simulations of new product introductions.

We also believe that as the basis is more realistic, the resulting new product predictions to be more reliable.

Additional research and experience is required to further test our hypothesis and the predictive capability of the CAPMOD concept. We are sufficiently encouraged by our experience over the last two years that we have committed to the ongoing development of the model and are currently applying it to a number of data sets including choice sets simulated from rank order data analyzed through CBC/HB. This allows for individual level utilities to be estimated for both main effects and interactions at the individual respondent level and, hence, capturing much of the best of traditional OLS- and MNL-based models with the added realism of individual current product images and maintained individual level choice thresholds.

CUTOFF-CONSTRAINED DISCRETE CHOICE MODELS

Curtis Frazier

Millward Brown IntelliQuest

Michael Patterson

Compaq Computer Corporation

INTRODUCTION

Traditional discrete choice conjoint analysis typically does not provide estimates of individuals' utilities since respondents do not provide enough choice data to accurately estimate their individual utilities. Rather, data from discrete choice conjoint are analyzed in the aggregate by pooling data across respondents, either at the total sample level or by various *a priori* segments. By treating choice data from all respondents as if it were derived from a single respondent requires the assumption of homogeneity across respondents – in other words, that all respondents have the same utility function and hence place the same relative “value” on each level of each attribute. Obviously in many cases this assumption is not tenable.

Introducing heterogeneity into the analysis and reporting of discrete choice data has several practical advantages. First, by using individual level utilities, it is possible to determine if different segments of respondents place more importance on certain attributes compared to other segments. For example among purchasers of notebook computers, one segment of respondents might value a notebook computer that offers the highest level of performance regardless of price, whereas another segment may value price over all other attributes. Individual level utilities also allow product bundles to be developed that are “optimal” in terms of likelihood of adoption. For example, a telecommunications provider might discover that individuals who have a strong preference for voice mail (which let's assume offers low margins) also value caller ID (which offers high margins) and hence develop a bundling strategy which offers the two products together and maximizes profitability. Individual level utility estimates may also improve model fit and predictive validity by allowing more accurate estimates of utilities.

Recently, several new approaches have been introduced in an attempt to introduce heterogeneity into the analysis of choice data including:

- Sawtooth Software's ICE (Individual Choice Estimation) Module
- Random parameters logit
- Hierarchical Bayesian methods
- Soft penalty cut-off models

Each of these approaches is briefly described below.

Sawtooth Software's ICE (Individual Choice Estimation) Module permits estimation of individual level utilities (Johnson, 1997). ICE can estimate these utilities by either using utilities from a Latent Class solution as a starting point or by estimating the utilities directly from choice data. When using the Latent Class solution, ICE estimates utilities by using weighted

combinations of the segment utilities. ICE does not, however, constrain these weights to be positive which produces more accurate estimates of individual's utilities.

Random parameters logit is a standard logit model where the utilities are allowed to vary in the population (McFadden & Train, 1998). This approach estimates the distribution of utilities across all respondents in the sample based on a variety of distributions (e.g., normally distributed, triangularly distributed, log-normally distributed, etc.). Use of random parameters logit has been shown to produce better fitting models compared to simple multinomial models (Brownstone, Bunch, Golob, & Ren, 1996)

Hierarchical Bayesian (HB) methods allow individual level estimates of utilities by combining information from individual's specific choices with the distribution of utilities across respondents. Rather than using the multinomial logit model to estimate utilities (at the aggregate level), HB uses the Gibbs Sampler. The Gibbs Sampler is a computer-intensive technique that utilizes random number generation to integrate the functions rather than traditional calculus derivations. The result is a set of utilities for each individual. HB has been shown to have better predictive validity compared to aggregate logit models (Garratt, 1998).

Soft penalty cut-off models permit respondent-level, "soft penalty" cutoffs to be incorporated into standard logit models (Swait, 1998). The general notion behind this approach is that when making purchases, consumers often think in terms of "cutoffs" or limitations. For example, when shopping for a car, an individual may say that they do not want to spend more than \$15,000. This maximum price is often flexible, however, so that they might be willing to spend \$15,500 or even \$16,000 if the car offers additional features. Paying more than originally intended creates some resistance however and makes the individual's decision more difficult. Since these limitations or exclusions are pliable they can offer insights into the decision making process. In addition, incorporating these "soft penalty" cutoffs into discrete choice models can result in a model that more accurately predicts a respondent's choices compared to aggregate logit models (Chrzan & Boeger, 1999). To implement the soft penalty cutoff model, a series of "penalty" questions is asked in addition to the discrete choice task and then a MNL model is estimated which includes aggregate level utility coefficients as well as penalty coefficients. Heterogeneity is thus introduced by taking into account each individual's responses to the penalty questions (see Appendix A for more detail).

STRENGTHS AND WEAKNESSES

ICE produces individual level utilities, is relatively easy to implement using Sawtooth Software's ICE module, and is computationally fast. It also produces fairly accurate results that have good predictive validity and that appear to compare closely to those obtained using HB (Huber, Johnson, & Arora, 1998). However, ICE requires that respondents complete a number of choice tasks (the ICE manual recommends at least 20) in order to generate reasonable utility estimates. ICE is also an extension of Latent Class and so it depends on the choice of segments and number of segments selected.

Random parameters logit can be implemented using commercial software packages (e.g., LIMDEP) however it does require that the user should have a good understanding of the model assumptions in order to choose the most appropriate underlying distribution.

HB produces individual level utility estimates that have good predictive validity and HB requires fewer choice sets than ICE to produce reasonable utility estimates. HB is also easy to implement using Sawtooth Software’s HB module. HB is computationally intensive however and can take several hours to estimate particularly with large data sets.

Soft penalty cutoff models are easy to implement by asking a series of questions to elicit respondent’s cutoffs (see Appendix A). These models have also been shown to produce better fitting models compared to aggregate logit in most cases (but not always) (Chrzan & Boeger, 1999) and they provide managerially relevant information about the proportion of individuals likely to reject certain attribute levels. However, utilities that are derived from soft penalty cutoff models often show unreasonable, counter-intuitive signs and including penalties in the model requires that the researcher increase the survey length since a cutoff question must be asked for each attribute.

The purpose of this paper is to make comparisons between Hierarchical Bayesian analysis, cutoff constrained choice models, and aggregate choice models in terms of their ability to predict holdout choice sets.

RESEARCH DESIGN

A disk-based, business-to-business study of a technology product was conducted with 450 respondents drawn from IntelliQuest’s Technology Panel. Eleven attributes of the product were identified as being important in previous research and were included in the present research. Five of the attributes had four levels and six of the attributes had three levels.

Respondents were presented with twelve choice tasks that presented four alternatives plus a “none” option. The design for the choice tasks was a randomized design that presented respondents with each of the attribute levels in a randomized fashion.

Respondents were also randomly assigned to one of four groups and were given two holdout questions that were unique for their particular group. The holdout questions were structured the same as the calibration choice sets (i.e., four alternatives plus none) and were randomly interspersed with the calibration choice questions. The holdout questions were not used with the calibration questions when estimating utilities, rather the holdout questions were used to test the predictive validity of the different analytic techniques.

Respondent counts and holdout identities were as follows:

Group	Number of Respondents	Holdouts
1	118	A & B
2	131	C & D
3	108	E & F
4	93	G & H

RESULTS

Utilities for the aggregate logit and penalty models were estimated using multinomial logit whereas HB utilities were estimated using Sawtooth Software's CBC/HB module.

A Share of Preference (logit) model was used with the HB and penalty model utilities to predict respondent's holdout choices since previous research has suggested that a Share of Preference model generally predicts shares more accurately than a First Choice model (Johnson, 1988). We "tuned" the scale factor (exponent) of the models in an attempt to minimize the Mean Absolute Error (Orme & Heft, 1999). However, these adjustments did not result in significant improvements in share predictions likely due to the fact that there were small differences in the variances of the predicted shares of the three models.

Two measures of predictive validity were computed, Mean Absolute Error (MAE) and Hit Rate. MAE is calculated by comparing the actual choice shares for each alternative for each holdout with the predicted choice shares. MAE provides a measure of predictive validity across respondents. The hit rate is the proportion of time respondents' holdout choices are correctly predicted. Hit rates give an indication of predictive validity at the individual level.

The table below shows the MAEs for the aggregate logit, HB, and penalty models. Comparing the MAEs for the aggregate logit (MAE = 3.50), HB (3.98), and penalty models (4.89), we do not find a significant difference between the three, $F(2,117) = 1.37, p > .05$.

In terms of hit rates, we find that HB (hit rate = 61%) more accurately predicts respondent's holdout choices than does the penalty model (hit rate = 44%), $Z = 7.22, p < .05$.

Analysis	MAE	Hit Rate
Aggregate	3.50	n/a
HB	3.98	61%
Penalty	4.89	44%

CONCLUSIONS

The results of this research reveal that in terms of predictive validity, individual level utilities estimated using hierarchical Bayesian methods predict respondent's holdout choices as well as, or better than, traditional aggregate logit models or models that incorporate individual's "soft penalty" cutoffs into the analysis. Specifically, we found that there was no significant difference between aggregate logit, HB, and penalty models when predicting choice shares at the total sample level (i.e., MAE). We did find that HB methods more accurately predicted which holdout question alternative an individual would select compared to penalty models.

These results suggest that when researchers are interested in examining heterogeneity in relation to discrete choice models, hierarchical Bayesian methods are likely to provide more accurate results than models that incorporate individual's "soft penalty" cutoffs. Soft penalty cutoff models can provide useful managerial information concerning which attribute levels individuals are likely to reject, and in previous research, they have been found to produce better fitting models compared to aggregate models. However, given that they increase the length of a

survey, often have coefficients which are the wrong sign, and appear to be less accurate at predicting individual's choices, soft penalty cutoff models seem to be less useful for introducing heterogeneity into choice models than HB methods.

REFERENCES

- Brownstone, D., Bunch, D., Golob, T., and Ren, W. (1996), "Transactions Choice Model for Forecasting Demand for Alternative-Fueled Vehicles," in S. McMullen, ed., *Research in Transportation Economics*, 4, 87-129.
- Chrzan, K., and Boeger, L., (1999), "Improving Choice Predictions Using a Cutoff-Constrained Aggregate Choice Model," Paper presented at the 1999 INFORMS Marketing Science Conference, Syracuse, NY.
- Garratt, M. J., (1998), "New Ways to Explore Consumer Points of Purchase Dynamics," Paper presented at the American Marketing Association Advanced Research Techniques Forum, Keystone Resort, Colorado.
- Huber, J., Johnson, R., and Arora, N., (1998), "Capturing Heterogeneity in Consumer Choices," Paper presented at the American Marketing Association Advanced Research Techniques Forum, Keystone Resort, Colorado.
- Johnson, R. M. (1997), "ICE: Individual Choice Estimation," Sawtooth Software, Sequim.
- Johnson, Richard M. (1988), "Comparison of Conjoint Choice Simulators—Comment," *Sawtooth Software Conference Proceedings*, 105-108.
- McFadden, D., and Train, K., (1998), "Mixed MNL Models for Discrete Response," working paper, Department of Economics, University of California, Berkeley, CA..
- Orme, B., and Heft, M., (1999), "Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results," *Sawtooth Software Conference Proceedings*.
- Swait, J. (1998), "A Model of Heuristic Decision-Making: The Role of Cutoffs in Choice Processes," working paper, University of Florida, Gainesville, FL.

APPENDIX A

Implementing soft penalty cutoff models

To implement the soft penalty cutoff model, we ask respondents a series of cutoff questions following the discrete choice questions to understand what constraints, if any, they might impose when making choices. Examples of cutoff questions for quantitative (e.g., price, processor speed for a PC) and qualitative attributes (e.g., brand, presence or absence of an attribute) are:

Quantitative cutoff question:

“The manufacturer is considering pricing the printer you just read about between \$200 and \$500. What is the highest price you would consider paying for the printer?” \$ _____

Qualitative cutoff question:

“Which of the following printer brands would you never consider purchasing (select all that apply)

- Hewlett Packard
- Lexmark
- Epson
- Canon
- I would consider them all

These cutoff questions are then incorporated into the discrete choice model as penalties. Thus for each attribute level, we create a variable that indicates whether (for categorical variables) or by how much (for quantitative variables) a choice alternative violates a respondent’s cutoff. All of these terms are included in the aggregate MNL model and we are able to estimate penalty coefficients for each quantitative attribute and each level of each qualitative attribute. For example, assume we are estimating the share of preference for a printer that costs \$450 and respondent A has a cutoff of \$350. The penalty for respondent A is therefore \$100 and this gets incorporated into the aggregate discrete choice model. The utility associated with the printer price is then multiplied by respondent A’s cutoff amount (\$100) when estimating respondent A’s total utility for a particular choice alternative. In this way, individual level penalty information is incorporated with aggregate level utility information.

COMMENT ON FRAZIER AND PATTERSON

Torsten Melles

Westfaelische Wilhelms-Universitaet Muenster

Dealing with heterogeneity is one of the most important and recently most popular issues in modeling choice behavior as in discrete choice experiments, both from a theoretical and substantive point of view. We need to understand the strengths and weaknesses of the several methods that can be used to introduce heterogeneity into the analysis of choice data. As Wedel and others (1999) point out we need more simulation studies and more empirical studies. Patterson and Frazier fill in this gap. They discussed four different methods that can be used to introduce heterogeneity and they tested the predictive validity of two of them. These models were compared to an aggregate model. Hierarchical Bayes provided better prediction of individual holdout choices than a soft penalty cutoff model.

The reason for this result lies in the rationale underlying the methods and their specificities. Hierarchical Bayes captures multiple (unknown) sources of heterogeneity, the soft penalty cutoff model incorporates one known source of heterogeneity. Moreover, there are two important differences between these methods: One regarding the statistical model used to estimate individual utilities and one regarding the additional information that is required from the respondents.

Hierarchical Bayes most often uses the Gibbs Sampler that is a special case of Monte Carlo Markov Chain methods to estimate utilities. A multinomial logit (MNL) model with a maximum likelihood method is used to estimate utilities in the soft penalty cutoff approach. As Patterson and Frazier pointed out Hierarchical Bayes has been shown to have better predictive validity compared to aggregate logit models. Some authors (e.g. Johnson, 1997) predict that Hierarchical Bayes may turn out to be the best way of analyzing choice data when processing times will decrease with faster computers.

Additional information is required from the respondents in the soft penalty cutoff model. The reliability and validity of the responses to these questions determine the efficiency of the information to provide insight into the heterogeneity of respondents. Due to only one cutoff question for each attribute these cutoffs may be unreliable. Moreover, the usefulness of this approach may depend on the heterogeneity in cutoff values between individuals and also depends on the number of attributes included in a study. The greater the heterogeneity in cutoff values, the worse will be the bias if cutoffs are not taken into account.

The number of attributes matters for two reasons: First, the additional questions increase the survey length and can lead to a higher amount of response error due to cognitive strain and motivational effects. The higher the number of attributes, the lower the expected reliability of responses to cutoff questions. Second, a more severe limitation is that cutoffs are imposed on each attribute. Of course, a decision maker may think in terms of "cutoffs" when reflecting each single attribute. But the main question is: Does he consider these cutoffs in choosing between multiattribute alternatives? Swait (2000) states that decision makers often violate self-imposed constraints when viewing a bundle of attributes. This leads him to considering cutoffs as to be "soft". The model he proposes is one that describes *individual decision making*. Problems may

arise when the utilities are first estimated on an *aggregate level* as it was done in the study of Patterson and Frazier. A short example may illustrate these problems: Suppose there are two individuals in the sample with different preferences and stated cutoffs (e.g. \$400) on an attribute (price). One individual (subject 1) may consider this attribute as important. Due to this, the difference between the values of two levels (\$200, \$500) is high. The other individual (subject 2) may estimate this attribute as less important and neglect it due to a high number of attributes in the multiattribute choice task. The hypothetical value difference would be zero in this case and the cutoff would be irrelevant. This is indicated by the solid line in figure 1. Estimating utilities on an aggregate level leads to a difference that can be higher than zero (see the dashed lines) and cutoffs that can have a significant impact. This estimation is biased because the heterogeneity is not taken into account. Because the true value difference between two levels is zero for subject 2, the bias regarding individual utilities corresponds to the value difference between two levels. The bias between \$400 and \$500 without a cutoff is indicated by e_1 . On this condition, introducing stated cutoffs can lead to a stronger bias. As price increases over the cutoff (\$400), the marginal disutility increases instead of being zero as should be the case if the attribute was not considered. Hence, the difference between the utilities that follows from increasing price over the cutoff is bigger than in the case of an aggregate estimation without introducing individual cutoffs. The same is true for the bias (corresponding to e_2). Using this estimation in order to predict individual holdout choices using full profiles would lead to a more severe deterioration than using an aggregate estimation without cutoffs.

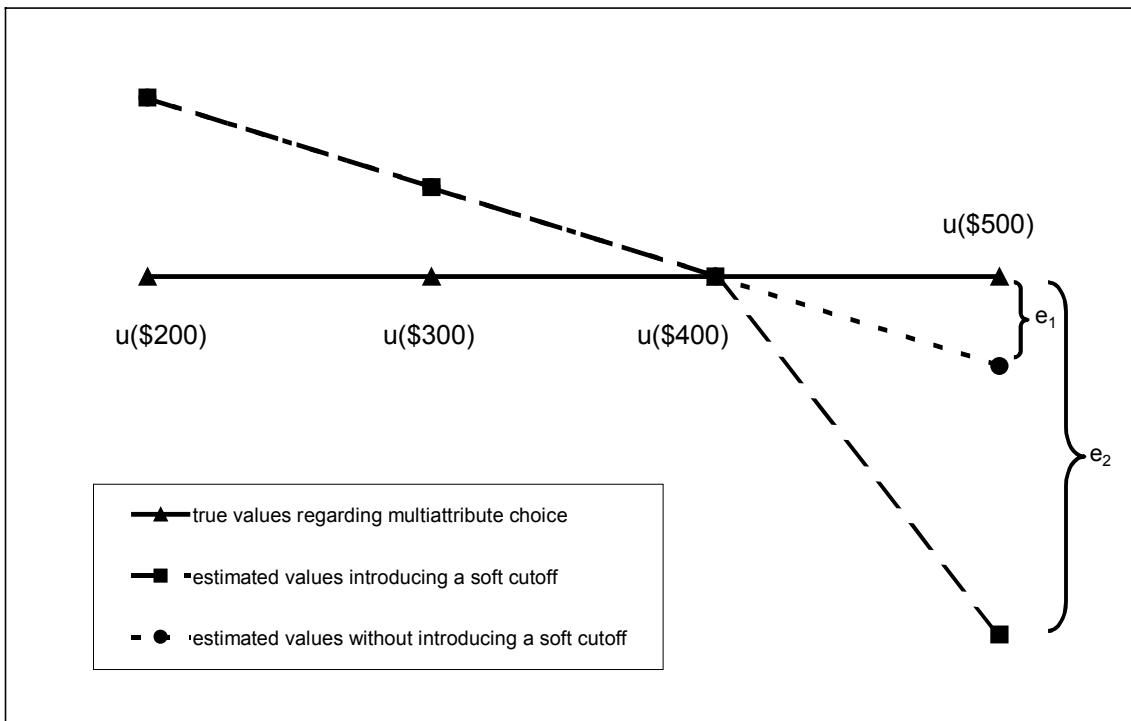


Figure 1: Introducing individual cutoffs into utility functions that are estimated on an aggregate level may lead to lower predictive validity regarding individual holdout choices.

Taking these differences between the methods into account, the question which method does better in predicting choice is still an empirical one. The results of Patterson and Frazier's study demonstrate the validity of two approaches under specific conditions. Further studies are necessary to test these methods under different conditions and compare them to other approaches. Moreover, Hierarchical Bayes and the soft penalty model are not rivals, they can complement one another. Hierarchical Bayes provides an efficient estimation of individual parameters that can also include individual cutoffs. On the other hand, soft penalty cutoffs may complement to any estimation method since it incorporates a known source of heterogeneity (Swait, 2000).

Finally, the following questions should be a guideline for future research:

- On which conditions do what respondent descriptor variables adequately capture heterogeneity?
- On which conditions should we use Hierarchical Bayes methods to estimate individual utilities and in which can we do without it?
- On which conditions does which level of aggregation capture heterogeneity adequately?

REFERENCES

- Johnson, R.M. (1997). *ICE: Individual Choice Estimation*. Technical paper, Sawtooth Software Inc.
- Swait, J. (2000). A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research B*, forthcoming.
- Wedel, M., Kamakura, W., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., Johnson, R., Lenk, P., Neslin, S. & Poulsen, C.S. (1999). Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10 (3), 219-232.

CALIBRATING PRICE IN ACA: THE ACA PRICE EFFECT AND HOW TO MANAGE IT

Peter Williams and Denis Kilroy
The KBA Consulting Group Pty Ltd

ABSTRACT

The tendency of ACA to underestimate the importance of price has been widely recognised over the last few years. Dual conjoint methodologies have been developed to address this issue. This paper proposes an alternative to dual conjoint. This new technique overcomes the “ACA price effect” by integrating ACA utility scores with the output of a series of explicit holdout choices.

Preference segments are developed from explicit choices made by respondents. A weighting factor for price utility is calculated for each segment. This is achieved by adjusting price utility so that ACA simulations closely align with holdout results.

Unadjusted ACA utilities match holdout responses very well for price insensitive respondents. However, significant adjustments are required for more price sensitive preference segments.

INTRODUCTION

Objectives

This paper has three objectives:

- To provide a brief description of the pricing problems that occur in ACA and to discuss potential methodologies for countering them
- To introduce a new method of adjusting ACA utility data to compensate for any inaccurate price signals that may exist
- To provide a starting point for further work and discussion

THE ACA PRICE EFFECT

Conjoint analysis has been utilised for a wide range of market research purposes over the last twenty years. One of its main applications has been to predict the potential demand for new products or services, and to establish the price that customers are willing to pay for them (Wittink et al, 1994).

A very popular conjoint technique is Adaptive Conjoint Analysis (ACA). ACA was introduced by Sawtooth Software in 1987 (Johnson, 1987), and is used extensively by marketing professionals in both the USA and Europe (Wittink et al, 1994).

One of the main advantages of ACA is that it allows the researcher to study more attributes than a respondent can evaluate at one time. This avoids the problem of “information overload” which can occur in full-profile studies when the number of attributes is greater than five or six (Green and Srinivasin, 1990). A typical ACA study uses between eight and fifteen attributes (Orme, 1998).

One of the most important outputs of a conjoint study is related to price. Understanding price utility allows researchers to:

- Forecast the effect of changes in price on customer demand for either a new or an existing product or service
- Quantify in dollar terms the benefits that individual product or service features provide to customers, and compare these with the cost to provide them

Over the last few years researchers have found that the importance of price is underestimated in many ACA studies (Pinnell, 1994; Orme, 1998). This is obviously of great concern, and a number of methods have been developed to counter this effect (hereafter referred to as the “ACA price effect”).

Most pricing studies make use of either traditional full-profile conjoint (for example Sawtooth Software’s CVA package) or choice-based conjoint (for example Sawtooth Software’s CBC package) techniques. However neither of these techniques is appropriate when the number of attributes to be studied is greater than about five or six. This problem has left researchers with a challenge to find a technique that has the ability to investigate large numbers of attributes, and still obtain accurate information about price utility.

A possible solution to the problem involves the use of dual conjoint methodologies (Pinnell, 1994; Sawtooth Software, 1999). If two conjoint studies are conducted in the one interview, then the first section can use ACA to obtain information about a large number of attributes, and the second section (utilising another conjoint methodology) can be used to obtain information about price and two to three other key attributes.

This paper proposes an alternative form of conjoint which integrates ACA utility scores with the outputs from a series of choice-based holdouts (CBH). The result of this is a set of calibrated utility scores that have had their price utilities adjusted to overcome the ACA price effect.

OVERVIEW OF PROPOSED METHODOLOGY

The study methodology is based on the combination of ACA and CBH. In this paper it is applied to a ten attribute study of approximately 1000 respondents which was completed for an electricity distributor. The decision process was highly involved and required respondents to carefully consider the impact that their decisions would have over both the short and long term. Interviews were strictly managed at central locations so that all respondents received a thorough introduction to the concepts and components of the questionnaire. This ensured that respondents carefully considered their options and made meaningful decisions. The prices of some options offered were significantly more than the respondents were currently paying.

ACA is best used when:

- The number of attributes involved in the study is larger than six
- The decision process being investigated is one in which consumers use substantial depth of processing (Huber et al, 1992)
- The number-of-levels effect needs to be minimised (Wittink et al, 1999)
- Individual level analysis is required

The study highlighted in this paper met all of these criteria. However due to the ACA price effect, ACA alone was not sufficient.

CBH (which is a series of explicit holdout choices) was structured so that the ACA price utility could be calibrated. As well as being the measure against which ACA was judged, it was also used to identify preference segments and to display results in a stand-alone manner.

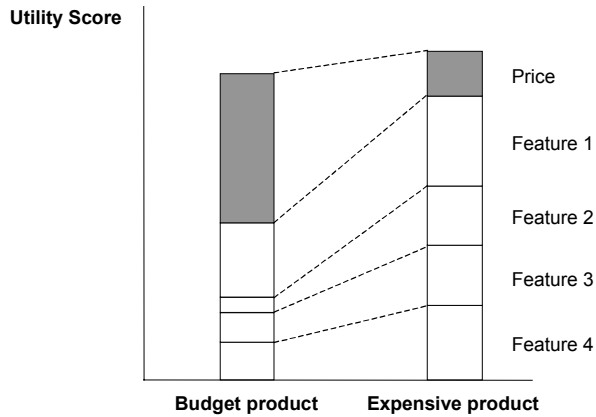
Since the core structure of CBH was for a client-specific purpose, the results presented in this paper are not as open to generalisation as would otherwise be the case. However many of the ideas behind the techniques demonstrated are still applicable in other situations, and could be used for calibrating “pure” dual conjoint studies. (This study is not a “pure” dual conjoint study, as utilities cannot be calculated from CBH in isolation.)

WHAT IS THE ACA PRICE EFFECT?

Recognising the ACA Price Effect

The clearest evidence of the ACA price effect is a simulation that severely over-predicts share for a feature-rich product (Pinnell, 1994). For a respondent who always tends to select the cheapest product available, the amount of extra utility that they place on a low price over a higher price level must exceed the utility from all the extra features that accompany the higher priced products. If it does not, then a simulation will incorrectly predict that the respondent will select the feature-rich product at a higher price.

As shown below, at the individual level the output of an ACA simulation may indicate that a respondent prefers the expensive feature-rich product to the budget product. However, this is a problem if the respondent actually selected the budget product when presented with an explicit choice. While the price level in the budget product has a significant amount of utility, it is not enough to counter the combined impact of the added features in the expensive product.



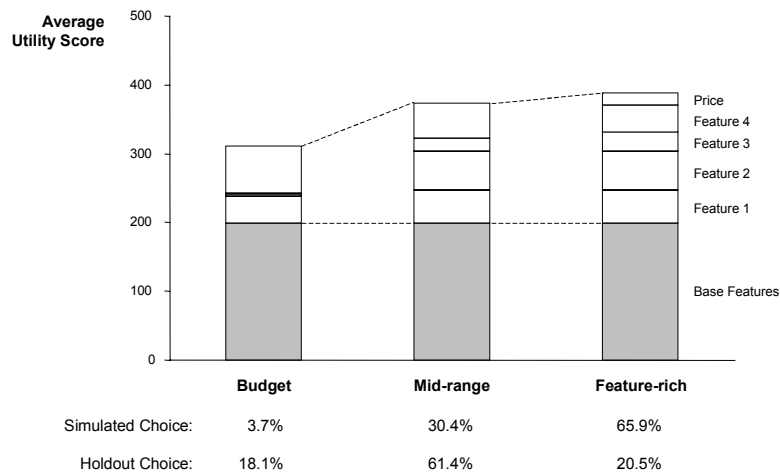
Holdout choices are an excellent means of identifying price underestimation. Both internal consistency and the underestimation of price utility by ACA can be examined by comparing ACA predictions with actual holdout choices (Johnson, 1997).

The ACA price effect can be best illustrated through the use of an example. An ACA simulation was compared with results from a holdout choice (consisting of three options) administered after the ACA questions were completed.

Example:

The three options consisted of a top-of-the-range feature-rich (expensive) product, a mid-range (mid-priced) quality product, and a budget-priced basic product. The results below show that while the simulation predicts that 66% of respondents would select the feature-rich product, only 20% of respondents actually did when presented with that holdout choice.

For this example, it is clear that using the results of ACA in isolation would result in an incorrect conclusion being drawn from the analysis, and ultimately would lead to flawed strategy development.



What Causes It?

There are a number of theories as to why the ACA price effect occurs. These include:

- Inadequate framing during importance ratings
- Lack of attribute independence
- Equal focus on all attributes
- Restrictions on unacceptable levels

Inadequate Framing During Importance Ratings

Perhaps the most difficult part of an ACA interview for respondents to understand and answer accurately is the section known as “importance ratings”. In this section, respondents are asked to indicate the level of importance they place on the difference between the highest and lowest levels of each attribute included in the study. The purpose of doing this is to refine the initial utility estimates before the trade-off section begins.

Assigning importance ratings to attribute levels in this way can be a difficult task – particularly when respondents do not know what other product attributes will be offered.

If an attribute related to the reliability of electricity supply has levels ranging from one blackout per annum to three blackouts per annum, then the respondent may rate the difference between one and three blackouts as being very important. If the next attribute tested is price, with levels ranging from \$0 to \$1,000 per annum, then the difference of \$1,000 would almost certainly be assessed as very important as well.

However, if both attributes are rated as very important, then ACA would initially assign utilities to these attribute levels consistent with the respondent being prepared to pay \$1000 to reduce blackout incidence from three to one per annum. Clearly if the importance of different attributes is to be captured accurately, respondents must be provided with some context so that they can frame their responses correctly.

If the respondent knew in advance that the next question was going to ask them about a large price difference, then they would probably realise that the number of blackouts per annum is not as important as they might otherwise have thought.

It can also be beneficial to order the importance rating questions so that they are structured in a similar manner to the “calibrating concepts” section. Therefore, the researcher may show what they believe is the most important attribute first and the least important attribute second.

While respondents should not be told this (as everybody may have different opinions), at the very least this ordering will help to better define the boundaries of what “importance” means. This “framing” will also help to ensure that the initial pairwise trade-offs are meaningful.

Lack of Attribute Independence

If an ACA study is conducted on the price, performance and colour of cars, then a respondent who rates the colour “red” highly because they believe that red cars go faster (ie. superior performance), has contravened the main-effects assumption. When a simulation is run, the effect of performance is double-counted because the respondent has attributed performance-related utility to both the colour and performance attributes.

This error serves to underestimate the importance of price. However, the effect can also work in reverse. If the respondent assigns a low rating to performance because they believed that a high performance car would always come at a high price, then the effect of price is effectively being double-counted, and its importance will be overstated.

Either way, these errors are preventable and should be countered by clear explanations and examples at the beginning of the questionnaire. For a very difficult questionnaire it may be worth including a short dummy conjoint exercise at the start that is not used for utility calculation. The extra time that this takes could well be insignificant compared with the increased efficiency with which the main exercise is completed.

Equal Focus on All Attributes

The partial-profile design of ACA forces respondents to consider all attributes. In a full-profile study, respondents may well focus on only those attributes that are more important to them (and take less notice of those attributes that are less important). Respondents are less likely to be able to employ simplification strategies such as this with ACA. Consequently, the importance of each attribute is likely to be more similar with ACA (Pinnell, 1994) than with full-profile techniques.

For example, if the price range tested is quite large, then price will naturally be one of the most important attributes. If ACA forces respondents to place more focus on other attributes than they would have in full-profile, then this will serve to underestimate the importance of price.

Restrictions on Unacceptable Levels

ACA permits the respondent to narrow the focus of a study by indicating which attribute levels are unacceptable. For most studies it is not appropriate for the researcher to allow price levels to be deemed unacceptable. In others it may be important to permit levels to be deemed unacceptable, but there is a risk that in doing this errors will be introduced.

It is difficult to ask respondents to rate a price level as unacceptable or not if they are not fully aware of the benefits of the product they are evaluating. This is particularly true in the case of new products or services. Many respondents may be quite clear about the maximum that they would pay for a new car or personal computer. However, it is much harder for them to put a limit on the price they would pay for products or services that they have never been exposed to, or that they have not been asked to consider before – such as value-added services provided by utility companies.

The problem for the researcher is two-sided. It is clearly undesirable for a respondent to be permitted to rate a large number of attribute levels as unacceptable at the start of the survey – particularly when demand for new or unfamiliar products or services is being studied. Nevertheless if respondents are not given the right to deem a particular price level to be unacceptable, then they may be forced to consider price levels that in reality they would never be prepared to pay.

If a respondent is not allowed to rate a price level as unacceptable, then this level will receive more utility than if it was rated as unacceptable. This will serve to understate the importance of price.

OVERCOMING THE ACA PRICE EFFECT

Quantifying the ACA Price Effect

The key to overcoming the ACA price effect is to be able to quantify it. Since the price utility calculated by ACA may be suspect, another methodology must be used which accurately assesses the impact of price.

One method for achieving this is through the use of “dual conjoint”. Dual conjoint can be described as two consecutive conjoint studies utilising different methodologies to obtain information on the same subject.

The first study may use ACA, with the second study usually focussing on price and two or three other key attributes. The first study enables the researcher to collect detailed information on a variety of attributes (which may or may not include price), while the second study can be used to calculate accurate pricing information.

Full-profile methodologies are often used as the second study in dual-conjoint projects. These methodologies include:

- Traditional full-profile conjoint (eg. CVA)
- Choice-based conjoint (eg. CBC)

The two studies can be either compared, or the results can be combined or calibrated to form one set of utilities. It must be noted that on many occasions researchers undertake dual conjoint studies to compare the different perspectives offered by ratings/rankings-based versus choice-based techniques. As each technique has different strengths and weaknesses, many researchers (if time and money permit) employ both techniques to “cover their bases”. However this usage of dual conjoint is not addressed in this paper.

This paper focuses specifically on the use of holdout choices (“Choice Based Holdouts” or CBH) for the second study. While CBH is not actually a conjoint methodology, it is used to calibrate conjoint utility scores from ACA. This differs from standard holdout choices which are purely used to check the predictive ability of the conjoint model.

At a general level, there are three advantages to selecting CBH as a method for countering the ACA price effect.

- CBH allows full control over the choices presented to respondents. It is undesirable to have restrictions (prohibited pairs) on the attribute levels that can appear when using traditional full-profile or CBC techniques. However restrictions are hard to avoid if certain combinations of attribute levels naturally tend to be associated with one another (eg. higher quality products should appear with higher prices so that the choices make sense to respondents). CBH allows the researcher to present realistic choices that may exist in the marketplace (Johnson, 1997). CBH cannot be used to calculate utilities, but can be used to calibrate those generated by ACA.
- CBH can be used in its own right. If the researcher is able to formulate choices that are realistic and meaningful, then the output from the choice questions can be presented (ie. results such as “when presented with three options, 28% of respondents selected option A”). While some managers may never really come to terms with the somewhat abstract

nature of “utility” derived from ACA, CBH is something that they can view as unambiguous. However such analysis is obviously restricted to the choices shown in the questionnaire, and the flexibility provided by ACA to simulate a variety of potential product offers is not available. If CBH is used as the primary method of research, ACA can be used to help explain why respondents made the choices that they did.

- The nature of CBH means that it is less time-consuming (and therefore cheaper) and less complex than many dual conjoint methodologies. Depending on the nature and purpose of the research, this factor may have a strong influence on the choice of methodology.

METHODS FOR CALIBRATING PRICE UTILITY

Introduction

In the absence of an actual purchase decision, most studies use CBH as a proxy for customer behaviour. Therefore any conjoint model which produces results significantly different to CBH has a problem. For example, if a CBH question has three options (feature-rich, mid-range, and budget), then an ACA simulation with three products should produce a similar outcome (assuming a “none” option is not appropriate). If the simulation severely over-predicts the share for the feature-rich product (relative to CBH), then the ACA price effect exists.

A number of methodologies for calibrating price utilities can be used:

- Compare the share predicted by ACA and CBH and apply a single weight to all respondents’ price utilities so that the ACA simulation and CBH results match at the aggregate level
- Identify utility importance segments using cluster analysis and use the method above to adjust each segment
- Use regression to adjust each respondent’s price utility individually to better predict their CBH results

However, there are a number of problems with each of these methodologies.

Comparing overall ACA simulations with CBH aggregate results does not identify any lack of internal consistency within the questionnaire. It is possible for the overall ACA simulation to match CBH results (to a satisfactory degree of accuracy), but at the individual level for the predictive validity to be much lower.

For example, in a study of 140 respondents, both ACA and CBH may predict that 100 respondents prefer product A, and 40 respondents prefer product B. A problem exists if the two methodologies do not agree on which respondents would select each product. If ACA and CBH only overlap in their prediction of preference for product A by 80 respondents, then the two methodologies only match for 80% of choices for product A, and for only 50% of choices for product B (the model performs no better than a simple coin toss).

Identifying segments that are influenced differently by the ACA price effect using utility importance is a difficult task. Not only do different respondents have different sensitivities to price, they may also be influenced differently by the ACA price effect. This makes identifying a homogenous segment very difficult.

Calibrating price at the individual level is theoretically the best way to align ACA and CBH. However the presence of any reversals (such as when a respondent appears to prefer a higher price level to a lower one) in the data makes calibration difficult.

If a weighting factor is applied to price utilities that are in the wrong order, then the magnitude of the reversal will increase. When these respondents are combined with other “clean” respondents, then aggregate utilities will appear more distorted than before the calibration took place. While the reversals could be artificially removed before calibration, this approach has implications that are beyond the scope of this paper.

Using CBH Segments

A potential segmentation methodology involves the use of CBH data. Examining the pattern of CBH responses can identify the importance of price to respondents.

A respondent who consistently chooses high-priced feature-rich products in CBH is unlikely to have the importance of price underestimated by ACA. Quite simply, price isn't particularly important to them. However, a respondent who makes all their choices based upon the cheapest price available is a strong candidate for the ACA price effect. Unless the lower price levels have extremely high utility, then simulations may predict that the respondent would choose a more expensive option if it contained enough features. If the mid-priced option is one price level more expensive than the cheap option, then the difference in utility between these two price levels must exceed the total difference in utility for each attribute that is enhanced in the mid-priced option.

A simple criterion for identifying price sensitivity segments is to count the number of times that the respondent chooses an option at a certain price (assuming that multiple CBH questions are asked on similar products). If each choice contains a high-priced (H), mid-priced (M) and low-priced (L) option, then a respondent may be considered relatively price insensitive if they choose the high priced option the majority of times. This allows the H segment to be identified. Similarly, respondents who choose the other two options the majority of times may be characterised as belonging to either the M or L segments.

Once segments have been identified, an extremely important check is to calculate the average utility importance for each one. If there is no obvious difference in utility importance between segments, then the study has very little internal consistency between ACA and CBH, and is of questionable value.

The main focus is to examine the ability of ACA to predict CBH results for each of the three price segments. When evaluating ACA's predictive ability, the most important factor to monitor is the percentage of respondents for whom ACA correctly predicts the CBH response. If the hit rate is high, it automatically follows that the predicted market shares will be similar. However, if the predicted shares are similar, it does not necessarily follow that the hit rate (ie. the internal consistency) is high.

The predictive ability of ACA for each of the segments is likely to be quite different. The H segment is not price sensitive, and will probably require little or no adjustment to the ACA utilities. Some adjustment will need to be made to the M segment, and it is likely that substantial adjustment will need to be made to the L segment. For the L segment, the utility for low price

levels will have to be high enough so that it outweighs all the extra utility available from more expensive options.

The above methodology is best illustrated through the use of a simple example.

Example:

Segments were identified by analysing the overall pattern of respondent preference to nine of the holdout choices presented. Of the ten attributes in the study, each of the nine key holdout choices presented comprised only five of these attributes. The remaining attributes were set to be equal (and were examined in separate holdouts not detailed in this paper). The utility importance determined using ACA for each of the three segments is shown below (using Points scaling).

The number of respondents in each segment was 231 in H, 592 in M, and 146 in L.

The utility importance of each of these five attributes is “pointing in the right direction”.

- The price attribute shows that the L segment places the most importance on price
- The importance of attribute 1 is flat across all segments, as attribute 1 was a fundamental attribute that wasn’t seen as a luxury
- The importance of attributes 2, 3 and 4 indicates that the H segment values them more than the M segment, which values them more than the L segment. These three attributes are all luxuries, so this trend in the level of importance is to be expected.

Attribute	Segment		
	H	M	L
Price	59	67	77
Attribute 1	45	45	45
Attribute 2	64	58	40
Attribute 3	28	24	21
Attribute 4	59	35	29

Price trend as expected

Similar importance due to common attribute levels

Additionally, of the three options presented in each choice, some of the attribute levels were common. This allowed ACA’s predictive ability to be scrutinised further.

For example, of the five attributes in each option, attribute 2 always appeared in the high priced and mid priced options at the best level, while attribute 4 always appeared in the mid priced and low priced options at the worst level.

- As attribute 2 is always set at the best level in the options used to identify the H and M segments, then it would be expected that it is one of the reasons why they chose those options. Respondents who fitted into the L segment would be expected to do so partly because they didn’t really value the best level of attribute 2. As can be seen by the similar

utility importance for attribute 2 in the H and M segments, ACA supports the preference segments identified.

- A further example of this is provided by attribute 4. Respondents who fitted into the H segment did so because they really valued the top level, while those who chose M or L didn't place as much value on that level. The ACA importance for the M and L segments demonstrates their relative indifference to attribute 4.

The price sensitive respondents (segment L) place more importance on the price attribute and less on the other attributes available. However when a simulation is run which combines price with these attributes, the magnitude of the price importance is not high enough to cancel out the utility available from the other four attributes. The importance of the price attribute reported by ACA is "too flat".

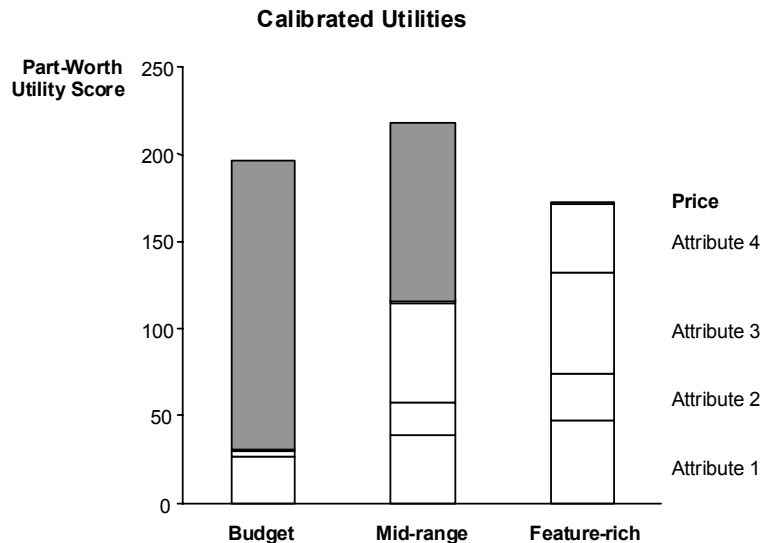
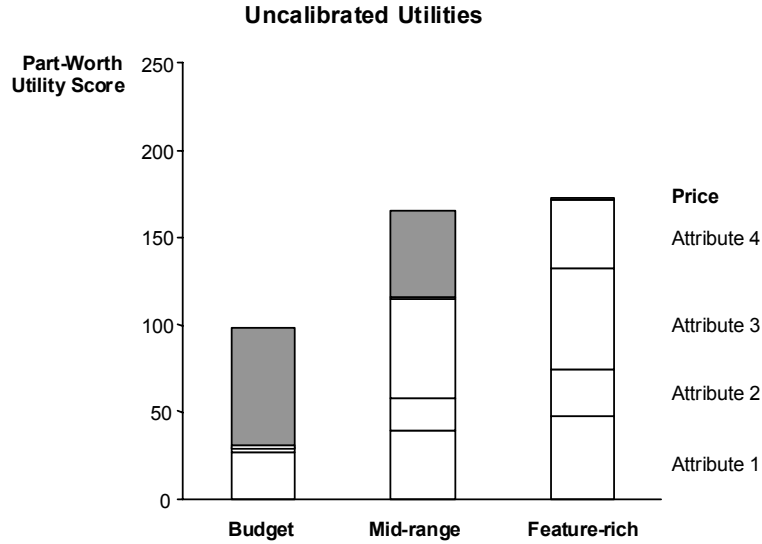
To adjust the utility levels so that the results from a simulation are similar to CBH results, the price utilities must be enhanced by a scaling factor. The factors found to maximise the predictive ability of ACA were:

- H: no adjustment needed
- M: scaling factor of 2
- L: scaling factor of 4 (or greater)

These factors were determined by looking at the hit rate of ACA when applied to the nine key CBH questions. The scale factors were adjusted until ACA best predicted (at an individual level) the CBH results for each segment. This process ensured that the integrity of each individual interview was maintained. After the three segments were analysed and calibrated, they were brought together.

The average utility function for one of the simulated CBH questions is shown below – before and after calibration. The average importance of the calibrated price attribute is about two times greater than before calibration. This is obviously significant – especially when calculations based on the dollar value of utility differences are performed.

When checking the validity of ACA simulations with CBH results, it is important to note that some options in CBH will have similar utility when simulated using ACA. If the utility of two options at the individual level is very close, then it is unreasonable to expect ACA to correctly predict the option that each respondent chose. A sensitivity allowance must be built into the calibration model to account for this effect.



Internal Consistency

A useful check of internal consistency is to repeat at least one of the CBH questions (Johnson, 1997). ACA results will not be able to predict CBH results if the CBH results themselves are inconsistent. However much care should be taken when repeating CBH questions. Respondents may detect that questions are doubled up, and therefore become suspicious about the motives of the questionnaire.

While in many cases it is desirable to repeat holdouts before and after the main conjoint task, this is not suitable for the ACA/CBH approach. As ACA relies on respondents making main-effects assumptions, any CBH questions shown before ACA will only serve to confuse them. CBH questions are designed to reflect reality, so they will contain options that have lots of features and high prices, versus others with minimal features and low prices. It is undesirable for

respondents to make any associations between levels (such as that a high price implies more features) before undertaking ACA, as this violates main-effects assumptions.

Another extremely important factor to consider is the order of CBH questions. If a respondent who prefers the feature-rich option is shown a choice which has this option at a high price, and then another choice which has the option at a low price, they will probably pick the feature-rich option both times. However, if they are then shown another choice with the feature-rich option at a high price once again, they may not select it, as they know that it is potentially available for the cheaper price. The respondent has had their preference “framed” by the range of prices previously shown.

OTHER ISSUES

Four other issues that must be addressed when designing an ACA/CBH survey are:

- Presenting meaningful choices
- The “number of attributes effect”
- The range of CBH questions shown
- Accuracy of utility estimates for non-price attributes

Presenting Meaningful Choices

CBH relies on presenting respondents with meaningful choices. It is obviously easier to construct realistic choices after a study is complete. For this reason, it may be worth running a pilot ACA study that can be used to help formulate choices. When the main ACA research program takes place, the researcher can be sure that the choices that they are presenting are appropriate and will provide meaningful information. Alternatively, the researcher may already have a strong idea of the particular product configurations that they are hoping to test.

Number of Attributes Effect

If the choices used in CBH are not full profile (ie. only contain a subset of the attributes used in the study), then the weighting factor to be applied to the price utilities may be influenced by the number of attributes present in the choice. The weighting factors used in the previous example were based on holdout choices that contained five attributes of varying levels, and five attributes set at a constant level (according to main-effects assumptions). However, the weighting factor would probably be different if only two attributes varied, and eight were kept constant. It is therefore important that CBH is structured to reflect reality as closely as possible.

Range of CBH Questions

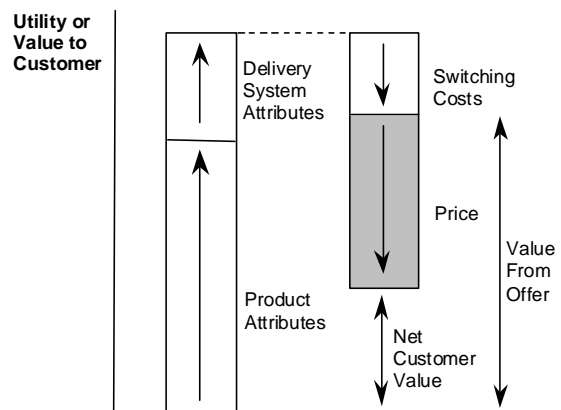
The range of CBH questions asked must be sufficient so that all price levels are covered. The calibration process effectively turns price utilities into a “plug” which adjusts a simulation so that it matches CBH results. If only one CBH question is asked, then it is possible to apply a weighting factor that implies that the respondent has a really strong preference for a particular option. While ACA is correctly predicting the option that the respondent chose, the calibration process has overwhelmed the fact that the respondent may have only just preferred this option. It is difficult to assess which option the respondent would have chosen if the pricing levels used

were slightly different, as a single CBH question gives no sense of how close the respondent was to choosing a different option.

However if multiple CBH questions are asked at different price points, then it is possible to assess the “strength of preference” that they possess for a particular pricing level. If ACA price utilities are calibrated so that simulations accurately predict a variety of CBH questions at different price points, then the researcher can be confident that the price utilities are quite robust.

Accuracy of Utility Estimates for Non-Price Attributes

Many of the problems with ACA that impact on the price utility estimation can also apply to other attributes. However the impact of these problems is likely to be different (and not as damaging) for non-price attributes. Price is a unique component of an ACA study, as in many cases it can be viewed as the only means of “taking utility away” from the respondent.



The product and delivery system attributes may be formulated based on 10-12 features which all add to the utility that the customer derives from a value proposition. If these attributes are enhanced, the net customer value will also increase unless other attributes are made less attractive. In most cases it costs a company money to provide extra product features. The price must be increased to cover this cost. This means that price is usually the only attribute that is used to subtract utility from an enhanced product offer (as switching costs are often too intangible to include).

The price attribute also comes to the fore when performing calculations such as the dollar value of attribute enhancements. The utility of price is the critical factor when converting the utility values of other attributes into more tangible units such as dollars.

While price may not be the only attribute which needs adjustment in an ACA study, it is often the attribute which most needs to be accurate. The effect of incorrectly estimating utility for other attributes is in many cases not likely to significantly impact on the findings of the study (although this of course is dependent on the purpose of the study). It is the unique role of price that makes it such an issue.

CONCLUSION

While there is much written about the merits of a dual conjoint approach, there is little documentation available on the mechanics of performing this technique. This is unfortunate, as

many researchers simply cannot condense the number of attributes to be tested down into a form that can be used exclusively by choice-based or traditional conjoint, and still meet their research aims.

The techniques used in this paper can be summarised as follows:

- In the absence of an actual purchase decision, CBH choices provide a strong basis for predicting respondent choices. By asking the respondent to choose between a range of options, real-world purchase decisions can be simulated. The pricing signals evident from these choices can be used to calibrate pricing signals emerging from the ACA study.
- The dual ACA/CBH approach is useful as it allows the researcher to present managers with the responses from CBH. ACA utility data can then be used to illustrate the drivers of these choices. If ACA is calibrated so that it accurately predicts CBH, ACA can be more confidently used to present results from hypothetical choice simulations that the respondent did not directly evaluate.

The dual ACA/CBH approach was developed to meet a specific aim for a particular project. In many ways it is a “fix”, and it may not be suitable for all projects. However as ACA will undoubtedly continue to be used to address projects with large numbers of attributes, it is important that researchers are able to achieve correct pricing signals. The dual ACA/CBH approach enables ACA to be used to develop robust strategies involving many attributes. Until other methodologies are further developed, this approach provides a sound basis for researchers and managers.

REFERENCES

- Green, P. E., and V. Srinivasan (1990), “Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice.” *Journal of Marketing*, 54 (October), 3-19.
- Huber, J., D. Wittink, R. Johnson, and R. Miller (1992), “Learning Effects in Preference Tasks: Choice-Based Versus Standard Conjoint.” *Sawtooth Software Conference Proceedings*.
- Johnson, R. (1997), “Including Holdout Choice Tasks in Conjoint Studies.” Working Paper, Sawtooth Software.
- Orme, B. (1998), “Which Conjoint Method Should I Use?” *Sawtooth Solutions*, Winter 1998.
- Pinnell, J. (1994), “Multistage Conjoint Methods to Measure Price Sensitivity.” Conference Paper, Advanced Research Techniques Forum, Beaver Creek, Colorado, June.
- Sawtooth Software (1999), “The CBC System for Choice-Based Conjoint Analysis.” Working Paper, Sawtooth Software, January.
- Wittink, D. R., and P. Seetharaman (1999), “A Comparison of Alternative Solutions to the Number-of-Levels Effect.” *Sawtooth Software Conference Proceedings*.
- Wittink, D. R., M. Vriens, and W. Burhenne (1994), “Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections.” *International Journal of Research in Marketing*, Vol. 11, pp. 41-52.

COMMENT ON WILLIAMS AND KILROY

Dick McCullough
MACRO Consulting, Inc.

INTRODUCTION

This paper, to my mind, epitomizes the Sawtooth Software Conference: it has real-world applicability, it is thorough and rigorous in its analysis and it is presented in such a straightforward and clear manner that it is accessible to almost any reader. It is a useful and interesting paper that raises many important issues.

The paper discusses two main topics:

- Sources of the ACA price effect
- A proposed adjustment to eliminate the ACA price effect

I will make a few comments on both these topics.

SOURCES OF THE ACA PRICE EFFECT

It would be desirable to minimize, or ideally eliminate, the ACA price effect by removing as much of the source of the effect as possible before making any post hoc adjustments.

One source of the effect identified by Kilroy and Williams is attribute additivity. Due to the large number of attributes that may be included in an ACA study, it is possible for a number of attributes, each with fairly low utility, to, in sum, overwhelm a price attribute that has a fairly large utility. For example, a product with nine attributes each with level utility of .2 (and a price level utility of .2) will have greater total utility than a product with a price utility of 1.5 (and nine attributes with level utilities of 0).

Attribute additivity can be a serious problem that will affect any trade-off method that attempts to accommodate a large number of attributes. One approach to counteract this effect is to limit the number of attributes included in the calculation of total utility (in the model simulation stage) for each individual to that individual's top six most important attributes. That is, if three products are being modeled simultaneously in a market share model and 10 attributes are included in product specification, for each product and each individual, include only those top six (out of the total 10) attributes in the calculation of total utility for that individual.

The rationale would be similar to that of limiting the number of attributes in a full-profile exercise to six: respondents cannot consider more than six attributes at a time when making a purchase decision. By limiting the number of attributes to six in the simulator, the attribute additivity problem would be diminished and the purchase decision process may be more accurately modeled.

Another source of the ACA price effect identified by Kilroy and Williams is attribute independence. Some attributes may interact with others, violating the main effects assumption of ACA. For example, in the importance ratings section of ACA, a respondent may be asked how

important the color red is versus the color blue in the context of new cars. Their true opinions, however, may depend on what type of car the color is applied to. They may prefer red on a high-priced car (such as a sports car) and blue on a lower priced car (such as a family are).

This is an extremely serious problem for all trade-off methodologies that involve some form of direct questioning or self-explicated scaling, not just ACA. The larger question that should be raised is whether self-explicated scaling is appropriate for all types of attributes. Kilroy and Williams have identified a problem with attributes that are dependent on other attributes, i.e., interact with other attributes. But can we determine if there are other types of attributes that are also inappropriate for self-explicated scaling? Are there other, operationally convenient ways to characterize inappropriate attributes? This issue deserves additional attention in the literature and I am very happy that Kilroy and Williams have raised it here.

Kilroy and Williams also cite attribute framing as a potential source of the ACA price effect. Without knowing what attributes are coming next, a respondent might give the strongest importance rating to the one presented first. For example, if the price attribute follows any other attribute, i.e., is not the first attribute to be rated in the importance section, then it may be rated as only equally important to another attribute that the respondent, in reality, does not feel is as important as price.

A simple antidote to this problem would be to review all attributes with the respondent prior to conducting the importance rating exercise. I believe that in most commercial applications that this would be feasible with the possible exception of telephone surveys.

PROPOSED ADJUSTMENT TO ELIMINATE THE ACA PRICE EFFECT

Given an ACA price effect, the authors have developed a surprisingly straightforward method for recalibrating the price utility to more accurately reflect the magnitude of respondent price sensitivity.

Their approach is to adjust each respondent's price utility so that predicted choice optimally matches a set of choice-based holdouts. They segment the sample population into three groups: those that need no price adjustment (that is, those whose predicted choices closely match their holdout choices), those that need some adjustment and those that need a lot of adjustment. They chose not to recalibrate price utility at the individual level due to the high incidence of reversals commonly found in conjoint data.

Reversals in conjoint data are commonplace, often involving up to 40% of the total sample. Like attribute independence, this issue is not unique to ACA. If there are reversals in the data set, it appears to me that they can be caused by one of only four factors:

- The data accurately reflect the respondent's values (and we simply are unwilling to understand or accept the possibility)
- Within attribute level variance is large due to the respondent being confused or fatigued when answering the survey, causing unstable level utility estimates
- Within attribute level variance is large due to limited sample size or experimental design issues, causing unstable level utility estimates
- There is an anomaly in the utility estimation algorithm

Whatever the cause of the reversals, recalibrating at the segment level does not avoid the problem, it just ignores it. The reversal respondents are included in the three segments and receive the same adjustment as the other respondents. I have no simple solution to this significant problem. I suspect that a good percentage of what we typically call reversals is simply accurate reflections of human behavior. Humans are clearly and frequently irrational in their buying behavior (and in every other aspect of their lives as well). Rather than attempt to force respondents to be rational just so that our models perform better, I suggest we look for ways to better model their sometimes irrational behavior. I also suspect that much of reversal data is due to confused or tired respondents. Making the interview as simple and brief as possible may help minimize reversals. I would like to see more research into the cause of reversals and possible ways to handle reversals in conjoint data sets without constraining respondent answers to conform to our assumptions.

The choice-based holdouts on which the price utility recalibration is based varied five of the 10 attributes and held five constant. The authors correctly point out that the recalibration scalars may be affected by the number of attributes included in the holdouts. For example, if only two attributes are included in the holdouts, the price recalibration scalar will most likely be smaller than if eight attributes are included because the attribute additivity problem will be greater with eight attributes than with two.

This appears to me to be a very serious problem with their proposed approach because the ACA simulator is designed to include all 10 attributes. Thus, one could recalibrate price to optimally predict holdouts and still underestimate price in the simulator. One possible solution would be to include all attributes in the holdout exercise but more often than not there would be too many attributes in the study to make this approach practical.

The suggestion made earlier for addressing the attribute additivity problem appears to me to also be a potential solution to the number of attribute holdouts problem as well. If the number of attributes included in the holdout exercise is six and if the simulator selects the top six attributes per person to calculate total utility, the recalibration scalar will be near the appropriate magnitude as long as the net relative importance of the six attributes in the holdout exercise is relatively similar in magnitude to the net relative importance of the six attributes selected in the simulator.

The authors make a very strong case for selecting price as the most important attribute to recalibrate. I would, in general, strongly agree. However, I suspect that there may be other attributes that would also be excellent candidates for recalibration, depending on the issues at hand. Although brand was not included in the Kilroy and Williams study, it is commonly included in conjoint studies and would be a strong candidate for recalibration because of the obvious lack of attribute independence coupled with its typically high degree of importance. It would be interesting to explore possible ways to simultaneously recalibrate two or more attributes, using the Kilroy and Williams approach.

It would also be interesting if guidelines could be developed to assist the practitioner in determining:

- The ideal number of holdout tasks needed for recalibration
- Which product configurations to include in the holdout tasks
- How many products to include in the holdout tasks

SUMMARY

The choice-based holdout tasks have appeal for numerous reasons:

- Fills client need for realistic alternatives
 - Increases model credibility in client's eyes
 - Powerful presentation/communications tool to unsophisticated audience
 - Cheaper and simpler than dual conjoint
- The Kilroy and Williams method for removing the ACA price effect is:

- A useful technique
- A sound approach
- Easy to understand
- Relatively easy to apply

Overall, I found this paper very useful and interesting. The authors raise many important issues and provide a practical solution to a significant shortcoming in many ACA models.

USING EVOKED SET CONJOINT DESIGNS TO ENHANCE CHOICE DATA

Sue York and Geoff Hall
IQ Branding (Australia)

BACKGROUND

This case study is based on a Brand Price Trade-Off study. In the study, the brands, which needed to be included, were drawn from the premium, sub-premium, mainstream and budget categories of the market, and in addition, a large range of pack sizes needed to be included. As a consequence, a number of questions arose during the design stage of the project.

Specifically, there were concerns about how realistic and reasonable it would be to ask respondents to trade-off brands from the different market categories. In particular, we had concerns about how realistic it was to place respondents in a situation where they may be asked to make a 'trade-off' between brands from the extreme high end (the premium category) and the extreme low end (the budget category) of the market. These concerns were exacerbated by the need to include a large range of pack sizes in the research. This potentially could create a situation in which a respondent may be asked to trade-off brands from different categories in quite different pack sizes, which may result in the different 'products' in the choice tasks having large differences in their prices. The key concern being that the choice set given to the respondent may not represent the choice situations that occur in the real world, and thus impact on the validity and quality of the data.

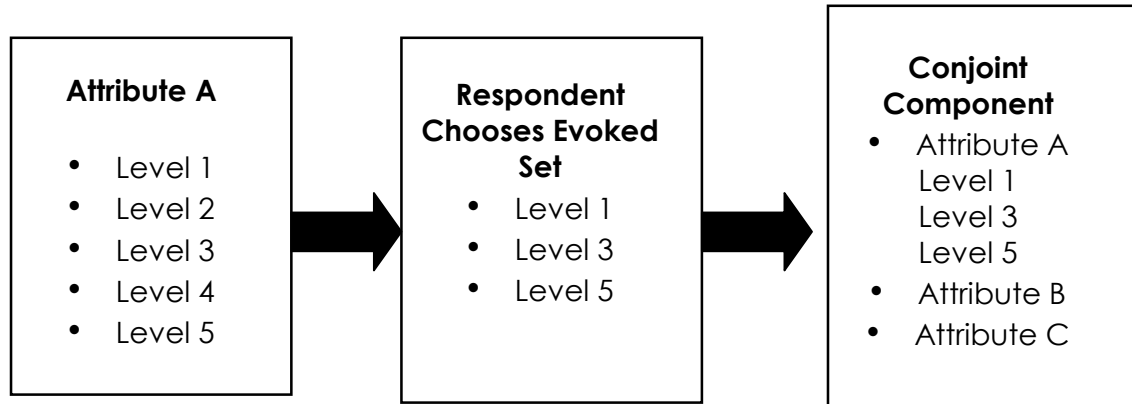
Whilst generating and discussing potential solutions, what was considered to be a strong alternative approach to the standard Choice Based Conjoint approach emerged. The hypothesis was that an 'Evoked Set' choice design would be able to address our concerns, by tailoring the choice exercise to only include the respondents' evoked set of both brands and package types. In effect, this design would create choice tasks for the respondent that were of relevance to them, and therefore, realistic. We then decided to test this hypothesis in a real life experiment.

AN EVOKED SET DESIGN – A DESCRIPTION

An Evoked Set design is an adaptation of a standard conjoint design derived from the marketing concept of consumers having 'Evoked' or 'Consideration' sets.

An Evoked Set design is one in which respondents choose those levels of an attribute which are most pertinent and only these are carried forward to the conjoint design, as is shown in Diagram 1 overleaf.

Diagram 1 – Evoked Set Design



EXPERIMENTAL APPROACH & OBJECTIVES

The experiment was designed to enable the comparison of the results obtained from two different conjoint designs, referred to as either i) the Standard design or ii) the Evoked Set design. The differences between these are detailed later in this paper.

The study had two key objectives.

The first objective was to assess whether the use of an ‘Evoked Set’ conjoint design offers improvements in either:

- i) Aggregate level utilities;
- ii) Individual level utilities derived using Hierarchical Bayes;

relative to a standard Choice Based Conjoint (CBC) design.

Two evaluation criteria were used to assess the performance of the designs. These criteria were the ability of the conjoint utilities to predict the:

- I) First choice in the hold-out tasks, reported as the mean hit rate; and
- II) Preference shares relative to the hold-out tasks, reported as the mean absolute error (MAE).

The second objective was to examine the effect of utilising “Unacceptables” information to tune utility values.

The impact of this was evaluated via improvements in First Choice Hit Rates.

EXPERIMENTAL DESIGN

At the data collection stage respondents were allocated to one of two matched samples. Respondents in the two samples completed essentially the same questionnaire, which varied only in the form of the conjoint trade-off design they completed. One sample completed the Standard CBC design, while the other completed the Evoked Set design. Respondents in the two samples

completed the same hold-out tasks and made ‘Unacceptables’ judgements for two of the attributes – brand and package type.

Two alternative approaches were then used to code the choice data from the Evoked Set design into a file format suitable for analysis in Sawtooth Software’s CBC program and CBC/HB module. This created two individual choice data files for the Evoked Set design.

Aggregate logit and individual level HB utilities were then calculated for the two Evoked Set design choice data sets and the Standard design data set. Comparative analysis of the results, in the form of first choice hit rates and the mean absolute error of share of preference predictions, was then conducted on the three choice data sets. As a final step, the ‘Unacceptables’ judgements were then analysed to determine whether these can be used to tune the utilities to improve the first choice hit rates.

CONJOINT DESIGN

The basic conjoint design was the same for both the Standard and Evoked Set designs.

Study Attributes and Levels

The base design included three attributes – Brand, Package (Type & Size) and Price. The Brand attribute contained 14 levels and the Package attribute contained 12 levels. The Price attribute included 13 levels in the Standard Design and 5 levels in the Evoked Set design.

Price was specified as being ‘conditional’ on the level of the package attribute; that is, different minimum and maximum price values were used for different package types and sizes.

The Standard design included 13 levels of price, while in the Evoked Set design the price attribute contained 5 levels. The minimum and maximum price points were the same for both designs. The number of levels for the price attribute differed for each design so that the number of levels of each attribute, within the individual design, were approximately equivalent. This was done to minimise a potential number of levels effect. To accommodate the difference in the number of price levels between both designs, price was treated as a linear variable or ‘linearised’ in the estimation stages.

Number and Complexity of Tasks

The number and complexity of tasks were the same for both treatments.

Each respondent completed 15 choice tasks, with each task containing 3 alternatives and respondents chose one of these. Each alternative included one (1) level of each attribute. A “None” option was not included.

Design 1 – Standard Design

In the standard design, all levels of the three attributes were included in a standard CBC Conjoint design, which utilised the conditional pricing facility.

Design 2 – Evoked Set Design

The second or Evoked Set design is a modification of the standard design in which respondents choose an Evoked Set of both Brands and Package type / size, independently, prior to the trade-off exercise. The respondents’ Evoked Set is determined by asking the respondent to

imagine they are in a retail outlet about to purchase this product and are asked to indicate which brands (or packages) they are most likely to purchase. The Evoked Set members, for both attributes, are the only levels (of these attributes) which are included in the trade-off exercise.

The Evoked Set methodology used in this study is a straightforward one, with the number of members of each Evoked Set (Brand & Pack) fixed at seven (7) members. This approach was used so that the design itself, and the comparison of the Evoked Set design with the Standard design, was less likely to be corrupted by other issues, such as the number of levels effect.

This design was programmed using Sawtooth Software's Ci3 program to produce a randomised design, with conditional pricing. The choice data was later coded into a 'CHO' type file for analysis in Sawtooth Software's CBC package and the associated Hierarchical Bayes module. A description of the file structure can be found in the CBC User Manual Version 2.

An extension to the assessment of the performance of this design, was an exploration of alternative methods of coding the choice data. This paper examines two alternative treatments. The first approach (referred to as Evoked Set Coding Option 1) includes information about the choice tasks the respondent was shown and the subsequent choices made. The potential deficiency of this approach is that it fails to 'communicate' to the algorithm that for a particular individual (respondent) the brands or packages not included in the Evoked Set are deficient to those included in the Evoked Set.

In an attempt to overcome this potential problem a second method for coding the data was examined. This approach communicates additional information about which brands and package types are not included in the Evoked Set.

This was achieved by including two additional choice tasks. The first additional task is used to communicate which brands were not included in the Evoked Set. The task includes all the levels of the brand attribute which were not in the respondent's Evoked Set and utilises the "None" option to indicate that the respondent did not choose any of these brands. Thereby, indicating that the brands that are not in the respondents' Evoked Set are of lower worth to the respondent than those in the Evoked Set. Within this additional task, the levels of the other two attributes were held constant. The second additional task is essentially the same as the first, but communicates which package types were not in the respondents' Evoked Set.

HOLD-OUT TASKS

The hold-out tasks are the internal measure against which the conjoint predictions are evaluated. The same hold-out tasks were included in both the designs. The hold-out tasks contained 14 options, with no "None" option available. The tasks were designed to be deliberately difficult to predict, to ensure that they discriminated well between individuals. Each task included every brand once, at its average market price, for a given package type.

'UNACCEPTABLES'

Both designs included a section in which respondents were asked to indicate which, if any, brands or pack types and sizes they considered to be 'unacceptable'. This was determined by asking respondents which brands or pack types and sizes (independently) they would never consider buying under any conditions.

These were included to allow an investigation as to whether ‘unacceptables’ data can be used to tune conjoint utilities to enhance the predictive ability of the model.

THE SAMPLES

The samples for the two designs were matched on :

- i) Key market demographics;
- ii) Product consumption levels;
- iii) Brand preferences; and
- iv) ii) & iii) within key demographic segments and sub-segments.

The sample sizes for both designs were in excess of 350.

RESULTS

A. First Choice ‘Hit Rates’

The performance of the alternative designs was firstly assessed using First Choice ‘Hit Rates’, that is, the ability of the utilities to predict the first option chosen in the hold-out tasks.

Each of the hold-out tasks contained 14 alternatives, as such, if we were to rely on chance to predict the results of the hold-out tasks we would have a 7.1% chance of correctly ‘guessing’ a respondent’s first preference. In this way, hit rates also provide a measure of how much better the utilities enable us to predict claimed market behaviour than chance alone.

1. Aggregate Logit

We would not anticipate aggregate logit, given that it is of an aggregate nature, to be particularly good at recovering individual level preferences, such as those obtained from the hold-out tasks. However, it is of interest to know which of the designs performed better at this level.

The first choice hit rates detailed in the table below, show that the Standard Design achieved a statistically significantly higher mean hit rate than the Evoked Set design, irrespective of the coding treatment.

Table 1: Aggregate Logit First Choice Hit Rates for Alternative Designs and Data Treatment Options

	Standard Design	Evoked Set Coding Option 1	Evoked Set Coding Option 2
Mean Hit Rate	0.18 *	0.12	0.12

2. Hierarchical Bayes

The individual level utilities were estimated using Sawtooth Software’s CBC/HB module. A total of 30,000 iterations were performed, with 20,000 iterations before saving and 1000 draws saved per respondent.

The results indicate that across both designs and the two coding treatment options Hierarchical Bayes provided statistically significant improvements in the mean hit rates achieved by Aggregate Logit.

The Evoked Set design, coded to include the additional information of which brands and packages were not included in the Evoked Set, (Coding Option 2) achieved a statistically significantly higher hit rate than the Standard CBC design and performed 4.6 times better than chance alone.

Table 2: Hierarchical Bayes First Choice Hit Rates for Alternative Designs and Data Treatment Options

	Standard Design	Evoked Set Coding Option 1	Evoked Set Coding Option 2
Mean Hit Rate	0.27	0.29	0.32 *

B. Share of Preference Predictions

As each respondent completed the same fixed hold-out tasks it is possible to compare the share of preference predictions to the hold-out task results. Share of preference predictions were calculated for both designs and the two coding treatment options using both the Aggregate Logit and Hierarchical Bayes utilities.

1. Aggregate Logit

The aggregate logit predictions show that the Standard Design has a lower mean absolute error relative to the hold-out tasks, than either of the Evoked Set Coding Options.

Table 3: Aggregate Logit Mean Absolute Error of Share Predictions Relative to Hold Out Tasks for Alternative Designs and Data Treatment Options

	Standard Design	Evoked Set Coding Option 1	Evoked Set Coding Option 2
MAE	2.1	4.2	3.9

2. Hierarchical Bayes

The share of preference predictions derived from the Hierarchical Bayes utilities indicate that the standard design achieves a lower mean absolute error than the Evoked Set design, under either coding option. The second point of note is that the use of Hierarchical Bayes offers a substantial improvement in the MAE level for the Evoked Set Coding Option 2, which declines from 3.9 to 2.87.

Table 4 : Hierarchical Bayes Mean Absolute Error of Share Predictions Relative to Hold Out Tasks for Alternative Designs and Data Treatment Options

	Standard Design	Evoked Set Coding Option 1	Evoked Set Coding Option 2
MAE	2.1	3.7	2.9

C. 'Unacceptables'

The 'unacceptables' judgements appear to have low reliability. Nine percent (9%) of respondents chose a brand in the hold-out tasks that they indicated to be unacceptable to them, while 29% chose a package type from the hold-out tasks that they considered unacceptable.

Using 'Unacceptables' Data to Tune Conjoint Utilities

Analysis of the unacceptables data for both brand and package type and size indicates that in this particular study the use of unacceptables data to tune the utilities may be of greater value and more appropriate for the brand attribute. Some 85% of respondents indicate that at least one brand is unacceptable to them, compared to only 50% of respondents for the package type and size attribute. In addition, (as discussed above) there is a higher level, in relative terms, of reliability for the brand unacceptables judgements than those for package type.

On this basis, a two-stage approach has been used in the process of using unacceptables data to tune the conjoint utilities. In the first instance, only the information about unacceptable brands was used to tune the utilities, while in the second instance information about both unacceptable brands and pack types was used.

Two alternative approaches for using the 'unacceptables' information to tune the utilities were examined.

- i) The first approach was to adjust a respondent's utility for an 'unacceptable' brand or pack type to the value of -9.9999 . This approach replicates that of Sawtooth Software's ACA (Adaptive Conjoint Analysis) package.
- ii) The second approach was a more extreme treatment in which the utility for an 'unacceptable' brand or pack type is adjusted to an extreme negative value, in this case -9999.99 .

The results, shown in Table 5 overleaf, show that for the Standard Design the two different approaches for using the 'Unacceptables' data to tune the Hierarchical Bayes utilities offer statistically significant improvements in the achieved mean hit rates. Similarly, in the Evoked Set Design (Coding Option 2) tuning of the brand utilities offers an apparent improvement in the mean hit rate, although this is not statistically significant. In this design, the tuning of the pack utilities actually decreases the mean hit rate, but not significantly.

The results suggest that using 'Unacceptables' data to tune utilities is of value when either the attribute has a very high proportion of respondents finding at least one of the levels unacceptable, and / or respondents display a high level of consistency between claimed unacceptables and claimed product choices.

Table 5: First Choice Hit Rates for Alternative Unacceptables Tuning Approaches (Hierarchical Bayes Utilities)

Mean Hit Rate	Standard Design	Evoked Set Design Coding Option 2
No Unacceptables Tuning of Utilities	0.27	0.32
Tune Utilities for Unacceptable Brand (-9.9999)	0.29 *	0.33
Tune Utilities for Unacceptables Brands & Packs (-9.9999)	0.29 *	0.31
Tune Utilities for Unacceptable Brand (-9999.99)	0.29 *	0.33
Tune Utilities for Unacceptables Brands & Packs (-9999.99)	0.29 *	0.31

CONCLUSIONS

'Unacceptables' Judgements

The findings suggest that there may be gains in the predictive ability of choice models to be achieved by using 'unacceptables' judgements to tune conjoint utilities. However, the more immediate concern is examining alternative approaches for collecting this information, improve the reliability of such judgements, and therefore, their usefulness for such a purpose.

The Evoked Set Design

The results of this study indicate that using an 'Evoked Set' design can be useful if predicting respondents' first choices is the objective. In this study, the Evoked Set design, using Hierarchical Bayes to estimate individual utilities achieved better, and good, first choice hit rates than the Standard CBC design. It did not, however, produce better share of preference predictions.

The approach can also be of use if many levels of an attribute need to be studied.

If an Evoked Set approach is utilised, the use of Hierarchical Bayes to estimate individual level utilities is recommended. In this study, Hierarchical Bayes achieved better first choice hit rates and share of preference predictions than aggregate logit for the Evoked Set design.

This paper examined two alternative approaches for coding the choice data collected using an Evoked Set approach. The option, which is referred to as Coding Option 2, is the superior alternative for prediction of either first choices or shares of preference. In this approach, additional "tasks" are added to the respondents' choice tasks in the coding phase. These tasks are used to indicate which attribute levels the respondent has not included in their Evoked Set, and as such are implicitly of lower utility to the respondent than those that are included.

At this point, the Evoked Set approach has been examined from an internal validity perspective. The next recommended step would be to conduct an external validation, comparing

the results of an experiment such as this, with either market share and scanner data, or both if possible. It would also be beneficial to ascertain if additional studies could replicate the findings of this one. In the future, it may be of benefit to evaluate the performance of the approach in different markets, with different market structures and in different product categories.

REFERENCES

CBC User Manual, Sawtooth Software, Inc., Sequim, WA.

COMMENT ON YORK AND HALL

Bryan Orme

Sawtooth Software, Inc.

Sue has done an impressive amount of work here. The Ci3 programming, data processing and HB runs involved in her research take considerable time and talents. All in all, I think her findings should come as good news. Researchers with the ability to program adaptive, customized CBC questionnaires may get improved results (at least, in this case, with respect to individual-level hit rates). But the standard procedure with CBC also works well, and avoids the added complexity. It seems likely, however, that as the number of levels in an “evoked set” attribute grows quite large (say, two-dozen brands), the added value of evoked set designs increases.

Reducing the design by customizing attribute levels taken into conjoint evaluations has a rich history. Researchers have been doing it since the 1980s with Adaptive Conjoint Analysis (ACA). The question has always remained as to how to treat the levels not taken into the conjoint evaluations. After all, we need to estimate utility values for those. Sue’s research (along with other previous work) casts doubt on whether “unacceptables” questions can consistently work the way we hope they should. For “unacceptables,” the default ACA value of -9.999 has seemed to be too extreme for Sue’s data set, as well as for ACA in general. For the “most likely” levels, simple interpolation or extrapolation is the default ACA procedure to compute the utilities. Coupled with a clever coding procedure, HB seems to have done a nice job with Sue’s data set, using information from other respondents to help estimate the missing levels for her experiment. The ACA/HB procedure by Sawtooth Software also generally does a better job than the traditional ACA estimation for “filling in the blanks.”

One interesting result with Sue’s paper is that the customized design seems to improve individual hit rates, but makes aggregate share predictions worse. Perhaps the customization procedure focuses the effort on estimating the most desirable levels of each attribute correctly, at the expense of measuring the relative value of the worst levels well. The standard procedure where all attributes are taken in the choice-based tasks more evenly spreads the effort over all levels. Perhaps that is why the customized design excels in predicting respondents’ most preferred options in the holdout tasks, whereas the standard approach works better in predicting choice shares for all alternatives in the holdout sets.

An interesting benefit of customized designs is the ability to measure many levels for an attribute such as brand, while potentially reducing the number-of-attribute-levels bias. Consider a design containing 24 brands, whereas all other attributes have four levels. If the evoked set for brands has four elements, then the design is symmetric for each respondent.

There are cases in which customized designs might be useful or even mandatory, and it is nice to know that they can work reasonably well. If you assemble the collective experimentation of three papers at this conference: Sue York’s (customizing levels), Jon Pinnell’s (customized utility balance) and Joel Huber’s (customized level balance plus partial profiles), you see the

common thread of trying to improve CBC results through adaptive strategies. As more is learned about how to impose individual-level utility constraints (priors) within HB estimation, it may turn out to be possible to achieve more efficient interviews, cleaner individual-level estimates and more accurate share predictions using customized CBC designs.

PRACTICAL ISSUES CONCERNING THE NUMBER-OF-LEVELS EFFECT

Marco Hoogerbrugge
SKIM Analytical

WHAT IS THE “NUMBER-OF-LEVELS EFFECT” IN CONJOINT ANALYSIS?

Let us start with an example. Suppose someone tells you the following: define a conjoint study with two levels for a certain attribute (e.g. top speed 120 km/h and 200 km/h) and measure the importance of the difference between the levels. Then do another study, in which you add a third level *in-between* (160 km/h) and measure the importance of the difference between 120 and 200 km/h again. One would be inclined to believe that this would not make any difference because you still measure the same thing, the difference between 120 and 200. However, people have shown for a long time that it does make a difference: the difference between 120 and 200 actually becomes larger when you add an intermediate level. This is the number-of-levels effect: just because of the fact that more levels have been defined, the importance between two fixed levels becomes larger.

The great contributor to the number-of-levels (NOL) discussion is Professor Dick Wittink. He has been putting a lot of effort into making the “conjoint world” aware of the NOL effect and making us aware of the seriousness of it¹. He has also had a long discussion – still not ended – with Rich Johnson about whether the NOL effect has a *psychological* or an *algorithmic* cause². This discussion has been focused particularly on ACA, Adaptive Conjoint Analysis. Rich Johnson advocates that by showing more variation in the number of levels, respondents will have the impression that this attribute may be more important than they would have thought otherwise. This is the psychological cause. Dick Wittink thinks that much of the NOL effect is related to ACA’s algorithm, because this software has a very special way of designing the trade-offs.

As the impact of the NOL effect is in practice generally not taken very seriously; at the same time most literature on this subject stems from one source. That is the reason why this paper re-examines the NOL effect using completely different approaches than usually are applied.

INTRODUCTION

This paper consists of one large part that is based on three ACA studies and one smaller part based on one CBC, Choice-Based Conjoint, study. Each study described here consists of a mix of European and North American respondents.

The three ACA studies have the same subject in common, and we can more or less compare the various results. Other advantages of these three studies are the large number of respondents and the anticipated high involvement of the respondents.

¹ e.g. Solving the number-of-attribute-levels problem in conjoint analysis, Dick Wittink, Sawtooth Software Conference 1997.

² e.g. A comparison of alternative solutions to the number-of-levels effect, Dick Wittink and P.B. Seetharaman, Sawtooth Software Conference 1999.

The studies deal with blood glucose meters - meters that diabetics use to measure their glucose level so that they can take action in case their level is either too high or too low. For someone who uses such a meter, the thing is obviously of crucial importance.

- The first study is from 1995 and consists of 1000 diabetics.
- The second study is also from 1995 but does not consist of the diabetics themselves but rather of professionals, who recommend meters to diabetics, mostly diabetes nurses, sometimes diabetologists. All together these will be referred to as “nurses”. They were interviewed in order to obtain information about what kind of meter (with what features) they would recommend to diabetics. The sample size was 450 nurses.
- The third study was, again, among diabetics but in 1997, and consisted of 1150 respondents.

The three studies were not completely identical. The 1995 studies among diabetics and nurses had an overlap in attributes of 80% and the levels of the relevant attributes were defined identically. Although the studies among diabetics in 1995 and 1997 had many attributes in common, quite a few levels of these attributes had been dropped in 1997, which makes it more tricky to compare them.

The CBC study is a split-run among 300 specialists for a certain medical disease. The choice tasks were about what kind of drug they would prescribe to a patient that suffers from the disease. We had 5 cells and most cells consist of about 50 respondents, excepting one cell that consists of about 100 respondents.

THE ACA STUDIES: METHODOLOGY

Adaptive Conjoint Analysis has a big advantage that most other conjoint techniques do not have: it starts with a self-explicated part. In this part respondents are asked to rate the importance of the difference between the best and the worst attribute level. Based on this part initial utility values are assigned. Since with this approach only two levels of every attribute are shown, we can easily assume that this measurement must be free of number-of-levels effect. By comparing the final utility values (that are probably affected by the NOL effect) with the initial utility values (NOL effect-free) we have a measure of strength of the NOL effect.

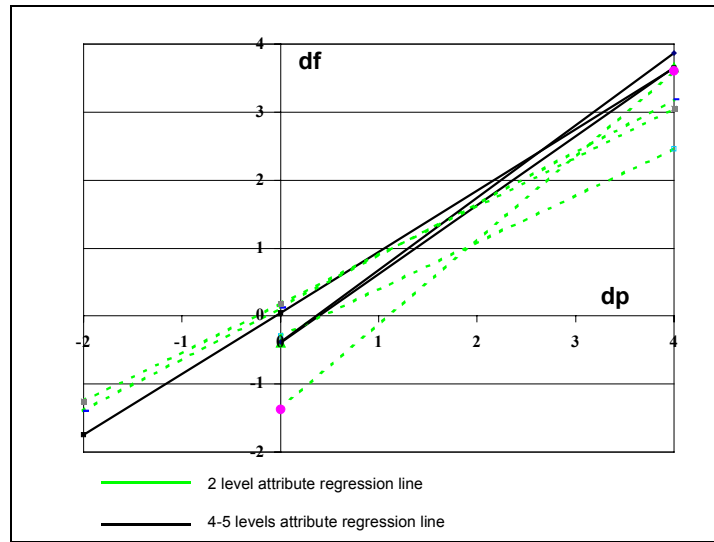
To operationalize this comparison, the difference between the prior and final utility values of two levels has first been calculated. Because a utility value is worth nothing in itself, it is only the difference between two utility values that counts. So for each attribute the difference of utility values of the two levels that were generally acknowledged as the best and the worst level has been taken. By the way, the word “generally” is used here on purpose because there were always occasional individual exceptions. In the end, for each individual respondent (i), for each attribute (a), a difference between prior utility values (dp_{ai}) and a difference between final utility values (df_{ai}) has been calculated.

As the next step a linear regression approach seems quite straightforward: for each attribute (a) the following model should apply:

$$df_{ai} = \alpha_a + dp_{ai} * \beta_a + \epsilon_{ai}$$

So for each attribute there will be estimations for a constant and a slope.

In the first study (1995 with diabetics), typical examples of regression curves are the following:



Note that the large range of dp_{ai} values is 3 or 4, which makes it mostly relevant to analyze the regression curves at the right hand side of the figure. As was to be expected, the 2-level attributes are below average (particularly at the right hand side), while the 4 and 5-level attributes are slightly above average. Many attributes do not have negative dp_{ai} values at all because they were prior-ranked. With other attributes a small minority of respondents have negative values, which is why the regression curves for these attributes have been drawn into the negative quadrant.

For two reasons a straightforward methodology like this does not suffice.

In the first place, for a number of attributes it is possible to have negative values. Assuming that the effects on the negative side of the y-axis and the positive side of the x-axis are the same, the regression curves will automatically be forced through the origin. For most attributes this is not the case, which makes comparison between the attributes impossible. Besides it is useless to have the curves through the origin because it is, among others, the value of the constants α_a that is most interesting. Consequently, in all cases in which dp_{ai} has a negative value, the values are reversed: the dp_{ai} value to positive and the df_{ai} value from negative to positive or vice versa.

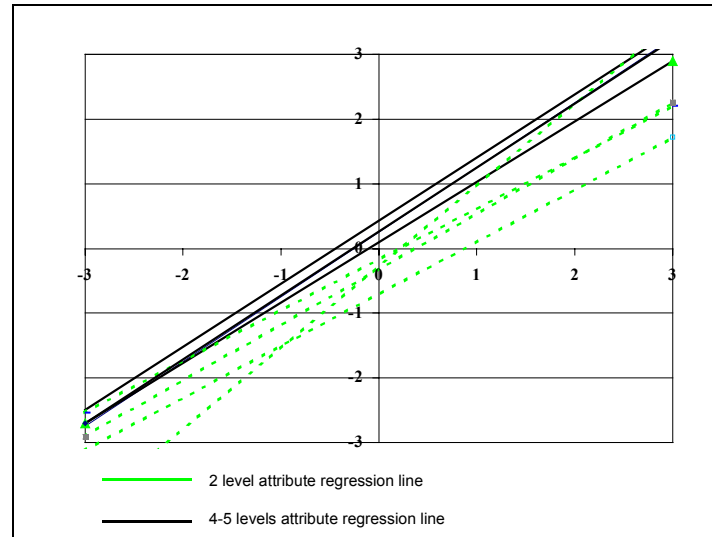
In the second place there is the problem that the dp_{ai} value has different meanings across the respondents. This is most particularly the case for the value 3 (on a scale from 1-4). Some respondents use the value of 3 really as the (upper-)center of the scale, as it was intended, but for other respondents the value of 3 is the bottom since they do not use the ratings of 1 and 2 at all. This latter situation is not as intended but in practice it occurs quite frequently. This problem of different scale usage is of course inherent in a self-explicated rating section. But if the value of the explanatory variable is not univocally interpretable, a regression curve does not make a lot of sense. To overcome this problem, all dp_{ai} values and df_{ai} values per respondent are *centered*. For example, the following conversion takes place:

Original dp_{ai} / df_{ai} value

Converted dp_{ai} / df_{ai} value

Respondent nr	Attribute number									
	1	2	3	4	5	1	2	3	4	5
1	4	2	1	3	2	1.6	-.4	-1.4	.6	-.4
2	3	4	3	4	4	-.6	.4	-.6	.4	.4

After these two refinements the new regression results are much more pronounced:



THE ACA STUDIES: RESULTS IN DETAIL

We can be very short about the *slopes* β_a of the curves. They are on average exactly 1 (which is not by definition the case, as we shall see later) and they are not affected by the number of levels. There are two attributes that have a really different slope, one attribute substantially larger than 1, the other smaller than 1, but both are 2-level attributes.

By the way, for the one attribute with a slope significantly larger than 1 there is a good explanation that has nothing to do with the subject of the paper. The attribute had been a-priori ordered, which has not fully been justified by all respondents - some started with a prior utility value of +1 (which was the lowest value allowed) and ended negatively. While this is one more indication that one has to be conservative with a-priori ordering in ACA, the figure also shows that this “small error” in the design is automatically adjusted during the trade-offs.

More interesting are the constants α_a that differ significantly, depending on the number of levels per attribute. The two-level attributes have an average value of -0.372, the three-level attributes have 0.088, and the four and five-level attributes together have an average constant of 0.262. It is most important to realize that the difference between the two-level and three-level attributes is *greater than the double* of the difference between the three-level and the 4/5-level attributes! We will return to this issue in the following examples as well.

Before switching to other examples, we will first come up with some further practical questions. In the first place, are certain types of attributes more vulnerable to the NOL effect than others? It is clear from the graph that there are big differences within the two-level attributes (the

constants vary from -0.708 to -0.169). Why has a certain attribute such an extremely low value and why are other attributes more modest?

One would have thought that there would be a difference between attributes with nominal or numeric levels. All two-level attributes are nominal in this case (they are all “present/absent” attributes) so that does not tell a lot. Some three or more-level attributes are a mixture of present/absent and then if present they can be defined in a numeric or ordinal way. That is why we have to check two columns together in the following table: we need to crosscheck with the occurrence of an “absent” level and with the nominal or numeric scale of levels.

# of levels	Characteristics of Attribute				
	a-priori ordered	“absent” level	Scale of levels	constant	slope
2	Yes	Yes	NOMINAL	-0.282	1.262
2		Yes	NOMINAL	-0.167	0.791
2	Yes	Yes	NOMINAL	-0.706	0.808
2		Yes	NOMINAL	-0.325	0.862
3	Yes	Yes	ORDINAL	0.074	1.078
3	Yes	Yes	Numeric	-0.118	0.969
3		Yes	Numeric	0.026	1.038
3		Yes	NOMINAL	0.228	0.931
3	yes		Numeric	0.255	1.001
3	yes		Numeric	-0.024	0.991
3			NOMINAL	0.215	0.982
3			Numeric	0.060	0.892
4	yes	Yes	Numeric	0.093	0.933
4	yes		Numeric	0.266	0.990
5			NOMINAL	0.434	0.976

If we concentrate on the 3 and more-level attributes, just by sight it is clear that there is no effect of an “absent” level or an effect of scale type.

There is another effect that does exist though. There is an effect between a-priori ordered attributes and individually ranked attributes. A-priori ordering causes a systematic decline in regression constant, a decline with average 0.125. As we have mentioned above with one particular attribute in mind, this is not really a surprise.

Secondly, we have the comparison between 1995 and 1997. Although the studies were not quite the same, the definition of the one attribute with regression constant of -0.708 did not change. And in 1997 the regression constant appeared to be -0.558. From this we can conclude that there is a *large and systematic* effect that causes this particular attribute to drop down. The only problem left is: what is this systematic effect? This problem has not been resolved yet, and there is no clue from the attribute description anyway.

We can reverse the reasoning here, since the difference between final utility values and prior utility values is, by definition, the effect of the trade-off section. So what we have found here is actually: we do not do the trade-offs in vain! After the self-explicated part of the ACA interview there are a number of factors that change the utility values systematically. “Just one” of these factors is the NOL effect. Another one is the a-priori ordering. And there are other (unidentified) factors at stake that are not merely random.

Another issue is whether the aggregate NOL effect that obviously exists, works out somehow at the individual level. The obvious question now is: are certain (types of) respondents responsible for this number-of-level effect or is it something “general”? A measure of individuals’ NOL-effect sensitivity has been calculated as follows:

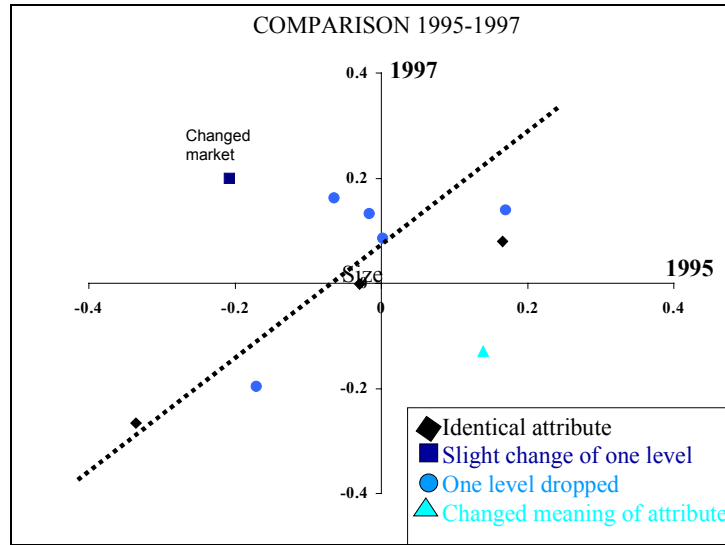
$$\text{eff}_i = \sum_{a(4-5 \text{ levels})} (\text{df}_{ai} - \text{dp}_{ai}) - \sum_{a(2 \text{ levels})} (\text{df}_{ai} - \text{dp}_{ai})$$

If certain respondents are responsible for the number-of-levels effect we would expect the distribution of eff_i to have two modes: one mode at zero for those respondents who do not have a number-of-level effect bias, and another mode at some positive value for those respondents who do. In addition, if certain types of respondents are responsible for this effect (perhaps more knowledgeable respondents have a smaller number-of-levels effect), we would expect the means of eff_i to be different if broken down by certain categorical variables. However, neither of the two appears to be the case: the distribution of eff_i is normal, it has only one mode with a positive value and none of the means shows different values for certain categories. There is absolutely no support for the hypothesis that the number-of-levels effect is respondent-specific.

This result is certainly no proof in the case of Wittink versus Johnson as mentioned in the beginning of the paper. But it might be explained as support to Wittink’s case; namely, if Johnson would be right, one might rather expect some (types of) respondents to be more vulnerable to the psychological cause than others.

Comparison Diabetics Studies 1995 and 1997

Just mentioned was the fact that one attribute in particular showed a similar result across 1995 and 1997. The figure below shows the comparison in time of all attributes to have a total picture, however. The two studies differed in quite a few respects though. First of all, a few attributes had been dropped and a few new attributes had been added (see figure). More important is that a large number of attributes had fewer levels in 1997, because of the NOL effect! We decided in 1997 that in the previous study a variation from 2 to 5 levels per attribute was too big a span, and decided to now keep the limit to 2 to 3 levels per attribute. To allow for a fair comparison (in case an attribute had been decreased from 5 to 3 or from 3 to 2 levels) the figure is corrected for the NOL effect. That is, the average value of the attributes with the same of number of levels has been subtracted, both in 1995 and in 1997.



Another difference is that the application of one attribute changed completely: in 1995 this feature was supposed to give better information to the doctor, while in 1997 the feature was presented to provide better information to the diabetic him- or herself. This is the triangle in the right lower corner. The fact that this attribute is situated there does not mean the feature has become less important because of the changed meaning (there is actually little difference in this respect). It merely means that the difference between final utility value and prior utility value is now negative instead of positive (implying that the prior utility values have increased a lot in 1997).

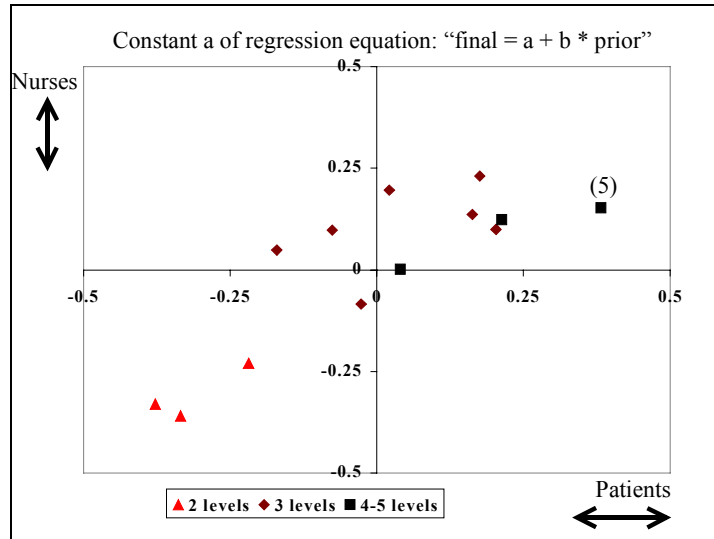
A final difference was with one attribute for which an intermediate level had been quantified differently but the number of levels and the description of the two extreme levels remained the same; this is the attribute situated most upper left. However, more important than the fact that the description has changed marginally is the fact that this feature had been introduced on the market in the period between 1995 and 1997. This latter fact makes it rather tricky to make a statistical comparison between those two years.

My strong suggestion is to ignore these two last attributes. If we do so, there remains a straight linear relationship between 1995 and 1997, thus confirming a stable systematic effect of the trade-offs on the utility values.

Comparison Diabetics with Nurses (1995)

There is another comparison available that works better in terms of attributes and levels used, but differs with respect to the type of people who are interviewed: in 1995 we had also a study among nurses who recommend meters to diabetics. An important difference in attributes, however, is that the attribute with the most extreme negative value had not been included for the nurses.

The results are striking: had we already concluded for the diabetics that the step from 3 to 4 or 5 levels is not as large as the step from 2 to 3 – now with the nurses we see there is no difference at all between 3 or 4-5 levels! The step from 2 to 3 levels is comparable to the diabetics' results, however.



Averaging the outcomes of diabetics and nurses, we have strong evidence that the more levels you have as a minimum across attributes, the less you have to care about a NOL effect.

There is not only a striking similarity but also a striking difference. With the patients the slope of the regression equations was on average 1.0 with the patients in the first study, 0.9 with the patients in the second study, and is on average as little as 0.8 with the nurses in the first study. Although when preparing this paper it has not specifically been the intention to find other values than 1 for the slopes, the possibility that such thing happens is not excluded by this approach.

The implication of a lower slope value than 1 implies that the most important attributes systematically become less important during the trade-offs while the least important attributes grow more important. This could be in line with the generally accepted hypothesis that ACA underestimates the most important attributes (in particular price), relative to CBC. However, as we have seen here, the regression slope is sometimes less than 1, but sometimes equal to 1, so this is not conclusive.

THE CBC STUDY: METHODOLOGY

A split-run CBC was conducted by varying the number of levels per attribute and by varying the number of concepts per choice task. We had:

- 2 concepts + number of levels: 5, 8, 3, 3, 3, 9
- 3 concepts + number of levels: 5, 8, 3, 3, 3, 9
- 4 concepts + number of levels: 5, 8, 3, 3, 3, 9 (double sample size)
- 2 concepts + number of levels: 9, 5, 3, 5, 3, 5
- 3 concepts + number of levels: 9, 5, 3, 5, 3, 5

So, apart from the number of concepts per choice task we had two different designs with regard to the number of levels. In one design we had two attributes with a higher number of levels, two with the same number of levels (both 3) and two attributes with a lower number of levels than in the other design. Note, however, that in all cases the minimum number of levels

across the attributes is 3. Furthermore, we test the difference between 3 and 5 levels once, and we test the difference between 5 and 8 or 9 levels three times.

As we shall see later, it is very difficult to draw conclusions on aggregate CBC results. So in addition two more “sophisticated” approaches will be used. In both approaches *individual* utility values are calculated, either by Hierarchical Bayes (HB-CBC) or by Individual Choice Estimation. The next step is a regression analysis with all these individual values. For each attribute we take the difference in the utility values between the two extreme levels as the variable to be explained and we take the two design factors (number-of-levels and number-of-concepts) as explanatory variables.

However, it is not as straightforward as it seems to undertake such a step. In the regression analysis as proposed, we will have to assume there is only *between-respondents statistical error* and we will have to disregard the *within-respondents statistical error*. This was of course also the case in ACA, but there is good reason to believe that in CBC the latter error that we ignore is more serious. This is because we have less choice tasks than trade-offs in ACA, and choice data instead of rating data. Both HB and ICE have ways to bypass these data scarcity problems in the estimation process: HB by assuming a normal distribution of utility values across respondents, ICE by a fitting the individual values in cross-respondent factors, and also by its allowance to put order constraints on utility values. Both methods make strong statistical assumptions (they *must* make strong assumptions of course because of the scarcity of the data) but those assumptions may not be correct. In other words, statistical error that is reported in the regression results will certainly be underestimated. We run the risk of unjustly concluding a significant effect.

After this warning one other remark about the methodology: the variables to be explained (attribute importances) have been standardized. Otherwise we would draw the conclusion that the importance of *all* attributes increases with the number of concepts that was shown in the choice tasks (where it is actually more concepts result in more data result in better fit result in more extreme utility values).

THE CBC STUDY: RESULTS

The first results are based on the aggregate utility values, by means of subtracting the lowest value per attribute from the highest values. In the table below the values of the high-number-of-level attributes are bold-faced. Numeric attributes with expected monotone decreasing or increasing utility values that in practice appeared to have a completely inconsistent pattern are marked by an asterisk. The attributes are numbered in the same order as they were shown on the screen during the interview.

Attribute number	NUMBER OF LEVELS		MODULE 1			MODULE 2	
	MODULE 1	MODULE 2	2	3	4	2	3
	concepts	concepts	concepts	concepts	concepts	concepts	concepts
1	5	9	2.04	1.54	3.06	2.19	2.96
2	8	5	*	0.85	1.21	*	0.34
3	3	3	0.43	0.11	0.07	0.24	0.65
4	3	5	0.26	0.22	*	0.78	*
5	3	3	2.24	2.52	2.29	1.89	2.31
6	9	5	1.28	*	1.20	0.93	1.15

The conclusions to be drawn from one attribute to another deviate enormously. The number-of-levels effect seems to exist for the second and fourth attribute but only in particular cases (while the asterisks in the other cases do make the result convincing). The NOL effect is inconsistent with the first attribute (non-existent with two concepts, large with three concepts) and finally there is no NOL effect in the last attribute.

Checking for the other variable, the number of concepts per choice task, there are also varying patterns. With the first attribute there is a definite increase in importance, the second attribute increases also, but only convincingly in one case (with 4 concepts) and the four other attributes are apparently not influenced.

An easy explanation for this phenomenon is the following: the more concepts, the less effort respondents undertake to read everything (especially with six attributes), and the more they focus on the attributes that are accidentally on top of the screen.

By the way, taking all combinations together there is one cell of which the results really do not fit with the other cells, namely the cell with module 1 and three concepts. The first attribute has a surprisingly low value, while the last attribute fades away completely. My suggestion is to leave this cell out when drawing conclusions. When one does so, the evidence of the NOL effect becomes very thin: it becomes only based (a little) on the fourth attribute.

As said earlier, drawing conclusions on aggregate utility values is a very rough approach. So here we have the ICE and HB-based regression results. In the table an asterisk is shown if a design effect does not have a significant influence on the importance of the attribute, else a t-value is shown.

	ICE		CBC-HB	
	#levels	#concepts	#levels	#concepts
Attribute 1	*	+5.4	*	+2.6
Attribute 2	-2.7	*	*	*
Attribute 3	+2.1	*	*	*
Attribute 4	*	*	*	*
Attribute 5	*	*	*	-2.6
Attribute 6	*	*	*	*

Combining the two methods and reconsidering the fact that we underestimate statistical error we can conclude that the number-of-levels effect can definitely not be shown in this example with 300 respondents. The number-of-concepts effect seems to be significant in the first attribute though.

CONCLUDING REMARKS

This paper ends with one conclusion (to keep it simple) and in addition a kind of hypothesis for future research. The conclusion is, the NOL effect clearly exists and is large when comparing 2-level attributes with more-level attributes, but the effect is questionable when you only have attributes with at least 3 levels.

A question is left for future research, namely how to explain this result. A start for a hypothesis that we would like to share with you is that the explanation lies in the fact that 2-level attributes are generally absent-present attributes (all attributes in my example studies were) while

more-level attributes are either numeric-defined attributes or discrete-choice attributes. My feeling is that just a yes-no, just a present-absent choice is so different that it distorts the trade-offs. Note on the other hand, however, one may argue that this hypothesis is not particularly confirmed by the results of the more-level attributes in my ACA studies: if one of the levels of these attributes is an “absent” level there is no effect. The hypothesis is then that the trade-off between “absent”, “present at low level” and “present at high level” is fundamentally different from a trade-off between “absent” and “present”. If this hypothesis would be right, there would be – after all – a psychological explanation for the NOL effect, as Rich Johnson has put forward for a long time.

AN EXAMINATION OF THE COMPONENTS OF THE NOL EFFECT IN FULL-PROFILE CONJOINT MODELS

Dick McCullough¹
MACRO Consulting, Inc.

ABSTRACT

The existence of the number of levels effect (NOL) in conjoint models has been widely reported since 1981 (Currim et al.). Currim et al. demonstrated that the effect is, for rank-order data, at least partially mathematical or algorithmic. Green and Srinivasan (1990) have argued that another source of this bias may be behavioral. Although NOL can significantly distort study findings, no method for eliminating NOL, other than holding the number of attribute levels constant, has been discovered.

In this paper, we confirm the existence of both algorithmic and psychological components of NOL for full-profile metric conjoint, examine the time decay of the psychological component and further develop a solution originally proposed in McCullough (1999) to completely eliminate NOL effects from full-profile trade-off models.

INTRODUCTION

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

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

The NOL effect behaves somewhat differently for rank-order data and metric data. No NOL effect has so far been detected by simply removing levels from metric data in Monte Carlo simulations. However, there appears to be some question of whether or not there can be an algorithmic component of NOL for metric data derived from human responses, if the assumptions of normal, independent error terms are not met.

¹ The author wishes to thank Rich Johnson and Dick Wittink for their invaluable assistance with the design of this study as well as Jayme Plunkett and Jamin Brazil for their help in collecting and analyzing the data.

On the other hand, for rank-order data, it has been widely reported since Currim et al. that an NOL effect can be detected that is at least partially algorithmic by removing levels. Thus, in this strict sense of algorithmic component, the NOL effect from rank-order data may have both an algorithmic and psychological component but the NOL effect from metric data may have only a psychological component. The question is still open as to whether or not an algorithmic component for metric data exists when the data are derived from human responses.

It is generally agreed that the NOL effect is a serious problem that can and often does significantly distort attribute relative importance scores, utility estimates and market simulation results. And largely due to the fact that the only known method for removing this effect has been to hold the number of levels constant across attributes, it has often been ignored in commercial studies. McCullough (1999) suggested an approach that may eventually prove practical in eliminating NOL effects in full-profile conjoint studies. This paper further develops the concepts originally proposed there.

METHODOLOGICAL OBJECTIVES:

The objectives of this paper are:

- For full-profile metric conjoint, confirm (or deny) the existence of and estimate the separate magnitudes of the algorithmic and psychological components of NOL
- Confirm (or deny) the existence of and estimate the magnitude of the order effect potentially present in the two-stage conjoint approach (see McCullough (1999))
- Measure the effect of time on the psychological component of NOL
- Quantify the learning effect of exposure to level specifications prior to conjoint exercise
- Suggest a potential solution to eliminate NOL
- Validate the key assumption of that solution, i.e., that the psychological component diminishes rapidly over time when viewed in conjunction with an order effect

RESEARCH OBJECTIVE:

The objective of the study is:

- Identify key drivers in Web survey banner ad solicitations

STUDY DESIGN:

Overall, a multi-cell study design has been constructed to isolate the effects of several potential biases to trade-off models, using a web survey for data collection. The potential biases addressed by this study are:

- Algorithmic component of NOL
- Psychological component of NOL

- A time-lagged effect of the psychological component: exposure during a conjoint exercise to an attribute with a large number of levels may have a lingering psychological effect on subsequent conjoint exercises that contain that attribute
- An order effect: in a two-stage conjoint study, that is, a study with two separate conjoint exercises, the existence of one exercise prior to the second may create an order bias
- A learning effect: exposing respondents to attribute levels prior to a conjoint exercise may create a bias which is referred to here as a learning effect

To be able to analytically isolate and measure the magnitude of each of the above effects the study was split into four cells. The survey outline for each cell is as follows:

Cell1 = DQ || F, 2, demo's
 Cell2 = DQ || 2, F, demo's
 Cell3 = DQ, 2&F mixed, demo's
 Cell4 = DQ || F, demo's, 2

Where:

- DQ denotes direct questioning to identify exterior levels of incentive attribute,
- || denotes a two-day pause between stages (the assumption being that learning effect can be eliminated by delaying subsequent stages),
- 2 denotes 2-levels, that is, exterior levels trade-off, a trade-off exercise containing only the two exterior levels of each attribute,
- and F denotes full-levels trade-off, that is, a trade-off exercise containing all levels of all attributes.

The data collection protocol was as follows:

- Email invitation to split-cell web survey:
 Potential respondents were invited to participate in the survey via email.
- nth respondent receives survey to cell $n \bmod 4$:
 Every respondent that came to the survey website was routed through a counter which assigned respondents to each of the four cells in rotating order.
- Sample frame generated from email panel:
 Sample frame was purchased from an email list supplier. Opt-in names only were purchased. Opt-in lists are comprised of people who have previously agreed to allow themselves to be contacted for surveys. A small portion of the sample was obtained from the Bauer Nike Hockey website where visitors to that site were invited, via a banner ad, to participate in the survey.
- Metric full-profile conjoint study conducted via web survey:
 The trade-off exercises were designed using Sawtooth Software's CVA program. Conjoint measurement was pairwise ratings on a 9 point scale. There were 20 paired ratings in the full-levels trade-off exercises (D efficiency = 95%) and 4 paired ratings in

the two-levels trade-off exercises (D efficiency = 100%). Individual level utilities were estimated for both full and exterior-only exercises using Sawtooth Software, Inc.'s CVA software.

- Sample size:

Approximately 8,500 potential respondents were invited to participate. Roughly 70% of those completed the direct questioning segment of the survey and roughly 40% of those returned and completed the second segment of the survey. In cell 3, there was no time delay between the direct questioning segment and the remainder of the survey. Consequently, initial sample size for cell 3 was 1,474. Sample sizes in all cells were reduced to those whose direct questioning exterior levels matched perfectly their derived exterior levels (see Analysis section below). Roughly 22% of the completed surveys (174 per cell, on average) in all cells had matching direct questioning exterior levels and derived exterior levels. Additionally, samples were rescreened to eliminate only those respondents whose derived utility weights for their claimed exterior levels were statistically significantly different from their derived exterior levels. For these statistically matched samples, average sample size was approximately 600, or 85% of initial completes. Both data sets are discussed in the Analysis section below.

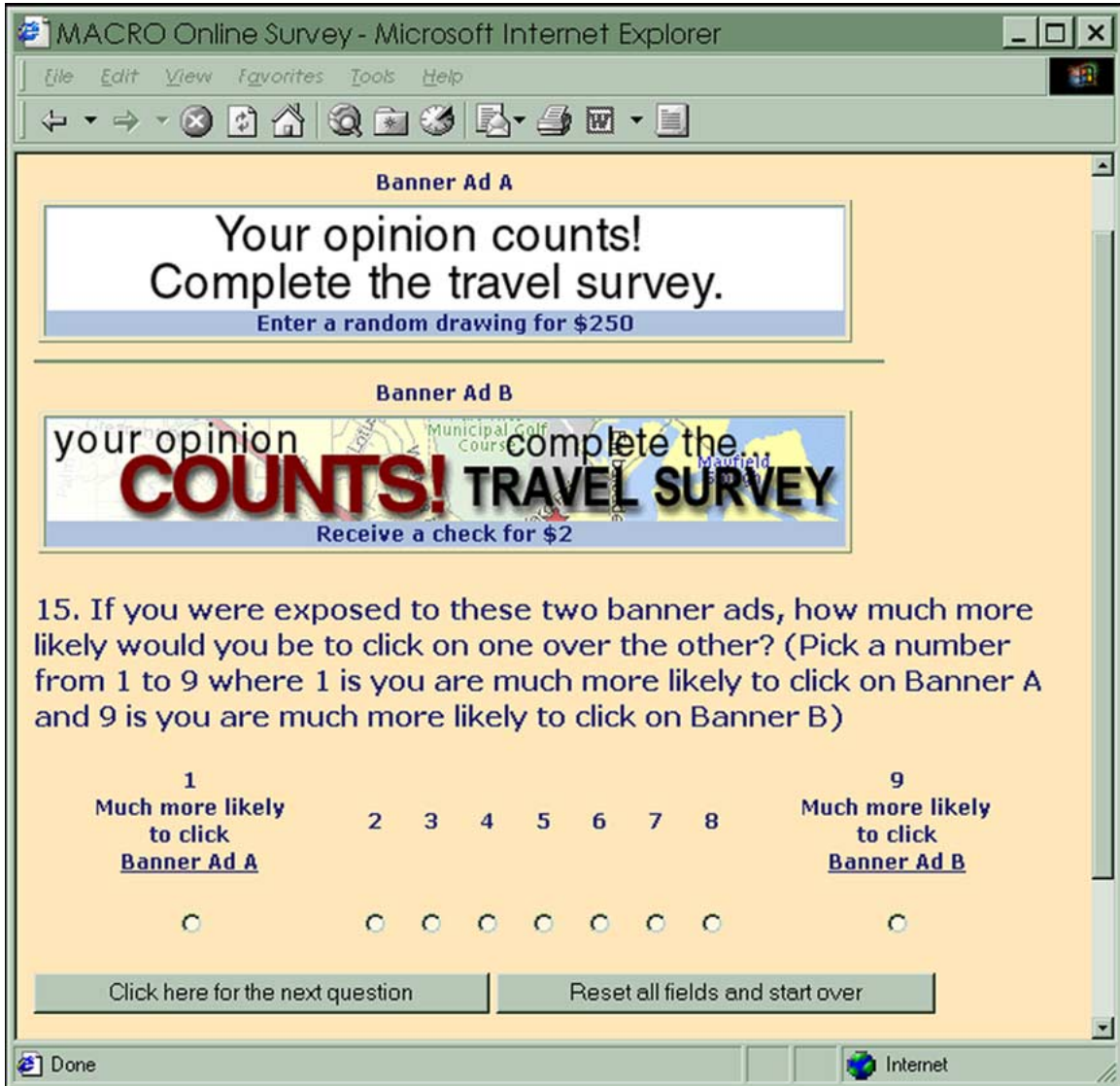
Data collection took place November 29 through December 18, 1999.

Trade-off attributes used in all eight conjoint exercises and the levels used in the four full-levels conjoint exercises were:

- Text only vs. graphics and text (2 levels)
- Animation vs. static (2 levels)
- Incentive (9 levels):
 - Random drawing for \$250 cash
 - Random drawing for \$2,500 cash
 - Random drawing for an Italian leather briefcase
 - Random drawing for a week in Hawaii for two
 - Each respondent receives a check for \$2
 - Each respondent receives a check for \$25
 - Each respondent receives a letter opener
 - Each respondent receives a Swiss Army knife
 - No incentive

All respondents were directed to a URL which contained all or part of the web survey for the cell they were assigned to. An example of the screen shown for a typical pairwise ratings question from one of the conjoint exercises can be found in Figure 1.

Figure 1.



Note: 36 versions of the two-levels design were needed (nine levels taken two at a time). Each respondent saw the version which contained his/her exterior levels for the incentive attribute.

Let $Cell1(2)$ = relative importances from 2-levels conjoint in cell1 and $Cell1(F)$ = relative importances from full-levels conjoint in cell1, similarly for all other cells.

And let:

A = Algorithmic component,

P_i = Psychological component at time i

O_{ij} = Order effect when conjoint exercise i precedes conjoint exercise j

L = Learning effect (due to exposure to levels during direct questioning)

Note that there are three P_i : P_0 , P_1 and P_2 . P_0 is the psychological component of NOL during a trade-off exercise that contains unequal numbers of levels across attributes. P_1 is the psychological component of NOL immediately after a trade-off exercise that contains unequal numbers of levels across attributes. P_2 is the psychological component of NOL a brief time after a trade-off exercise that contains unequal numbers of levels across attributes. Thus, in cell 1, where the full-levels trade-off is followed immediately by the two-levels trade-off, P_1 is the form of psychological component that would be affecting the two-levels trade-off. In cell 4, where the full-levels trade-off is followed by a demographics battery and then the two-levels trade-off, P_2 is the form of psychological component that would be affecting the two-levels trade-off.

Also, there are two forms of O_{ij} potentially at work: O_{F2} and O_{2F} . O_{F2} is the order effect when full-levels precedes two-levels. O_{2F} is the order effect when two-levels precedes full-levels.

Each of these eight different trade-off exercises (with the exceptions of Cell1(F) and Cell4(F)) will have a different combination of these various sources of bias operating. Table 1 below summarizes the sources of bias operating on each of the different trade-off exercises.

Table 1.

<u>Cell</u>	<u>Bias Sources</u>
Cell1(F)	A and P_0
Cell1(2)	O_{F2} and P_1
Cell2(F)	A, P_0 and O_{2F}
Cell2(2)	nothing
Cell3(F)	L, A and P_0
Cell3(2)	L and P_0
Cell4(F)	A and P_0 (same as Cell1(F))
Cell4(2)	O_{F2} and P_2 (similar to Cell1(2))

Jayme Plunkett and Joel Huber have both verbally expressed the opinion that the psychological effect may be short term². So short term that it may not appear or at least not appear fully when full-levels precedes two-levels in a two-stage design. If so, and if order effect is negligible, then it seems a good solution to NOL would be to do full-levels, calculate utils on the fly, insert derived exterior levels into a 2-levels conjoint (all with same respondent and within the same interview) and omit direct questioning altogether. This would avoid the discarded sample problem discussed in McCullough (1999). In cells 1 and 4, varying amounts of demo's (from none to some) have been inserted between full-levels and 2-levels to measure the effect of time on the psychological effect, if it is indeed short-term.

If $Cell1(2) = Cell2(2)$, then $P_1 + O_{F2} = 0$, the psychological/order component is very short-term and the on-the-fly solution should be viable. Note that this argument assumes that P_1 and O_{F2} have the same sign.

If $Cell1(2) \neq Cell2(2)$ but $Cell4(2) = Cell2(2)$, then $P_2 + O_{F2} = 0$, the psychological/order component is short-term, but not very short-term, and the on-the-fly solution should be viable, if the second trade-off exercise is delayed for a short amount of time by the insertion of other

² Verbal discussions were held at the 1999 Sawtooth Software Conference.

survey questions, such as a demographics battery. Again note that this argument assumes that P_2 and O_{F2} have the same sign.

The above design and analysis should allow us to:

- Confirm (or deny) the viability of the on-the-fly solution
- Isolate and estimate magnitudes for A, P_0 , O_{2F} and L
- Measure the time decay of P and O in combination, that is, measure $P_1 + O_{F2}$ and $P_2 + O_{F2}$

ANALYSIS

In each cell, respondents were asked directly what were their most and least preferred incentive levels. These claimed exterior levels were used in the two-levels trade-off exercise. Both of the other attributes had only two levels so no direct questioning was required to identify their exterior levels. Respondents whose claimed exterior levels, based on direct questioning were different from their derived exterior levels, based on the full-levels trade-off exercise, were excluded from this analysis. Recall from above that “different” is defined two ways: not perfectly, i.e., numerically exactly, matched and statistically significantly different.

Table 2 shows the frequency and incidence breakdowns by cell, for both perfectly matched and statistically matched samples.

Table 2.

<u>Cell</u>	<u>Invitations</u>	<u>1st part completes</u>	<u>2nd part completes</u>	<u>Perfectly Matched</u>	<u>Statistically Matched</u>
1	2,000	1,386/69%	536/39%	138/26%	447/83%
2	2,000	1,359/68%	544/40%	127/23%	462/85%
3	2,500	1,474/60%	1,474/100%	286/19%	1,224/83%
4	2,000	1,374/69%	546/40%	144/26%	476/87%

Attribute relative importance scores were calculated for each attribute within each trade-off exercise for each individual by taking the absolute difference between the attribute level with the highest utility weight and the attribute level with the lowest utility weight (for the same attribute), summing the absolute differences across all attributes in the trade-off exercise, dividing that sum into each absolute difference and multiplying by 100. These individual attribute relative importance scores were then averaged across all respondents.

To measure the magnitude of various sources of bias, mean attribute relative importance scores for the incentive attribute are differenced. Since the other two attributes have only two levels, any NOL-related effect will be reflected entirely in the attribute relative importance scores for the incentive attribute.

Using this difference, the various bias sources can be estimated. For example, the magnitude of the algorithmic component of NOL, i.e., A, is defined as the attribute relative importance score for the incentive attribute in Cell3(F) minus the attribute relative importance score for the incentive attribute in Cell3(2), since Cell3(F) is affected by L, A and P_0 and Cell3(2) is affected by L and P_0 (see Table 1). In Table 3 below, several bias sources are defined in terms of the cells of this study.

Table 3.

<u>Source</u>	<u>Definition</u>
A	Cell3(F) – Cell3(2)
P ₀	(Cell1 (F) – Cell2(2)) – (Cell3(F) - Cell3(2))
O _{2F}	Cell2(F) – Cell1 (F)
L	Cell3(F) – Cell1 (F)
P ₁ + O _{F2}	Cell1 (2) – Cell2(2)
P ₂ + O _{F2}	Cell4(2) – Cell2(2)
● ³	Cell1 (F) – Cell4(F)

Statistical significance of the differences in two sets of attribute relative importance scores has been tested using both anova and t-tests.

RESULTS

Table 4a lists the attribute relative importance scores for all attributes for the perfectly matched samples.

Table 4a: Perfectly Matched Samples.

	Cell1(n=138)		Cell2(n=127)		Cell3(n=286)		Cell4(n=144)	
	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>
Text	7.17%	2.09%	6.12%	3.68%	6.06%	5.68%	6.59%	3.50%
Animation	8.47%	1.96%	5.19%	2.42%	6.64%	6.04%	7.22%	1.77%
Incentive	84.36%	95.94%	88.69%	93.90%	87.30%	88.28%	86.19%	94.73%

Table 4b lists the attribute relative importance scores for all attributes for the statistically matched samples.

Table 4b: Statistically Matched Samples.

	Cell1(n=447)		Cell2(n=462)		Cell3(n=1,224)		Cell4(n=476)	
	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>	<i>Full</i>	<i>Exterior</i>
Text	8.95%	6.81%	6.75%	8.80%	11.00%	14.29%	8.82%	6.95%
Animation	9.06%	5.05%	7.21%	8.30%	7.82%	12.04%	8.46%	4.94%
Incentive	81.99%	88.14%	86.04%	82.89%	81.15%	73.67%	82.73%	88.10%

Based on these data, the following calculations of differences in incentive attribute relative importances were made:

³ ● is listed as a bias source for methodological confirmation purposes only. It is known that Cell1(F) and Cell4(F) have exactly the same biases operating on them regardless of what those biases are, since these two cells have been implemented exactly the same way. Therefore there should be no statistically significant difference in their attribute relative importance scores.

Table 5a: Perfectly Matched Samples.

<u>Source</u>	<u>Definition</u>	<u>Difference</u>
A	Cell3(F) – Cell3(2) =	-.98 p.pts.
P ₀	(Cell1 (F) – Cell2(2)) – (Cell3(F) - Cell3(2)) -9.54p.pts.- (-.98) p.pts.=	-8.56 p.pts. ⁴
O _{2F}	Cell2(F) – Cell1 (F) =	4.33 p.pts. ⁴
L	Cell3(F) – Cell1 (F) =	2.94 p.pts. ⁴
P ₁ + O _{F2}	Cell1 (2) – Cell2(2) =	2.04 p.pts.
P ₂ + O _{F2}	Cell4(2) – Cell2(2) =	.83 p.pts.
●	Cell1 (F) – Cell4(F) =	-1.83 p.pts.

Table 5b: Statistically Matched Samples.

<u>Source</u>	<u>Definition</u>	<u>Difference</u>
A	Cell3(F) – Cell3(2) =	7.48 p.pts. ⁴
P ₀	(Cell1 (F) – Cell2(2)) – (Cell3(F) - Cell3(2)) -0.9 p.pts.- 7.48 p.pts.=	-8.38 p.pts. ⁴
O _{2F}	Cell2(F) – Cell1 (F) =	4.05 p.pts. ⁴
L	Cell3(F) – Cell1 (F) =	-.84 p.pts.
P ₁ + O _{F2}	Cell1 (2) – Cell2(2) =	5.25 p.pts. ⁴
P ₂ + O _{F2}	Cell4(2) – Cell2(2) =	5.21 p.pts. ⁴
●	Cell1 (F) – Cell4(F) =	0.74 p.pts.

ANALYSIS OF PERFECTLY MATCHED DATA

These data in table 5a show that, among this sample, there is a statistically significant psychological component of NOL but not an algorithmic component (at least not one detectable with these sample sizes). This psychological component is the largest of all the biases measured. Further, these data demonstrate that there is an order effect in the two-stage methodology that also significantly biases the attribute relative importance estimates. Also, there is a learning effect due to direct questioning that significantly biases the attribute relative importance estimates.

Intriguingly, these data also show that the combination of a time-delayed psychological effect and an order effect is negligible. This finding superficially suggests that a practical solution to eliminate the NOL effect is to do full-levels trade-off, calculate utils on the fly, insert derived exterior levels into a 2-levels trade-off (all with same respondent and within the same interview) and omit direct questioning altogether. Recall that this conclusion would require the assumption that P₁, P₂ and O_{F2} have the same sign. Given that fact that these data consistently and strongly suggest that the psychological component is negative and the order effect is positive, the validity of the two-stage approach as currently formulated must be questioned. That is, in this case, P₁ + O_{F2} = P₂ + O_{F2} = 0 but this finding does not generalize.

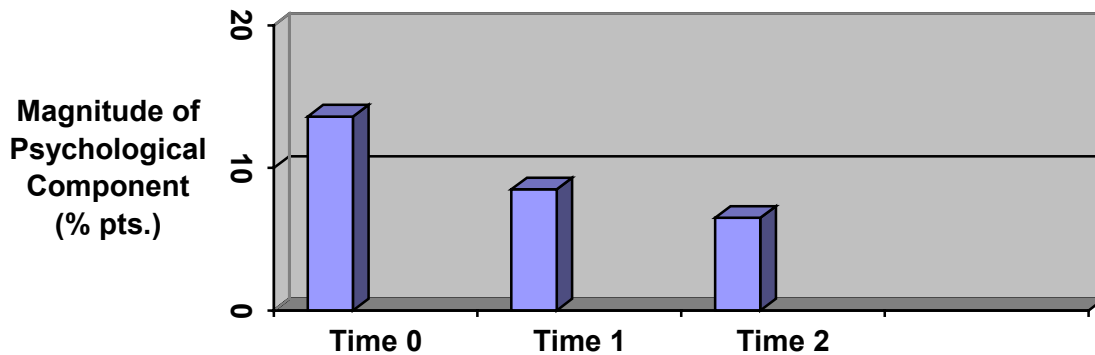
The fact that the attribute relative importances from Cell1(F) and Cell4(F) are statistically equal adds some face validity to the data collection process.

The data show that P₁ + O_{F2} is statistically equal to zero. P₀ is known to be large and negative. If O_{F2} is roughly the magnitude of O_{2F} and P₁ is negative, then the magnitude of P₁ is

⁴ Statistically significant at 95% confidence level.

roughly 8.5 percentage points. Similarly, P_2 would be roughly 6.5 percentage points. If true, this would allow us to chart the time decay of the psychological component. These data suggest such a chart might look like the one in Figure 2.

Figure 2.
Possible Time Decay of Psychological Component



As noted above, a careful review of Table 5a will show the surprising result that the psychological component is negative. That is, the large number of attribute levels in the incentive attribute cause incentive attribute relative importance to diminish, rather than increase. This result is consistent across all three cells which contain the psychological component P_0 in the full-levels exercise and not the two-levels, i.e., cells 1, 2 and 4. Recall that $P_1 + O_{F2}$ and $P_2 + O_{F2}$ were not statistically different from zero.

There are two possible explanations for this phenomenon. One possibility is that, in the full-levels exercise, respondents are exposed to the levels from both two-level attributes, graphics and animation, roughly four and a half times more often than they are to the one nine-level attribute, incentive. This exposure may sensitize them to the two-level attributes, resulting in greater importance for the two-level attributes in the full-levels exercise and lesser importance for the incentive (nine-level) attribute. Conversely, in the two-level exercise, where attribute levels are all exposed equally, the two-level attributes would have less importance than in the full-levels exercise and the incentive attribute would have more.

Another possible explanation is that respondents, when faced in the two-levels exercise, with banner ads that have either their most preferred incentive or their least preferred incentive, give polarized responses. That is, respondents may have tended to give more 1 or 9 ratings because the incentive attribute either had a level respondents liked strongly or disliked strongly. This is all the more likely given the fact that the incentive attribute is overwhelmingly most important of all three attributes tested. This behavior may be an example of the utility balance issue discussed in Wittink et al. (1992a), Wittink et al. (1992b) and again in Wittink et al. (1997). That is, the incentive attribute may have greater attribute relative importance in the two-levels case because the utility imbalance is extreme in the two-levels trade-off. Wittink has demonstrated that utility imbalance will increase the magnitude of the NOL effect.

It is also possible that the sign of the psychological component may be a function of the data collection method. Perhaps, for example, monadic ratings inspire a different psychological component in respondents than pairwise ratings.

ANALYSIS OF STATISTICALLY MATCHED DATA

The statistically matched data offer somewhat different results from the perfectly matched data. Similar to the perfectly matched data, the statistically matched data show a statistically significant and negative psychological component and a statistically significant order effect (O_{2F}). However, the statistically matched data also show a statistically significant algorithmic component, statistically significant $P_1 + O_{F2}$ and $P_2 + O_{F2}$ and a statistically insignificant learning effect.

The fact that $P_1 + O_{F2}$ is statistically equal to $P_2 + O_{F2}$ implies that $P_1 = P_2$. And if we assume that O_{F2} is statistically equal to O_{2F} , we can conclude that $P_1 = P_2 = 0$. Thus, for the perfectly matched sample, the time decay of the psychological component of NOL appears to be slow while for the statistically matched sample, it appears to be quite rapid. This appears consistent with the fact that for the perfectly matched sample, there was a significant learning effect but for the statistically matched sample, there was not.

The fact that the attribute relative importances from Cell1(F) and Cell4(F) are statistically equal again adds some face validity to the data collection process.

DISCUSSION

In summary, we have two different data sets that yield somewhat different results. However, regardless of the way the data are analyzed, it must be concluded that there is a sizable psychological component that, surprisingly, can be negative and that there is a significant order effect (O_{2F}).

These data also suggest that a two-stage conjoint design where respondents do full-levels trade-off, utilities are calculated on the fly, derived exterior levels are inserted into a 2-levels conjoint (all with same respondent and within the same interview) may be seriously flawed as a method of eliminating the NOL effect, due to the existence of a significant order effect (assuming $O_{2F} = O_{F2}$), unless some method of handling the order effect can be developed.

A final comment on matching respondents' claimed exterior levels (via direct questioning) to their derived levels. Across all cells, 22% of respondents had perfect matching. That is, 22% of respondents had claimed exterior levels that matched perfectly with the utility levels derived from the full-levels conjoint exercise. It appears possible that these respondents may be capable of more metric quality responses than those respondents who did not exactly match their claimed exterior levels and their derived exterior levels. If so, then the algorithmic component of the NOL effect measured with the perfectly matched data could be minimized by metric quality responses (refer to Steenkamp and Wittink). It may be the case that with a sample of less metric quality responses, such as the statistically matched samples, more statistically significant results would be obtained, i.e., an algorithmic component might be shown to exist, both because a larger NOL effect would exist and also due to the larger sample sizes. That is exactly what has occurred. Only the learning effect has diminished with the larger sample size.

Clearly, the negative psychological component is a surprising result. The fact that this result is consistently and clearly reflected in the data makes it hard to ignore. There are other results that are also puzzling:

- Why does an algorithmic component appear with the statistically matched data but not with the perfectly matched data?
- Why does the learning effect appear with the perfectly matched data but not with the statistically matched data?
- Why do $P_1 + O_{F2}$ and $P_2 + O_{F2}$ appear with the statistically matched data but not with the perfectly matched data?

One possible answer for the lack of algorithmic component among the perfectly matched sample may be that, for metric quality responses, the regression model error terms may not violate the assumptions of normality and independence. Conversely, it may be the case that the statistically matched sample generated non-metric quality responses and violated the error term assumptions.

The existence of a learning effect among the perfectly matched sample may again be influenced by non-metric quality responses. Would respondents capable of metric quality responses have greater recall of the direct questioning portion of the survey during the conjoint exercises and, therefore, be more influenced?

Interestingly, and perhaps related to the difference in learning effect results, perfectly matched samples appear to demonstrate a slower time decay of the psychological component than the statistically matched sample. Do metric quality respondents “remember” better? That is, does the psychological component of NOL decay more slowly for metric quality respondents than for non-metric quality respondents? Is the perfectly matched sample a sample of “super” respondents?

One possible explanation for the discrepancies between the results of the perfectly matched samples and the statistically matched samples is that the perfectly matched samples have been screened to retain only those respondents who are unusually “smart”. They provide metric quality responses (and normal error terms), they remember the direct questioning experience and they retain the psychological influence longer (perhaps again because of better recall).

However, there are several other potential factors that may affect these results, as well:

- The incentive attribute was nominal, not ordinal or metric.
- Data collection was paired comparison rather than monadic ratings.
- The learning effect associated with the direct questioning stage of the survey may alter responses in some unanticipated way.
- The number of levels is radically different across attributes (two versus nine).
- The relative importance of the incentive attribute is overwhelmingly dominant.
- Fatigue: the full-levels exercise involved 20 cards while the two-levels exercise involved just four.

One argument against the appropriateness of using the statistically matched samples in analysis is that the statistically matched samples would have more noise, more error, making it more difficult to obtain statistically significant differences. But the statistically matched samples in this study found more statistically significant differences than the perfectly matched samples. Thus, this argument does not seem to have merit.

If metric quality responses are playing a role in these data, then it would appear that the statistically matched data sets would be more appropriate for analysis. If the learning effect associated with the direct questioning stage of the survey was also involved, then again it would appear that the statistically matched data sets would be more appropriate for analysis, since the statistically matched samples were not affected by a learning effect. The other factors: nominal attribute, paired ratings, number of levels disparity, relative importance disparity and fatigue, would apply equally to both the perfectly matched samples and the statistically matched samples.

Thus, it would appear that the statistically matched samples would be more appropriate for analysis and the conclusions derived from those data should be given greater weight than the conclusions derived from the perfectly matched data.

Based on these findings in combination with earlier studies, a clearer perspective on the NOL effect is beginning to emerge. The following hypothesis is consistent with existing literature:

There are two sources for the NOL effect: a psychological component due to disproportionate exposure to selected levels and an algorithmic component due to non-metric quality responses making the data act similar to rank order data. The psychological component is, at least sometimes, negative. In general, the algorithmic component is bigger than the psychological. In the study cited in this paper, the large number of levels of the most important attribute may have exaggerated the magnitude of the psychological component. In all other studies reported in the literature, the total NOL effect has been reported as positive. This result could be explained by an algorithmic component which is generally larger than the psychological component under more typical circumstances. None of these earlier studies had separated out the algorithmic component from the psychological component. Thus, a negative psychological component would have simply made the total observed NOL effect smaller but it would have remained positive.

If the above hypothesis is true, then future studies should focus on how to remove each of the sources of the NOL effect separately. Perhaps the algorithmic component may be eliminated or minimized by asking questions in such a way that respondents are able to give metric responses. To combat the psychological component, perhaps there can be developed experimental design strategies that constrain each level of each attribute to be shown an equal number of times, without sacrificing the model's ability to estimate unbiased parameters. Another avenue for investigation is utility balance. As has been discussed earlier, Wittink has shown that by balancing total utility in pairwise ratings the NOL effect is diminished. Does utility balance affect the algorithmic component, the psychological component or both? The implication of a better understanding of the sources of the NOL effect is that we have new areas to examine for potential solutions.

SUMMARY

Both algorithmic and psychological components of NOL were confirmed to exist and quantified. The psychological component was shown to be negative, at least in this case. The psychological component also appeared to decay rapidly over time, for the more general statistically matched samples, assuming the two order effects, O_{2F} and O_{F2} , to be equal in magnitude.

A solution to the NOL effect continues to be an elusive target. While the two-stage approach remains potentially useful, it cannot yet be viewed conclusively as a valid method for eliminating NOL. It appears that there is an order effect inherent in the two-stage approach that must be accounted for. However, given the lack of learning effect demonstrated here (for the statistically matched samples), the solution proposed in McCullough (1999) may be viable if: 1) respondents are screened to have statistically matched claimed and derived exterior levels rather than perfectly matched claimed and derived exterior levels and 2) the order of the full-levels trade-off and the two-levels trade-off is rotated to minimize the order effect. The amount of lost sample not only diminishes dramatically with the alternative screening method but the sample derived may be more representative of the target population. This revised approach would not suffer from a learning effect or a time-lagged psychological component of NOL and order effect would be minimized. Further work needs to be done to verify that the time-lagged psychological component of NOL is zero, that is, confirm the assumption $O_{2F} = O_{F2}$, and understand the magnitude of the resulting order effect when the two trade-offs are rotated.

Additional work needs to be done to understand how different types of respondents may have different NOL effects, depending on the quality of their responses and, perhaps, even on their memory capacities or other mental attributes.

REFERENCES

- Currim, I.S., C.B. Weinberg, D.R. Wittink (1981), "The Design of Subscription Programs for a Performing Arts Series," *Journal of Consumer Research*, 8 (June), 67-75.
- Green, P.E., and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54 (October), 3-19.
- McCullough, Dick (1999), "The Number of Levels Effect: A Proposed Solution," *1999 Sawtooth Software Conference Proceedings*, 109-116.
- Steenkamp, J.E.M., and D.R. Wittink (1994), "The Metric Quality of Full-Profile Judgments and the Number-of-Levels Effect in Conjoint Analysis," *International Journal of Research in Marketing*, Vol. 11, Num. 3 (June), 275-286.
- Schiffstein, Hendrik N.J., Peeter W.J. Verlegh, and Dick R. Wittink (1999), "Range and Number-of-Levels Effects in Derived and Stated Attribute Importances," an unpublished working paper, Yale School of Management.
- Wittink, D. R., (1990), "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions," *1990 Advanced Research Techniques Forum Proceedings*, American Marketing Association.
- Wittink, D. R., J. C. Huber, J. A. Fiedler, and R. L. Miller (1992a), "The Magnitude of and an Explanation for the Number of Levels Effect in Conjoint Analysis," working paper, Cornell University (December).
- Wittink, D. R., J. C. Huber, P. Zandan, and R. M. Johnson (1992b), "The Number of Levels Effect in Conjoint: Where Does It Come From and Can It Be Eliminated?," *1992 Sawtooth Software Conference Proceedings*, 355-364.
- Wittink, D.R., L. Krishnamurthi, and D.J. Reibstein (1989), "The Effects of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters*, 1, 113-23.
- Wittink, D. R., William G. McLaughlan, and P. B. Seetharaman, "Solving The Number-of-Attribute-Levels Problem In Conjoint Analysis", *1997 Sawtooth Software Conference Proceedings*, 227-240.

CREATING TEST DATA TO OBJECTIVELY ASSESS CONJOINT AND CHOICE ALGORITHMS

Roy Poynter

Millward Brown IntelliQuest

When comparing the different merits of alternative Conjoint and Choice model approaches there is a great risk of confusing two quite separate issues. The first issue is the quality of the choice algorithm's ability to interpret valid responses. The second issue is the nature of the responses generated (ie are the respondents able to give us valid answers).

The main topic of this paper is to propose a method for investigating the first of these topics in isolation from the second. However, the paper commences by discussing the other elements involved in the responses generated.

PRODUCTS AND PEOPLE DIFFER

There is a danger that in trying to research the world of Choice modelling there will be a tendency to look for a single best technique. This tendency can appear in different guises. For example an author might try to generalise from a specific survey, perhaps asserting that this test shows that this affect occurs with this technique (or even more worrying that this test shows that this affect does not occur with this technique). Alternatively an author might try to validate or extend an earlier piece of research using a potentially different context.

An example of this difficulty is illustrated by Vriens, Oppewal, and Wedel in their 1998 paper "Ratings-based versus choice-based latent class conjoint models". The paper compares their research with a project published by Moore, Gray-Lee, and Louviere. The authors highlight just one difference between the projects. Namely in the Moore et al paper the order of the ratings and choice tasks are not controlled for (in Vriens et al this is controlled for). The discussion in the paper pays no cognisance of the subject of the research or of the respondents. In Vriens et al the topic is Drip coffee makers (a consumer durable), apparently conducted in the Netherlands. In Moore et al the product was toothpaste (an FMCG product), in a different continent.

Whilst it is possible that the differences in the two papers relate to the task order, it is also possible they relate to the different consumers or the different product types.

Different consumers think about products in different ways, this difference is the basis of many market segmentations. Different types of products are thought about in different ways. For example, most consumers would consider their next PC purchase in much more detail than whether to buy a glass or carton of milk. Some purchases are very conjoint friendly (such as the configuration of PCs), other markets are not (such as why people prefer Coke or Pepsi).

When we evaluate different techniques and different effects we need to look at the people and markets as well.

SOME QUESTIONS ARE HARDER TO ANSWER

For a Choice technique to be successful it must be possible for respondents to answer the questions. There are several reasons why people may not be able to provide usable answers:

- The researcher may not ask the questions clearly enough
- The respondent may be unable to visualise the attributes
- There may be too many questions
- The respondent may not know the answer
- The preferences may not be stable (take me to a restaurant, ask me what I want to eat and my preferences for the food options will change every 30 seconds during the ordering process – including how many courses and what to drink!)
- The asking of the question may change the responses given.

A good example of a research technique affected by these limitations is Brand Price Trade-Off. This algorithm is very efficient in sorting a small number of responses in a way that allows a wide range of pricing questions to be analysed and answered. However as Johnson (1996) and Poynter (1997) have pointed out the technique fails because people can't answer the questions in a way that produces useful answers. Something about the process causes the responses to distort.

ACCURACY, THE PRODUCT OF METHOD AND CONTEXT

The accuracy of the models we build is dependent on two elements. One element is the ability of the algorithm to correctly interpret responses. The second element is the ability of the research context to generate the correct responses.

If researchers try to evaluate different techniques by concentrating on the algorithm and not controlling variability in the context it is likely that contradictory and unsatisfactory results will occur.

The appropriate research programme is one which compares different algorithms using comparable data and in the context of different situations. This paper argues that the best way to achieve this is by using Test Data sets to separate the two influences (algorithm and context) into controllable elements.

THE CASE FOR TEST DATA SETS

The use of hypothetical Test Data sets is quite common, for example Johnson 1997. However, in many cases these data sets are random in their nature and are designed specifically to work with a specific project.

The author proposes that Test Data sets should be created such that they are based on real data and with sufficient information that they can be used with a wide range of procedures. For example an ideal Test Data set should include answers for all attributes, for all levels that might be examined, and should include additional information such as demographics. In addition the

Test Data sets should include market information to allow the researcher to assess the implications of different markets.

The call in this paper is for a collection of such Data Sets to be created. These Data Sets should reflect different markets and different assumptions about how those markets work. Such a collection would allow the algorithmic element of different techniques to be assessed independently of other influences.

An understanding of the algorithmic effects should allow real world studies to be more appropriately used to examine issues such as:

- Respondents’ ability to answer different types of questions
- Differences between market situations (consumer durable vs Fast-Moving Consumer Good)
- Differences between respondents (students vs housewives, one culture vs another)

A TEST DATA SET

In order to explore further the ideas set out above a data set was created. The data were based on a financial services product researched in the UK. This product was relatively simple in that it comprised a charge/price element plus 8 features. Each of the features was a service option that the respondent could utilise if it were offered. Consequently the data are in theory a priori ordered on all 9 attributes.

The Attributes are as follows:

Price	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
5 Levels	2 Levels	2 Levels	2 Levels	2 Levels	3 Levels	3 Levels	2 Levels	2 Levels

Table 1

Since this study mostly comprised attributes with two levels, plus two with three levels, and one with five levels it should be possible to explore the Number of Attribute Levels Problem.

All of the attributes are of increasing utility. In order to create the data set an existing project was used (with data that had been collected using ACA plus additional modules for price and to verify scaling effects). The 200 ‘cleanest’ data items were selected (clean in terms of: high internal correlation, low levels of reversal in the utility values, sensible answers to the demographics). Four simple demographics were retained to facilitate their inclusion in tests and analyses. The preparation steps then included:

- Remove all references which could identify the original product and client;
- Add a small amount of random noise to the data, enough to protect the commercial information, but small enough not to change the ‘shape’ of the data;
- Clean the data to ensure that there are not utility reversals in the data (the object of the exercise is to create data which can be readily understood as an input to a conjoint exercise – not one which accurately reflects the outcome of a conjoint exercise)

- Apply scalings to the utilities to make them more reflective of the market, using values established in the original research;
- Convert the utilities into a form which makes them easy to compare, in this project the Points transformation, as previously defined in Sawtooth’s ACA manual was used (although the reasons for Sawtooth Software’s recent move away from this particular transformation must be acknowledged).

Table 2 shows the aggregate importance weights and utilities of the Test Data. It is these values which various tests will seek to replicate.

One feature of this data is the unequal steps in the utilities revealed in the price variable. The nature of this financial product is such that for many people if the price is too high they will use the product in a different way. This tends to result in utilities that only increase gradually until specific values are reached. Beyond this point there is a rapid increase in utility. This makes the utility structure very asymmetric – a feature that creates complexities later in the analysis.

Attribute	Importance
Price	178
F1	62
F2	80
F3	38
F4	63
F5	50
F6	72
F7	69
F8	52

Level	Utility
Price 1	0
Price 2	24
Price 3	57
Price 4	107
Price 5	178
F1 L1	0
F1 L2	62
F2 L1	0
F2 L2	80
F3 L1	0
F3 L2	38
F4 L1	0
F4 L2	63
F5 L1	0
F5 L2	17
F5 L3	50
F6 L1	0
F6 L2	30
F6 L3	72
F7 L1	0
F7 L2	69
F8 L2	0
F8 L1	52

Table 2

EXPLORING FULL PROFILE CONJOINT ANALYSIS

As a first utilisation of the Test Data it was decided that there should be an investigation of some of the key characteristics of Full Profile Conjoint Analysis. In order to do this various designs and treatments were configured using Sawtooth Software’s CVA and the Test Data. Seven Treatments were designed, each one using the values in the Test Data set to assess the impact of the Treatment.

In all but one Treatment the attributes were specified in both the design and analysis stages of CVA as a priori ordered. This means that reversals in the predicted utilities are not permitted and the results should be ‘more correct’ if the assumption of an a priori order is correct (which by definition it is in this Test Data).

CVA recommends that there should be 3 times as many tasks as unknowns. In most of the Treatments this recommendation has been followed, exceptions will be highlighted.

In all of the Treatments the method used is the rating of pairs of concepts. Where 1 indicates strong preference for the left concept and 9 indicates strong preference for the right concept. In all cases the utilities are calculated using OLS (Ordinary Least Squares).

In all but one Treatment the rating values (1 to 9) reflect the relative strength of the left and right concept (as defined by the Test Data utilities).

THE TREATMENTS

Table 3 shows a summary of the seven Treatments used in this exercise.

	Levels	Rating	Tasks	D-Efficiency
Treatment 1	All	1 to 9	45	98.9%
Treatment 2	All	1 to 9	23	95.5%
Treatment 3	All	1 to 9 with noise	45	98.9%
Treatment 4	All	2 to 9 with noise	23	95.5%
Treatment 5	All	1,5,9 only	45	98.9%
Treatment 6	2 per attribute	1 to 9	30	99.1%
Treatment 7	All, no a priori	1 to 9	45	98.8%

Table 3

Treatment 1

A CVA design of 45 tasks (the recommended number) was generated. A VBA (Visual Basic for Applications) program was created to take each of the 200 respondents from the Test Data and to use their utilities to evaluate each of the 45 pairs of concepts. The program evaluated the total utility of the left and right concept. The greater value was used as the numerator and the lesser value as the denominator of a rating term. This term was then rounded to the nearest integer and centred on 5. In a few cases the rating was outside the range 1 to 9, in these cases the rating was truncated.

Treatment 2

This treatment was identical to Treatment 1 except that it used just 23 tasks (the minimum recommended number, 1.5 times the number of unknowns).

Treatment 3

Treatment 3 starts the same as Treatment 1. Once the rating is identified a random number in the range 0 to 3 is added and another random number is deducted. The effect of this is to add noise to the scale on average of nearly one rating point. It should be stressed that the magnitude of the noise was selected arbitrarily. One anecdotal estimate of genuine error was an average of about 2 scale points.

Treatment 4

This extends Treatment 2 in the same way Treatment 3 extends Treatment 1, that the reduced number of tasks are subject to noise to see if the smaller number of tasks has more difficulty with problems created by the noise.

Treatment 5

In this treatment we start with Treatment 1, at the rating evaluation stage the following procedure is followed: if the left utility is higher than the right then the rating is set to 1, if the right is higher then the rating is set to 9, if they are equal the rating is set to 5. This process creates very 'blunt' ratings, the sort that are frequently seen in real studies.

Treatment 6

In this Treatment a new CVA design was generated, assuming that each attribute only had 2 levels, and using the recommended 30 tasks (3 times the number of unknowns). The VBA program was modified to use only the top and bottom levels from each attribute as the inputs to the left and right concepts.

Treatment 7

This treatment uses exactly the same process as Treatment 1 but with the difference that the CVA design and analysis make no assumptions of a priori ordering. This replicates the case where the researcher is not sure whether the attribute is naturally ordered and decides to 'play it safe'.

A summary of the results is shown in tables 4, 5, and 6.

	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5	Treatment 6	Treatment 7
Ave r-squared	0.97	0.96	0.89	0.81	0.87	0.99	0.97
Min r-squared	0.93	0.87	0.64	0.24	0.57	0.93	0.89
Average MAE	6	7	11	15	13	4	6
Max MAE	12	15	21	32	50	11	19
Prices reversals	0%	0%	0%	0%	0%	0%	42%

Table 4

The average r-squared values in Table 4 show the relationship between the utilities calculated by CVA and those in the Test Data used to create the responses. The minimum r-squared shows the lowest value relating one of the two hundred respondents in the Test Data to the predicted utilities. The price reversal figures show the percentage of cases where there is at least one reversal (where level n+1 has a lower utility than level n) in the predicted price utilities. In the first 6 Treatments price reversals are not possible.

The average MAE is the average across the 200 respondents of the Mean Absolute Error between the Test Data and the utilities calculated. The Max MAE is the value from the respondent with the highest MAE value.

Importance	'Real'	T 1	T 2	T 3	T 4	T 5	T 6	T 7
Price (5)	178	159	162	158	163	160	183	161
F1 (2)	62	61	62	63	62	66	60	57
F2 (2)	80	81	83	77	78	85	78	79
F3 (2)	38	37	39	38	40	36	39	37
F4 (2)	63	65	63	63	62	72	63	58
F5 (3)	50	53	52	55	53	56	50	50
F6 (3)	72	69	73	72	70	61	73	69
F7 (2)	69	69	68	68	67	76	70	68
F8 (2)	52	51	53	49	50	52	49	54

Table 5

Table 5 shows the importance values for the Test Data and for each of the seven Treatments.

Utilities	'Real'	T 1	T 2	T 3	T 4	T 5	T 6	T 7
Price 1	0	0	0	0	0	0	0	3
Price 2	24	30	22	30	27	27	*	31
Price 3	57	53	64	53	64	46	*	64
Price 4	107	108	105	105	108	106	*	
Price 5	178	159	162	158	163	160	183	161
F1 L1	0	0	0	0	0	0	0	1
F1 L2	62	61	62	63	62	66	60	57
F2 L1	0	0	0	0	0	0	0	0
F2 L2	80	81	83	77	78	85	78	79
F3 L1	0	0	0	0	0	0	0	2
F3 L2	38	37	39	38	40	36	39	37
F4 L1	0	0	0	0	0	0	0	2
F4 L2	63	65	63	63	62	72	63	58
F5 L1	0	0	0	0	0	0	0	4
F5 L2	17	27	20	28	24	21	*	15
F5 L3	50	53	52	55	53	56	50	50
F6 L1	0	0	0	0	0	0	0	1
F6 L2	30	37	35	41	33	34	*	34
F6 L3	72	69	73	72	70	61	73	69
F7 L1	0	0	0	0	0	0	0	1
F7 L2	69	69	68	68	67	76	70	68
F8 L2	0	0	0	0	0	0	0	0
F8 L1	52	51	53	49	50	52	49	54

Table 6

Table 6 shows the aggregate utilities for the Test Data and the seven Treatments. Note that in Treatment 6 there are no values for the intermediate levels of the Price attribute and Features 5 and 6. The values for Treatment 6 have been scaled to make them comparable with the other Treatments.

Before listing the four areas of investigation it should be noted that this research is based on just one set of Test Data. The findings produced could be unique to this Test Data or possibly

typical of only those cases very similar to these. This can only be established by repeating the analyses on alternative data sets. Therefore the findings in this paper should be taken as being no more than postulates.

Number of Attribute Levels Effects

The Number of Levels Effect is the phenomenon that has been observed that when an attribute is expressed in more Levels it appears to have more importance. There has been debate about whether the effect is created by the algorithms or by the behavioural changes (eg Wittink 1997).

If the Number of Levels Effect was (in the case of Full Profile Conjoint) caused by algorithmic elements we would expect the utilities for Price, and Features 5 and 6 to be overstated by comparison with the real cases, and consequently the remaining attributes to be understated.

Treatment 1 shows us that in this experiment there was no apparent Number of Levels effect. Whilst Feature 5 showed slight evidence of increased value, Feature 6 (the other 3 level attribute) showed neutral to reduction signs. In most cases the Price variable showed marked declines in importance between the Test Data and the Treatments. This price result is of particular interest because it also happened in the original study! In the original study a separate pricing module suggested that the ACA had underestimated price and it was re-scaled accordingly. At the time this effect was ascribed to respondents treating the price variable differently because it was price. This experiment suggests that there may be something about the Price utilities themselves that could contribute to the underestimation of their values by Conjoint Analysis.

To explore this issue further Treatment 6 was constructed using only the top and bottom levels of each attribute. This produced results that were largely similar to Treatment 1. In Treatment 6 Price appears to be more important than in Treatment 1, and very close to the Price value in the Test Data. This result raises several questions both about the Number of Levels Effect and also about whether there is something intrinsic in the shape and pattern of Price utilities that creates analysis problems.

These findings support those people who argue that the reason for the Number of Levels Effect may have its causes in behavioural reasons (eg Green 1990). However, even if this were the case the practitioner would be well advised to try and keep the number of levels equal for all attributes, until good methodological remedies to this phenomenon have been established.

Given the findings from this experiment it would be useful to repeat the analysis using other choice techniques and additional data sets to see if the findings can be reproduced and generalised.

Number of Tasks (and Errors)

Sawtooth Software recommends in their CVA manual that the number of tasks is three times as great as the number of unknowns. Indeed a warning appears in the software if the user tries to use fewer tasks than 1.5 times the number of unknowns. Practitioners are familiar with building in more redundancy to provide extra degrees of freedom to allow for unforeseen problems. The Test Data allows us to review some of the implications of doing this.

Treatment 2 reduces the number of tasks for the default case from 45 to 23 (the minimum recommended). The aggregate data in Tables 5 and 6 show that these fewer tasks, given perfect responses, produce results that are almost indistinguishable from those supplied by 45 tasks. If respondents were capable of perfect responses the cost and possible fatigue effects of 45 tasks would never be justified. The r-squared values and the Mean Absolute Error figures for Treatments 1 and 2 show that the fewer number of tasks slightly increases the risk that an individual respondent will be less adequately represented by the Conjoint process.

In an attempt to move closer to the real world Treatments 3 and 4 introduced error into the responses. Once the correct 1 to 9 values had been calculated (to evaluate a pair of concepts), a random element of noise was added to this rating score. The rating was randomly moved up or down by up to 3, with an average movement of 0.86. Most practitioners would consider this a modest amount of noise.

The introduction of noise made little impact on the aggregate estimation of importance and utility. But the average r-squared and MAE figures show that under the calm of the aggregate data the individual cases were being estimated much less well. Given that the noise that was added was random and 0-centred it is not surprising that the aggregate estimates remain largely unaffected. The difference between the 45 tasks and the 30 tasks is not particularly marked at the level of the average r-squared and MAE levels. However, the worst case situation for both r-squared and MAE is much worse in the case of 30 tasks, compared with 45 tasks.

This data suggest that increasing the number of tasks has relatively little impact in terms of aggregate utilities. However, if a study were to use the disaggregated data, for example for simulations or segmentation, the larger number of tasks would be expected to produce safer data.

A Priori

Most of the treatments take advantage of the fact that the data is known to be ordered in terms each of 9 attributes having increasing utility values for their levels. This information is used in creating the tasks and also in the analysis to bar reversals. However, in a real world study the practitioner often does not know whether attributes should be a priori ordered. If a practitioner could not decide whether an attribute truly has an a priori order he/she may well decide to simply leave the attribute as unordered.

In order to investigate the impact of not specifying attributes as a priori ordered Treatment 7 was created with CVA turning a priori off for all attributes and for both design and utility calculation.

In terms of Aggregate importance, average r-squared, and average MAE the decision to leave the attributes as not having an order has made no discernible difference. In terms of worst case results the unordered is slightly worse than Treatment 1, but not by an amount that would cause concern.

The aggregate utilities for Treatment 7 are very similar to those for Treatment 1. Although to the casual reader the fact that several of the 0 utilities have turned to values in the range 1 to 4 would cause comment.

However there is a serious concern for the modeller. In 42% of cases at least one of the 5 price levels is reversed (ie level n+1 has less utility than level n). It should be noted this is

entirely an artifact of the processing, the Test Data replied without errors and without changing their mind during the interview.

Should this finding be replicated with other data sets the inference for the researcher is that there is a penalty for incorrectly leaving an attribute as unordered when it should be ordered. When using a technique such as Full Profile the decision can be made post facto and different options explored. With a technique such as ACA the decision needs to be made before the main interview stage (increasing the need for a pilot?).

Blunt Responses to Rating Questions

Most practitioners will be familiar with those respondents who take a 9-point scale and use it as a 3-point scale (ie 1 prefers the left, 9 prefers the right, 5 for no preference and/or don't know). The author's experience is that different cultures tend to have larger or smaller proportions of respondents who answer in this simplified way. For example the author would assert that Japan has fewer extreme raters and Germany has more.

In order to assess the impact of these blunter ratings Treatment 5 was introduced. In Treatment 5 the initial rating score of 1 to 9 is simplified into 1, 5, or 9.

The results for Treatment 5 show that the aggregate importance and utilities are very similar to Treatment 1. The average r-squared and MAE values show that there is some deterioration in the values for Treatment 5 (about the same as introducing the random error in Treatments 3 and 4). However the worst cases for Treatment 5 is very poor.

If this result generalises to other data sets the inference is that researchers should try to minimise these blunt ratings and try to find analyses that are more tolerant to them. In cases where all the ratings are as blunt as Treatment 6 the researcher may question the use of OLS, but in cases where some of the ratings are 1 to 9 and some blunt would a researcher be happy to mix utility calculation techniques?

FURTHER RESEARCH STEPS

There are two steps that would be interesting to follow but that have not yet been pursued.

Firstly, the existing Test Data could be used to explore more Conjoint related topics. For example the test data could be used to examine all of the following:

- ACA versus Full Profile
- BPTO vs PEP (Purchase Equilibrium Pricing, Poynter 1997)
- Choice Based Conjoint versus conventional conjoint
- The extent to which Hierarchical Bayes can reproduce original responses

Secondly, additional Test Data sets need to be created. Test Data sets are needed to explore the experiences of practitioners. Market Researchers come across a wide range of research situations in terms of products, research objectives, markets, and cultures. A collection of Test Data sets would help the industry to understand if some techniques are algorithmically more suited than others to a specific need.

This paper is a call for the creation of a publicly accessible collection of Test Data sets, to allow fuller examination of the algorithms that underpin the daily life of the Choice Modeller.

Beyond these steps research is needed into finding out more about the behavioural and psychological issues which reduce the effectiveness of Choice Based techniques.

The use of Test Data will help us understand the algorithmic efficacy of a technique but we must never lose sight of the fact that algorithmic efficacy is necessary but not sufficient. The final result will be a product of the algorithm, the ability of the respondent to provide suitable answers, and the ability of the researcher to ask the right questions!

REFERENCES

- Green, P. and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, October 1990.
- Johnson, R. (1997), "Individual Utilities from Choice Data: A New Method," Sawtooth Software Conference, Seattle, USA.
- Johnson, R. and K. Olberts (1996), "Using Conjoint in Pricing Studies: Is One Pricing Variable Enough?" Sawtooth Software Technical Paper.
- Moore, W. L., J. Gray-Lee and J. Louviere (1998), "A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation," *Marketing Letters*.
- Poynter, R. (1997), "An Alternative Approach to Brand Price Trade-Off," Sawtooth Software Conference, Seattle, USA.
- Vriens, M., H. Oppewal and M. Wedel (1998), "Ratings-Based Versus Choice-Based Latent Class Conjoint Models – An Empirical Comparison," *JMRS* July 1998.
- Wittink, D., W. McLauchlan, and P. Seetharaman (1997), "Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis," Sawtooth Software Conference, Seattle, USA.

CLASSIFYING ELEMENTS WITH ALL AVAILABLE INFORMATION

Luiz Sá Lucas

IDS-Interactive Data Systems

DataWise-Data Mining and Knowledge Discovery

ABSTRACT

The paper describes a neural network classification procedure that combines classifying solutions from different sources. It also shows how different approaches for well known techniques such as Discriminant Analysis and Automatic Interaction Detection can be used to minimize the problems due to the different predictive capability when classifying elements into distinct groups.

INTRODUCTION

Segmentation is key issue in Marketing Research and Database Marketing. After a segmentation is obtained from an original sample, another problem naturally arises: how should we classify a new set of elements into the original segmentation?

This question may arise in the context of Marketing Research, for example, when screening customers for focus groups. This was, by the way, the real world problem that originated the technique described in this paper. We had previously built a 3-Group Segmentation for O GLOBO, a newspaper from Rio de Janeiro, Brazil. The idea was to go in more depth into some issues that arose in the segmentation step, analysing separately people from the three different groups. Another possible application is when we have a new wave of some survey and want to classify the new sample into the segmentation defined in the first wave. We could think even on a third possible application: as a predictive model for categories such as ‘user / non user’ of a product or as ‘light / medium / heavy’ product usage. In Database Marketing we can think on similar applications.

In the paper we initially highlight the main cluster analysis issues that may affect the ability to classify new elements into pre-existent clusters. Then we describe what we think are the main methods for classifying elements into any previously defined grouping. Finally we show how to integrate these results with the help of a neural network, illustrating the procedure with an example.

CLUSTERING TECHNIQUES

Unless some market is segmented by some objective, clearly defined criteria like ‘heavy / medium / light’ buyers, segmentation must be performed using some clustering method.

Clustering methods are described in several sources (see for example Anderberg (1973), Sá Lucas (1993) and CCA-Sawtooth Software(1996)). Basically what we aim with these methods is to ‘organize objects into groups so that those within each group are relatively similar and those in different groups are relatively dissimilar...we also hope that the groups will be “natural” and “compelling” and will reveal structure actually present in the data’ (CCA-Sawtooth Software).

Another way to look at a cluster analysis problem is as follows (Sá Lucas(1993)). Let's suppose we have a set of elements or objects (customers, for example). Let's suppose also that we have a measure of association between every pair of elements (a correlation or similarity coefficient, for example). The problem can be summarized in the sentence: "Based on the information given by the association coefficients, determine a partition of the initial set in a fixed number of groups, allocating each object to only one group, in such a way that similar objects are gathered in the same cluster while dissimilar elements are allocated to distinct groups".

We could roughly classify the clustering methods into three classes:

1. Hierarchical procedures
2. Iterative Relocation methods
3. Mathematical Programming methods

Hierarchical methods are quite common in its agglomerative variant. The technique starts with each one of the N elements forming a group (that is we have N groups). Then the most similar groups (or elements, in the first step) are joined, forming a total of $(N-1)$ groups. The process is repeated for $(N-2)$, $(N-3)$... groups until some pre-defined convergence criteria is satisfied. The method suffers from a 'chaining effect' problem: if a group is formed in a previous step, it cannot be rearranged afterwards in order to maximize homogeneity.

An Iterative Relocation method proceeds as follows: given an initial partition for a fixed number of groups, the algorithm generates a new partition that improves the overall homogeneity. The process is repeated until a pre-defined convergence criteria is satisfied. The number of groups remains fixed during computations.

Usually the center of the groups are averages (also called centroids) and homogeneity within a group is measured using some distance function between elements that belong to the group and the centroid of this group. The most common distance function is the euclidean distance function. The most common of these methods is the well known 'K-Means' method.

There is, although, an interesting variant of this method that, instead of using centroids as centers of groups uses the concept of 'median'. The median of a group is the element of the group such that the sum of distances to the remaining elements of the same group is minimum. The method is known as the 'K-Medians' method and is due to Mulvey and Crowder (1979). It is also described, for example, in Sá Lucas (1993). The advantage of the method is that it can be used with any kind of variable (nominal, ordinal, interval, ratio), provided we use an appropriate similarity coefficient where a distance function can be derived.

We can also broaden the scope of iterative relocation methods to include clustering / non supervised neural networks such as the Kohonen neural net (Kohonen (1997)). The Kohonen net has the advantage of not only providing a clustering solution for a given sample but also can be used as a classifier for new observations. In this sense it could be used as a substitute for the algorithm that is the subject of this paper or it can be used as another input for the final consolidating Feed Forward Neural Net model that we will describe later.

Finally Mathematical Programming methods implement optimization algorithms to perform the clustering. Here we could quote the Subgradient Optimization method due to Mulvey and

Crowder (1979) and the Greedy Algorithm (see for example Sá Lucas (1993)). Both methods have the ‘median’ feature described above.

Besides the fact that different methods can produce different solutions, and from a practical point of view, there are other two quite difficult and interrelated problems associated with the daily implementation of clustering / segmentation procedures in Marketing Research and DBM and that have a lot to do with our classifying procedure:

- The fact that the final clusters are ‘natural’, corresponding to clearly separated and well defined groups
- The fact that the cluster solution is ‘reproducible’, a concept that, for our purposes, can be defined as the ability to reproduce similar clustering solutions for different data sets.

Clearly ‘natural’ clusters must be the goal but the degree of ‘naturalness’ is quite difficult to assess in an exact mathematical way. More about ‘reproducing’ ability can be found in CCA-Sawtooth Software (1996). Anyway we understand that the ability of classifying new observations into a previous clustering scheme will depend heavily on ‘naturalness’ and reproducing ability of the clustering solution.

CLASSIFYING NEW ELEMENTS

Provided we have some partitioning scheme (either defined by some ‘heavy / medium / light’ definition or given by a clustering method such as those described in the previous section) we can classify new observations using several different methods such as, for example:

- Automatic Interaction Detection
- Discriminant Analysis
- Logistic Regression
- Classification by Chi-square Rule
- Neural networks

Automatic Interaction Detection – AID (here taken as a general term for the family of techniques that include CHAID, CART and other similar methods) define a decision tree, or equivalently, a Rule like:

```
IF                conditions1
THEN              Prob(Group1)=P11 / Prob(Group2)=P12/.../Prob(GroupK)=P1K
ELSEIF           conditions2
THEN              Prob(Group1)=P21 / Prob(Group2)=P22/.../Prob(GroupK)=P2K
.....
ELSEIF           conditionsM
THEN              Prob(Group1)=PM1 / Prob(Group2)=PM2/.../Prob(GroupK)=PMK
```

In this decision scheme any observation will certainly fall into some condition, so we can allocate the observation to the Group with greatest membership probability in that only satisfied condition. More details about the technique can be found in Berry and Linoff (1997). For more

general AID algorithms the classification variables can be of any nature from nominal to ratio scale.

Discriminant Analysis is a classic well known classification technique described, for example, in Tatsuoka (1971) and Cooley and Lohnes (1971). In Discriminant Analysis for each group a Fisher's Linear Discriminant Function is calculated, so that a new observation should be allocated to the group where the Fisher's function is largest. The classification variables must be continuous or binary.

Logistic Regression (Aldrich and Forrest (1984)) can be used when we have only two groups. As we will see later, this is not so restrictive as it may seem. Here the classification variables must also be continuous or binary. Anyway, results for two-Group classification use to be more reliable with Logistic Regression than with Discriminant Analysis (see for example Fichman (1999) or Pereira (1999)). The reason for this fact can be found in Ratner (1999): Discriminant Analysis is more sensitive to violations of its basic hypothesis than Logistic Regression, a more robust technique that should be preferred in the two-group case.

Classification by Chi-square rule is another classic technique described, for example, in Tatsuoka (1971) and Cooley and Lohnes (1971). Here the criteria is to assign an element to the group where the chi-square distance is smallest (the probability of group membership is largest). Again the variables should be continuous or binary.

Finally a quite flexible classifying scheme could be obtained with the use of neural network. Here we could use either a Feed Forward Net, a Probabilistic Neural Net – PNN or a Kohonen net (see Sá Lucas (1998) or Berry and Linoff (1997) for a brief description of these nets). Even here the variables should be continuous or binary.

A final word should be said about two-group or multiple group classification. Anyone involved with classification procedure must have felt that the predictive power is quite different for different groups. The more defined / natural / separate a group is, the better must be the predictive power of classifying an element into this group. We will illustrate that in our example. On the other hand, it should be easier to classify an element into a two group scheme than into a multi-group one. For example, suppose we have a 3-group solution. We can set a classification procedure that handles the 3 groups simultaneously or alternatively we can define three different problems like:

1. Element belongs to Group 1 / Element does not belong to Group 1
2. Element belongs to Group 2 / Element does not belong to Group 2
3. Element belongs to Group 3 / Element does not belong to Group 3,

provided we have a consolidating procedure like the one we will describe later. It is in this area that Logistic Regression can be attractive.

CONSOLIDATING DIFFERENT CLASSIFYING SOLUTIONS WITH A NEURAL NET

The different methods described above can produce different information that can be used as input for the Final Consolidating Net – FCN:

- AID, Logistic Regression and Classification by Chi-square Rule can produce group membership probabilities
- Discriminant Analysis usually produces Fisher’s Linear Discriminant values
- A Kohonen network will allocate the element to a single group (value of input equal to 1) so that for a K-Group solution all the remaining K-1 variables will be equal to 0 (zero)
- The results of a Feed Forward and PNN nets can also produce probabilities (if they are not normalized to sum up to 1, that can be done prior to input them into the final consolidating neural FCN net).

So let’s suppose that we have a training data set for a 3-Group problem where the group membership is known (as we said before, the cluster membership will be given by a clustering procedure or by a ‘heavy / medium / light’ classification). Let’s also suppose that we have estimated group membership with the following methods and derived input variables for the final net:

- D0-Discriminant Analysis with three groups – three variables with Fisher’s function – FF for each group
- D1-Two-Group Discriminant Analysis for Group 1 – one variable with FF for Group 1
- D2-Two-Group Discriminant Analysis for Group 2 – one variable with FF for Group 2
- D3-Two-Group Discriminant Analysis for Group 3 – one variable with FF for Group 3
- A0-Automatic Interaction Detection - AID with three groups – three variables with group membership probability – GPM for each group
- A1-One two-group AID for Group 1 – one variable with GPM for Group 1
- A2-One two-group AID for Group 2 – one variable with GPM for Group 2
- A3-One two-group AID for Group 3 – one variable with GPM for Group 3

We would have a total of 12 variables to input the final consolidating neural network. This is exactly what we have implemented in our real world case. The results, in a disguised version, are presented in the table below:

	D0	D1	D2	D3	A0	A1	A2	A3	FCN
Group1	93.2%	93.3%	---	---	82.7%	90.7%	---	---	94.0%
Group2	80.5%	---	82.3%	---	79.7%	---	88.0%	---	86.0%
Group3	78.6%	---	---	84.4%	82.7%	---	---	83.9%	88.0%
Overall	83.3%	---	---	---	80.7%	---	---	---	89.0%

The percentage figures correspond to the success rate when reclassifying elements from the original training data set using the several algorithms. In each column we have a heading where D0 stands for a the first approach for Discriminant Analysis, A1 for the second approach to Automatic Interaction Detection and so on.

Methods D1, D2, D3, A1, A2 and A3 are specific for each group (D1 and A1 for Group1, for example). We verify that this two group approach (D1 / D2 / D3 and A1 / A2/ A3) had a better success rate than the three group approaches D0 and A0.

On the other hand we can see that the success rate for each group in D0 and A0, the only methods that work with three groups at the same time, are not equal: they differ from one group to another (in D0 the success rate was 78.6% for Group 3 and 93.2% for Group1...).

We can also see that the Final Consolidated Net improved the overall success rate and made it more even among groups, although Group1 still has the best classifying performance.

CONCLUSION

The success rates presented in the preceding section may be a little optimistic about the performance of the FFN method, since the test was performed on the same training sample where the algorithm was calibrated. A test sample would be better, but our sample was quite small for that purpose. Anyway the issue here is to assess the improvement in the classification performance.

Finally we want to stress the question of the ‘naturalness’ of the previous grouping that is the basis for the subsequent classification. Surely ‘natural’ solutions should be the goal in any cluster analysis, but in daily work some problems may occur. For example:

- The most ‘natural’ solution is a two group solution, useless for Marketing purposes. Maybe a, say, 5 group less natural solution may be preferable
- In classifications such as ‘heavy / medium / light’ almost certainly these three groups would not correspond to a ‘natural’ solution
- The degree of homogeneity (‘naturalness’) among groups will vary in general

In other words, in daily work we will have to live with a certain degree of ‘non-naturalness’ of the solutions. This is the reason we have to try to have the best possible classifying procedure. If all solutions were clearly defined any simple algorithm would suffice...

REFERENCES

- Aldrich, J. and Nelson, F. (1984). *Linear Probability, Logit and Probit Models*. Sage Publications.
- Anderberg, J. (1973), *Cluster Analysis for Applications*, Academic Press.
- Berry, M. and Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley & Sons
- Cooley, W. and Lohnes, P. (1971). *Multivariate Data Analysis*, John Wiley & Sons
- Kohonen, T. (1997), *Self-Organizing Maps*, Springer
- Fichman, L. (1999), “Construção de um Modelo de Predição de Insolvência Bancária Baseado na Tipologia de Porter”, Departamento de Administração – PUC/RJ.
- Mulvey, J.M. and Crowder, H.P.(1979), “Cluster Analysis: an Application of Lagrangian Relaxation”, *Management Science*, Providence, 25(4):329-40
- Pereira, R. and Esteves W. (1999), “Análise de Solvência de Bancos Utilizando Análise Discriminante e Regressão Logística”. ENCE/IBGE-RJ
- Ratner, B. (1999) “Response Analysis” in David Shepard Associates, “The New Direct Marketing”, McGraw Hill.
- Sá Lucas (1993), “GCA Algorithms and their Application to Image and Segmentation Problems”, *Proceedings of the 2nd Latin American European Society for Opinion and Marketing Research -ESOMAR Conference, Mexico City (Mexico)*.
- Sá Lucas (1998), “Data Mining vs. Conventional Analysis: When Should We Use the New Techniques?”, *Proceedings of the 51st ESOMAR Marketing Research Congress, Berlin (Germany)*.
- Sawtooth Software (1996), CCA-Convergent Cluster Analysis System, <http://www.sawtoothsoftware.com>.
- Tatsuoka, M. (1971). *Multivariate Analysis*. John Wiley & Sons

AN OVERVIEW AND COMPARISON OF DESIGN STRATEGIES FOR CHOICE-BASED CONJOINT ANALYSIS

Keith Chrzan

Maritz Marketing Research

Bryan Orme

Sawtooth Software, Inc.

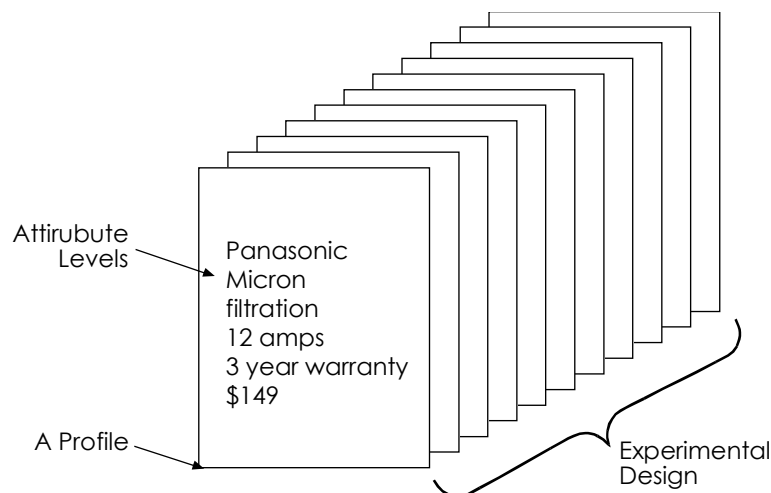
There are several different approaches to designing choice-based conjoint experiments and several kinds of effects one might want to model and quantify in such experiments. The approaches differ in terms of which effects they can capture and in how efficiently they do so. No single design approach is clearly superior in all circumstances.

This paper describes different kinds of design formats (full profile, partial profile), and different methods for making designs (manual, computer optimization, computer randomization) for choice-based conjoint designs. Over and above the plain vanilla generic main effects most commonly modeled in conjoint analysis, there are several types of “special effects” that can be included in choice-based models. The various ways of constructing choice-based designs are compared in terms of their ability to capture these effects. Using simulations and artificial data sets we also assess the statistical efficiency of the various design methods.

BACKGROUND

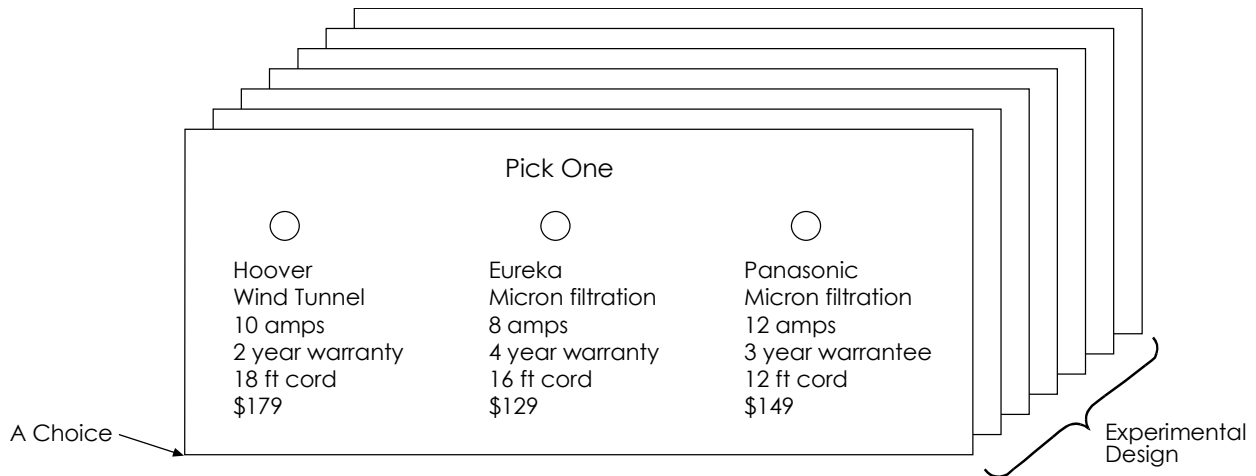
In traditional conjoint analysis (see Figure 1), experimentally controlled combinations of attribute levels called profiles are presented to respondents for evaluation (ratings or rankings). In a multiple regression analysis these evaluations then become the dependent variables predicted as a function of the experimental design variables manifested in the profiles.

Figure 1: Traditional Ratings-Based Conjoint



In 1983, however, Louviere and Woodworth extended conjoint analysis thinking to choice evaluations and multinomial logit analysis. In the choice-based world, respondents choose among sets of experimentally controlled sets of profiles and these choices are modeled via multinomial logit as a function of the experimental design variables.

Figure 2: Choice-Based Conjoint Experiment



As you might guess, the greater complexity of the experiment allows the researcher to think about designing and estimating many more interesting effects than the simple main effects and occasional interaction effects of traditional conjoint analysis (Louviere 1988, Anderson and Wiley 1992, Lazari and Anderson 1994).

In addition to focusing on the novel effects choice-based analysis allowed, other topics became important for choice-based analysis. Design efficiency became a topic of research because the efficiency of experimental designs for multinomial logit was not as straightforward as that for traditional linear models and their designs (Kuhfeld *et al.* 1994, Bunch *et al.* 1994, Huber and Zwerina 1995). Finally, still other researchers sought ways to make choice-based experiments easier for researchers to design (Sawtooth Software 1999) or for respondents to complete (Chrzan and Elrod 1995).

CHARACTERIZING EXPERIMENTAL DESIGNS

Stimulus Format

In choice-based experiments, stimuli can be either full profile (FP) or partial profile (PP). Full profile experiments are those that display a level from every attribute in the study in every product profile. Partial profile experiments use profiles that specify a level for only a subset (usually 5 or fewer) of the attributes under study. Full and partial profile stimuli for a 10 attribute vacuum cleaner study might look like this:

Figure 3 – Full vs Partial Profile Stimuli

Full			Partial		
Hoover	Eureka	Panasonic	Hoover	Eureka	Panasonic
9 amps	12 amps	10 amps	9 amps	12 amps	10 amps
12 ft cord	16 ft cord	24 ft cord	Edge cleaner	Edge cleaner	-
Dirt Sensor	-	-			
-	Micro filter	Micro filter			
1 yr warranty	2 yr warranty	6 mo warranty			
Edge cleaner	Edge cleaner	-			
Flex hose	-	Flex hose			
-	Height adj.	Height adj.			
\$249	\$199	\$299			

Generating Choice-Based Experiments

Three broad categories of experimental design methods for choice models are a) manual, b) computer optimized, and c) computer randomized.

Manual

Strategies for creating full profile designs start with traditional fractional factorial design plans. Consider a four-attribute conjoint study with three levels each, commonly written as a 3^4 experiment. (Note that this notation reflects how many possible profiles can be constructed: $3^4 = 81$ profiles, representing the full factorial.) The figure below shows a 9 run experimental design from the Addelman (1962) catalog for a 3^4 design, and how it would be turned into 9 profiles in a traditional full profile ratings or rankings based conjoint experiment. In this design plan, each column represents an attribute whose three levels are uncorrelated (orthogonal) with respect to each other. In a traditional conjoint experiment, each row would specify a single profile.

Figure 4: 3^4 Addelman Design for Profiles

Profile	V1	V2	V3	V4
1	1	1	1	1
2	1	2	2	3
3	1	3	3	2
4	2	1	2	2
5	2	2	3	1
6	2	3	1	3
7	3	1	3	3
8	3	2	1	2
9	3	3	2	1

Traditional fractional factorial designs were designed for creating sets of single profiles, so they need to be adapted if they are to be used to generate sets of choice sets. The original (Louviere and Woodworth 1983) methods are no longer in widespread use but there are three adaptations of fractional factorial designs that are.

The simplest of these comes from Bunch *et al.* (1994) and is called “shifting.” Here’s how shifting would work for an experiment with four attributes each at three levels:

1. Produce the 9 run experimental design shown above. These runs define the first profile in each of 9 choice sets.
2. Next to the four columns of the experimental design add four more columns; column 5 is just column 1 shifted so that column 1's 1 becomes a 2 in column 5, 2 becomes 3 and 3 becomes (wraps around to) 1. The numbers in column 5 are just the numbers in column 1 "shifted" by 1 place to the right (and wrapped around in the case of 3). Likewise columns 6, 7 and 8 are just shifts of columns 2, 3 and 4.

Figure 5: 3⁴ Shifted Design

Set	Profile 1				Profile 2				Profile 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	1	1	1	1	2	2	2	2	3	3	3	3
2	1	2	2	3	2	3	3	1	3	1	1	2
3	1	3	3	2	2	1	1	3	3	2	2	1
4	2	1	2	2	3	2	3	3	1	3	1	1
5	2	2	3	1	3	3	1	2	1	1	2	3
6	2	3	1	3	3	1	2	1	1	2	3	2
7	3	1	3	3	1	2	1	1	2	3	2	2
8	3	2	1	2	1	3	2	3	2	1	3	1
9	3	3	2	1	1	1	3	2	2	2	1	3

3. The four columns 5-8 become the second profile in each of the 9 choice sets. Note that the four rows just created are still uncorrelated with one another and that the value for each cell in each row differs from that of the counterpart column from which it was shifted (none of the levels "overlap")
4. Repeat step 2, shifting from the values in columns 5-8 to create four new columns 9-12 that become the third profile in each of the 9 choice sets.
5. Replace the level numbers with prose and you have a shifted choice-based conjoint experimental design.

Shifted designs are simple to construct but very limited in terms of what special effects they can capture (described later).

A second way of using fractional factorial designs is a "mix and match" approach described in Louviere (1988). A few more steps are involved. For the 3⁴ experiment, for example:

1. Use 4 columns from the Addelman design to create a set of 9 profiles. Place those in Pile A.
2. Use those 4 columns again, only this time switch the 3's to 1's in one (or more) of the columns and the 1's to 3s, etc., so that the 9 rows are not the same as in step 1. Create these 9 profiles and place them in Pile B.
3. Repeat step 2 to create a third unique set of profiles and a new Pile C.
4. Shuffle each of the three piles separately.
5. Choose one profile from each pile; these become choice set 1.

6. Repeat, choosing without replacement until all the profiles are used up and 9 choice sets have been created.
7. A freebie: you could have set aside the attribute “Brand” and not included it in the profiles. In step 4 you could label each profile in Pile A “Brand A,” each profile in Pile B “Brand B” and so on. The Brand attribute is uncorrelated with any other attribute and is a lucky side benefit of having constructed your design in this way. This freebie also allows designs to support estimation of alternative specific effects described below.

A very general and powerful way to use fractional factorial designs is called the L^{MN} strategy (Louviere 1988). One can use an L^{MN} design when one wants a design wherein choice sets each contain N profiles of M attributes of L levels each. For our small example, let’s have N=3, M=4 and L=3 (still the small 3^4 experiment with 4 attributes of 3 levels each). This approach requires a fractional factorial design with N x M columns of L level variables. It turns out that for such an experiment the smallest design has 27 rows (Addelman 1962). Taking 12 of the columns from the Addelman design and placing them in groups of four each is the hardest part of the design:

Figure 6: $3^{12} L^{MN}$ Design

	Profile 1				Profile 2				Profile 3			
Set												
1	1	2	1	3	2	2	3	1	1	3	3	2
2	3	2	3	1	3	1	3	1	2	3	1	2
3	1	2	3	3	2	3	2	2	2	1	2	3
.
.
.
.
27	3	3	2	1	3	2	1	2	1	1	3	3

The L^{MN} design now requires just one step, because all three profiles come directly from each row of the fractional factorial design: The first 4 columns become the 4 attributes in profile 1, columns 5-8 describe profile 2 and columns 9-12 describe profile 3. No shifting or mix and match are necessary.

The larger 27 choice set design in this example is typical of L^{MN} designs. This cost buys the benefit of being able to support “mother logit” analysis of cross effect designs described below (caution: this is true only if each choice set includes a “none” or “other” response).

For manual generation of partial profile stimuli a design recipe can be used. Design recipes for profiles with 3 or 5 attributes appear in the Appendix of a 1999 Sawtooth Software Conference paper (Chrzan 1999). A new design recipe for partial profiles with just two attributes, and suitable for telephone survey administration is available upon request.

Randomized Designs

Randomized designs are used in Sawtooth Software’s CBC product. A random design reflects the fact that respondents are randomly selected to receive different versions of the choice sets. Those choice sets are created in carefully specified ways. CBC allows the user to select one of four methods of design generation:

1. In Complete Enumeration, profiles are nearly as orthogonal as possible within respondents, and each two-way frequency of level combinations between attributes is equally balanced. Within choice sets, attribute levels are duplicated as little as possible (a property called “minimal overlap”), and in this sense this strategy resembles the shifting strategy described earlier.
2. In Shortcut, profiles for each respondent are constructed using the least often previously used attribute levels for that respondent, subject again to minimal overlap. Each one-way level frequency within attributes is balanced.
3. The Random option uses profiles sampled (randomly, with replacement) from the universe of possible profiles and placed into choice sets. Overlap can and does occur, though no two profiles are permitted within a choice set that are identical on all attributes.
4. Finally, the Balanced Overlap approach is a compromise between Complete Enumeration and Random – it has more overlap than the former and less than the latter.

Please see the CBC documentation for a description of these different kinds of randomized designs (Sawtooth Software 1999). Depending on the extent of overlap, these types of randomized designs are differently able to measure special effects and differently efficient at measuring main effects. It turns out that designs with little or no level overlap within choice sets are good at measuring main effects, while designs with a lot of overlap are good at measuring higher-order effects.

Computer Optimization

Kuhfeld *et al.* (1994) discuss how to use computer search algorithms in SAS/QC to assess thousands or millions of potential designs and then pick the most efficient. The authors find substantial efficiency improvements even in traditional conjoint analysis when those designs are asymmetric (when they have different numbers of levels). Computer optimization enables the researcher to model attributes with large numbers of levels or complex special effects. Huber and Zwerina (1996) add the criterion of utility balance to further improve computer optimization of designs. New SAS macros have been added specifically for generating efficient choice experiment designs (Kuhfeld, 2000). Please refer to these papers for further details.

SPSSTM Trial Run can be used to generate computer optimized designs (SPSS 1997) as can Sawtooth Software’s CVA (Kuhfeld 1997). Their design strategies are usually suitable for traditional (one profile at a time) conjoint designs, but their capabilities are limited when it comes to designing choice experiments.

TYPES OF EFFECTS

Before comparing these design strategies, the various types of effects on which they are evaluated require explication.

Generic, plain vanilla, main effects

The basic kind of effect in all types of conjoint studies is the utility of each level of each attribute. Designs that produce only main effects are, not coincidentally, called main effects designs. Each main effect measures the utility of that level, holding everything else constant (at

the average combination of other levels used in the study). Traditional conjoint analysis typically produces only main effects.

Interactions

Interactions occur when the combined effect of two attributes is different from the sum of their two main effect utilities. For example, being stranded on a deserted island is pretty bad, say it has a utility of -40. Attending a party hosted by cannibals is also a bad thing, say with a utility of -50. But, attending a party hosted by cannibals on a deserted island could be altogether worse, in grisly sorts of ways (utility -250). Or again, being naked is a modestly good thing (+3) and speaking at the Sawtooth Software Conference is a +10, but speaking naked at the Sawtooth Software Conference is a -37.

Alternative specific effects

When not all of the alternatives (say brands or technologies) in a choice experiment share exactly the same attributes or levels, the non-shared effects are said to be alternative specific. In the simplest case brands may have different levels, so that brands might range in price between \$500 and \$1,500, say, while generics range from \$300 to \$700. But, the more complex case may include alternatives having different levels and even different numbers of levels for an attribute.

The other kind of alternative specific effect allows different alternatives to have different attributes altogether. For example, I can walk to work, drive, or ride the train. All three alternatives have transit time as an attribute. Driving myself has gas cost and parking fees as attributes not shared with the other alternatives. Similarly, taking the train involves a wait time and a ticket fare as unique attributes. Driving, walking and taking the train have some attributes in common and some not. The attributes not shared by all three are alternative specific. The advantage of alternative specific effects is that they obviously allow modeling of a much wider range of choice situations than traditional conjoint analysis which requires all profiles to share the same attributes and levels.

Cross-effects

A vendor sells Coke, Sprite, and Miller Beer in 5:3:2 proportion. What happens if Pepsi becomes available and takes 10 share points? According to the simple logit choice rule, it will draw proportionally from each other alternative, taking 5 points from Coke, 3 from Sprite and 2 from Miller. Common sense, however, says that Pepsi is likely to take more share from Coke and very little indeed from Miller. But the multinomial logit model will not allow this unless you trick it, and the trick is to include what are called cross effects. A cross effect in this example would be a part worth utility that penalizes Coke, Sprite and Miller differently when Pepsi is present, so that Pepsi draws proportionally more share (say 8 points) from Coke, and proportionally less (say 2 and 0 points respectively) from Sprite and Miller. These cross effects are also called availability effects.

Cross effects can be used to permit asymmetric share draws from other attributes besides brand. In a study of personal computers, for example, one might expect asymmetric share draws to affect PC brand, price, microprocessor speed, etc.

COMPARISONS OF DESIGN STRATEGIES

Two comparisons of the above design strategies involve identifying which of the above special effects each design strategy can accommodate and quantifying the statistical efficiency of the various design strategies under different conditions.

The capabilities of the various design strategies are compared in Exhibit 1, where an “X” indicates that that design strategy may be used under various conditions or to estimate certain effects.

**Exhibit 1: Comparison of Capabilities
Design Method**

Effects	Full Profile							Partial Profile	
	FF Shift	FF Mix & Match	FF L ^{MN}	CBC Complete Enum.	CBC Shortcut	CBC Random	CBC Balanced Overlap	Computer Optimiz. Recipe/CBC	
Main Effects only	X	X	X	X	X	X	X	X	X
Interactions		X	X	X	X	X	X	X	?
Prohibitions				X	X	X	X	X	?
Alternative Specific Effects		X	X		X	X		X	?
Cross Effects			X			X		X	
Many attributes									X
Telephone administration									X

We assessed the efficiency of the various design strategies via a series of simulation studies. For each, we created data sets whose respondents (n=300) had mean zero vectors of utilities (random responses). We tested the efficiency of alternative manual, computer optimization and computer randomization (CBC software) design methods in estimating the several types of effects. We estimated parameters using CBC and LOGIT (Steinberg and Colla 1998) software.

Exhibit 2: Comparison of Relative Efficiencies Design Method

Effects	FF	FF Mix	FF	CBC		CBC		CBC	
	Shift	&Match	L ^{MN}	Complete Enum.	CBC Shortcut	CBC Random	Balanced Overlap	Computer Optimiz.	Recipe
Main effect FP, symmetric ¹	100%	ni	ni	100%	100%	68%	86%	100%	na
Main effect FP, asymmetric ²	99%	ni	ni	100%	100%	76%	92%	98%	na
Generic partial profile ³	na	na	na	na	100%	66%	na	ni	95%
FP, few interactions ⁴	ne	90%	ni	94%	94%	90%	97%	100%	na
FP, many interactions ⁵	ne	81%	ni	80%	80%	86%	88%	100%	na
FP, prohibitions ⁶	ni	ni	ni	100%	67%	90%	93%	96%	na
FP, alternative-specific effects ⁷	100%	ni	100%	na	100%	85%	na	ni	na
FP, cross-effects ⁸	ne	ne	74%	ne	ne	100%	ne	ni	na

- ¹ 3⁴, 18 choice sets in fixed designs, 18 choice sets per respondent, triples
- ² 5 x 4 x 3 x 2, 25 choice sets in fixed designs, 25 choice sets per respondent, triples
- ³ 3¹⁰, 60 choice sets in fixed designs, 10 choice sets per respondent, triples
- ⁴ 3⁴, 8 interactions, 81 choice sets in fixed designs, 27 choice sets per respondent, triples
- ⁵ 3⁴, 16 interactions, 81 choice sets in fixed designs, 27 choice sets per respondent, triples
- ⁶ 3⁴, 4 prohibitions, 18 choice sets in fixed designs, 18 choice sets per respondent, triples
- ⁷ 3⁴ common effects, A: 3³ B: 3² C: 3¹ alternative specific effects, 27 choice sets in fixed designs, 27 choice sets per respondent, triples
- ⁸ 3⁴, 36 cross effects, 27 choice sets in fixed designs, 27 choice sets per respondent, triples
- ne = effects not estimable
na = design strategy not available or not applicable
ni = not investigated

Efficiency is a measure of the information content a design can capture. Efficiencies are typically stated in relative terms, as in “design A is 80% as efficient as design B.” In practical terms this means you will need 25% more (the reciprocal of 80%) design A observations (respondents, choice sets per respondent or a combination of both) to get the same standard errors and significances as with the more efficient design B. We used a measure of efficiency called D-Efficiency (Kuhfeld *et al.* 1994). The procedure for computing the precision of a design and D-efficiency using CBC and SPSS software is explained in the appendix.

The relative efficiencies of the different design strategies appear in Exhibit 2. We have scaled the results for each row relative to the best design investigated being 100% efficient. Many of the designs were inestimable because they were inappropriate for the kind of effects included in the design and these are coded ne (not estimable). Other designs simply cannot be constructed by the method shown in the column – these are na. Finally, some results we did not investigate (ni) for reasons noted below.

For generic main effects estimation, minimal overlap within choice sets is ideal, whether or not attributes are symmetric or asymmetric:

Method	Symmetric Design	Asymmetric Design
	Efficiency	Efficiency
Shifted fractional factorial	1.00	0.99
Computer optimization	1.00	0.98
CBC complete enumeration	1.00	1.00
CBC shortcut	1.00	1.00
CBC random	0.68	0.76
CBC balanced overlap	0.86	0.92

When designs are symmetric, the orthogonal catalog-based designs with a shifting strategy (where each level is available once per choice set) produce optimal designs with respect to main effects, as do CBC Complete Enumeration, CBC Shortcut and SAS and CVA optimization. Other fractional factorial methods we did not investigate because they would be inferior in principle to shifting. With asymmetric designs, however, CBC's strategies (Complete Enumeration and Shortcut) can be slightly more efficient than fractional factorial shifting. This finding was shown even more convincingly than our particular example in a 1999 Sawtooth Software Conference paper (Mulhern 1999).

For partial profile designs, only four methods are available and one, SAS optimization, we found difficult to program. Of the three remaining, CBC Shortcut performed best, followed closely by the recipe approach and distantly by CBC Random. This confirms earlier findings (Chrzan 1998).

<u>Method</u>	<u>Partial Profile Efficiency</u>
Recipe	0.95
CBC shortcut	1.00
CBC random	0.66

For situations requiring interactions, computer optimization via SAS produces the most efficient designs:

<u>Method</u>	<u>Few Interactions Efficiency</u>	<u>Many Interactions Efficiency</u>
Mix & match fractional factorial	0.90	0.81
Computer optimization	1.00	1.00
CBC complete enumeration	0.94	0.80
CBC shortcut	0.94	0.80
CBC random	0.90	0.86
CBC balanced overlap	0.97	0.88

Balanced Overlap is the best of the CBC strategies in both of these cases. Interactions can be inestimable when shifting fractional factorial designs and the L^{MN} approach should be about as efficient as the fractional factorial mix and match approach. A practical advantage of using CBC for interactions designs is that the analyst need not accurately predict which interactions will be needed, as in SAS or in the fractional factorial designs.

The most efficient designs for alternative-specific attributes are CBC Shortcut and a fractional factorial approach that uses shifting for shared attributes and L^{MN} for alternative-specific attributes:

<u>Method</u>	<u>Efficiency</u>
Fractional factorial mix & match/ L^{MN}	1.00
CBC shortcut	1.00
CBC random	0.85

Computer optimization using SAS is possible, but we found it difficult to program.

Especially interesting was how poorly the fractional factorial L^{MN} design fared relative to CBC random for estimating a cross-effects design:

<u>Method</u>	<u>Efficiency</u>
Fractional factorial L^{MN}	.74
CBC random	1.00

It is worth noting that the cross-effect design had only 27 total choice sets. Assigning respondents randomly to designs selected in a random manner with replacement results in a very large pool of different profiles and ways those profiles can be assembled in sets. In the limit, it is a full factorial design, both with respect to profiles and the ways they can be combined (without duplication) into sets. When sample size is sufficient (our example used 300 respondents), the naïve way of composing designs in this situation wins out, which may come as a surprise to those who regularly design studies specifically to model cross effects. Again, optimization with SAS is possible, but requires a steep learning curve.

Interesting, too, was how well CBC’s random designs fared almost across the board – for all but the “many” interactions designs, one or more of the four CBC strategies is either optimal or near optimal.

If the researcher wants to prohibit combinations of levels from appearing together within profiles, it is very difficult to do with catalog designs. One simple but arbitrary approach has been to discard or alter choice sets that violate prohibitions. Randomized designs can do this automatically and in a more intelligent way, and as long as the prohibitions are modest, the resulting design is often quite good. In our study CBC Complete Enumeration and computer optimization gave the most efficient prohibitions design

<u>Method</u>	<u>Efficiency</u>
Computer optimization	0.96
CBC complete enumeration	1.00
CBC shortcut	0.67
CBC random	0.90
CBC balanced overlap	0.93

However, we’ve seen other cases in which the Shortcut strategy performed better than Complete Enumeration for CBC, so we caution the reader that these findings may not generalize to all prohibitions designs.

Interestingly, some prohibitions can actually improve design efficiency, as we will now demonstrate.

LEVEL PROHIBITIONS AND DESIGN EFFICIENCY

Sometimes the analyst or the client wishes to prohibit some attribute levels from combining with others when constructing product alternatives. Prohibiting certain attribute combinations (“prohibitions”) leads to level imbalance and dependencies in the design, which popular wisdom holds should decrease design efficiency.

For example, consider a four-attribute choice study on personal computers, each with three levels (3^4 design). Further assume that we prohibit certain combinations between two attributes: Processor Speed and RAM. Each attribute has three levels, and we can characterize a particular pattern of level prohibitions between Processor Speed and RAM using the following two-way frequency grid:

	32 Meg RAM	64 Meg RAM	128 Meg RAM
200 MHZ			
300 MHZ			
400 MHZ	X	X	

In this example, of the nine possible combinations of Processor Speed and RAM, two (the cells containing an “X”) are prohibited. Three-hundred respondents are simulated assuming part worths of 0. Error with a standard deviation of unity is added to the utility of alternatives (3 per task for 18 tasks) prior to simulating choices. The design efficiencies reported below are with respect to main effects only and are indexed with respect to the orthogonal design with no prohibitions:

Complete Enumeration	64.50%
Shortcut	43.10%
Random	57.81%
Balanced Overlap	59.81%
Optimized Search	61.82%

Note that we haven’t included an efficiency figure for orthogonal catalog plans. For main effect estimation, orthogonal designs often are not possible in the case of prohibitions. In practice, using optimized search routines is usually the more feasible approach.

We see from this table that the best design efficiency (Complete Enumeration) is only 64.50% as efficient as the design without prohibitions. Prohibitions in this example have led to a 35.5% decrease in efficiency.

We caution about drawing detailed conclusions from this example, as the pattern and severity of the prohibitions chosen will dramatically alter the results. However, the main points to be made are:

- Prohibitions can have a negative effect upon design efficiency. (In some cases, severe prohibitions can result in inability to measure even main effects.)
- Some design strategies in CBC are better able to handle particular patterns of prohibitions than others. (We suggest testing each strategy through design simulations.)
- Computer search routines can accommodate prohibitions. Orthogonal plans are much more difficult to manage for prohibitions.

Now that we have provided what at the surface may seem to be a convincing argument that prohibitions are damaging, we’ll demonstrate that they are not always detrimental. In fact, prohibitions in some situations can actually *improve* design efficiency. The prior example assumed no particular pattern of utilities. Under that assumption, prohibitions are by definition harmful to design efficiency. But in real-world examples, respondents have preferences.

At this point, we should mention another factor that impacts design efficiency: utility balance. Utility balance characterizes the degree to which alternatives in a choice set are similar in preference. Severe imbalance leads to obvious choices that are less valuable for refining utility estimates. Huber and Zwerina (1996) showed that by customizing the designs for each respondent to eliminate choice tasks that had a high degree of imbalance, they were able to generate designs that were about 10-50% more efficient than an unconstrained approach.

Let’s again consider the previous example with Processor Speed and RAM. Lets assume that respondents have the following part worth utilities for these levels:

200 MHZ	-1.0	32 Meg RAM	-1.0
400 MHZ	0.0	64 Meg RAM	0.0
500 MHZ	1.0	128 Meg RAM	1.0

The combinations of levels most likely to lead to utility imbalance are 200 MHz with 32 Meg RAM (-1.0 + -1.0 = -2.0) and 500 MHz with 128 Meg RAM (1.0 + 1.0 = 2.0). If we prohibit those combinations, the frequency grid (with utilities in parentheses) would look like:

	32Meg RAM (-1.0)	64 Meg RAM (0.0)	128 Meg RAM (+1.0)
200 MHz (-1.0)	X (-2.0)	(-1.0)	(0.0)
300 MHz (0.0)	(-1.0)	(0.0)	(1.0)
400 MHz (+1.0)	(0.0)	(1.0)	X (2.0)

If we assume no pattern of preferences (part worths of zero for all levels), such a prohibition would lead to a 13% decrease in design efficiency with respect to main-effects estimation, relative to the orthogonal design with no prohibitions. But, if we assume part worth utilities of 1, 0, -1, the pattern of prohibitions above leads to a 22% *gain* in efficiency relative to the orthogonal design with no prohibitions. Note that this strategy (prohibiting certain combinations for all respondents) works well if the attributes have a rational *a priori* preference order, such as is the case for Processor Speed and RAM. Otherwise, a more complex, customized design strategy might be developed for each respondent, as illustrated by Huber and Zwerina.

Often, prohibitions are dictated by the client. With respect to Processor Speed and RAM, it is more likely that the client would state that it is highly unlikely that a 200 MHz processor would be offered with 128 Meg RAM, or that a 400 MHz processor would be offered with 32 Meg RAM. Let's examine those prohibitions:

	32Meg RAM (-1.0)	64 Meg RAM (0.0)	128 Meg RAM (+1.0)
200 MHz (-1.0)	(-2.0)	(-1.0)	X (0.0)
300 MHz (0.0)	(-1.0)	(0.0)	(1.0)
400 MHz (+1.0)	X (0.0)	(1.0)	(2.0)

Note that these prohibitions have discarded the combinations with the best utility balance and retained those combinations leading to the least utility balance. The net loss in design efficiency for this combination of prohibitions relative to the orthogonal design with no prohibitions is -34%.

The main points to be made are:

- For attributes with *a priori* preference order, prohibitions that lead to utility balance can enhance the efficiency of main-effect estimation.
- The prohibitions that clients often suggest (to make product alternatives more realistic) can be very detrimental to design efficiency.

We should note that the utility-balancing strategies above for prohibitions probably should not be implemented for price attributes. A conditional pricing strategy can lead to improved utility balance without specifying any prohibitions. The equivalent of having alternative-specific prices, conditional pricing, in CBC involves the use of a “look-up” table. Price levels are defined in terms of percentage deviations from an average price. If a premium product alternative is displayed in a task, the look-up function references a correspondingly higher price range relative to average or discount product alternatives. We won’t take time in this paper to elaborate on this technique, as the details are available in Sawtooth Software’s CBC manual.

CONCLUSION

There are several different approaches to designing choice-based conjoint experiments and several kinds of effects one might want to model and quantify in such experiments. The approaches differ in terms of which effects they can capture and in how efficiently they do so. No one design approach is clearly superior in all circumstances, but the capabilities comparison and the efficiency comparisons give the practitioner a good idea of when to use which type of design.

Researchers with good data processing skills and access to software such as CBC and SPSS can simulate respondent data and compute design efficiency prior to actual data collection. We recommend that reasonable *a priori* utilities be used when simulating respondent answers, and that a variety of design strategies be tested. The simulation results we report here can serve as a guide for choosing candidate design strategies.

APPENDIX

Computing D-Efficiency using CBC and SPSS™ Software

1. Compute a set of logit utilities using CBC software. Under the advanced settings, make sure to specify that you want the report to include the covariance matrix.
2. Use SPSS software to compute the relative precision of the design. An example of SPSS matrix command language to do this follows, for a small covariance matrix for 4 estimated parameters. Paste the covariance matrix from the logit report into the syntax between the brackets, and add the appropriate commas and semicolons.

```
MATRIX.  
COMPUTE covm={  
  0.000246914 , -0.000123457 , -0.000000000 , -0.000000000 ;  
 -0.000123457 , 0.000246914 , -0.000000000 , -0.000000000 ;  
 -0.000000000 , -0.000000000 , 0.000246914 , -0.000123457 ;  
 -0.000000000 , -0.000000000 , -0.000123457 , 0.000246914  
}.  
COMPUTE fpeff=DET(covm).  
COMPUTE defffic=fpeff**(-1/4).  
PRINT defffic.  
END MATRIX.
```

Note that this procedure reads the covariance matrix into a matrix variable called “covm”. The determinant of that matrix is saved to a variable called “fpeff.” The precision of the design is computed as fpeff raised to the -1/4 power (the negative of the reciprocal of the number of rows in the covariance matrix). In another example with 24 estimated parameters, the inverse of the covariance matrix should be raised to the -1/24 power. The precision is printed.

The resulting output is as follows:

```
Run MATRIX procedure:  
  
DEFFIC  
4676.529230  
  
----- END MATRIX -----
```

This needs to be done for two designs, a test design and a reference design. The ratio of the precision of the test design to that of the reference design is the relative D-efficiency of the test design (Bunch *et al.* 1994).

REFERENCES

- Addelman, Sidney (1962) "Orthogonal Main Effects Plans for Asymmetrical Factorial Experiments," *Technometrics* 4, 21-46.
- Anderson, Donald A. and James B. Wiley (1992) "Efficient Choice Set Designs for Estimating Availability Cross-Effect Designs," *Marketing Letters* 3, 357-70.
- Bunch, David S., Jordan J. Louviere and Don Anderson (1994) "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes." Working paper UCD-GSM-WP# 01-94. Graduate School of Management, University of California, Davis.
- Chrzan, Keith and Terry Elrod (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes," *1995 Advanced Research Techniques Conference Proceedings*. Chicago: American Marketing Association (in press).
- Chrzan, Keith (1998) "Design Efficiency of Partial Profile Choice Experiments," paper presented at the INFORMS Marketing Science Conference, Paris.
- Chrzan Keith and Michael Patterson (1999) "Full Versus Partial Profile Choice Experiments: Aggregate and Disaggregate Comparisons," *Sawtooth Software Conference Proceedings*, 235-48.
- Huber, Joel and Klaus B. Zwerina (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research* 33 (August), 307-17.
- Kuhfeld, Warren F. (2000) *Marketing Research Methods in the SAS System, Version 8 Edition*, SAS Institute.
- Kuhfeld, Warren, Randal D. Tobias and Mark Garratt (1995) "Efficient Experimental Designs with Marketing Research Applications," *Journal of Marketing Research* 31 (November), 545-57.
- Kuhfeld, Warren, (1997) "Efficient Experimental Designs Using Computerized Searches," *Sawtooth Software Conference Proceedings*, 71-86.
- Lazari, Andreas G. and Donald A. Anderson (1994) "Designs of Discrete Choice Set Experiments for Estimating Both Attribute and Availability Cross Effects," *Journal of Marketing Research* 31, 375-83.
- Louviere, Jordan J. (1988) "Analyzing Decision Making: Metric Conjoint Analysis" Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-67. Beverly Hills: Sage.
- Louviere, Jordan J. and George Woodworth (1983) "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20 (November) pp. 350-67.
- Mulhern, Mike (1999) "Assessing the Relative Efficiency of Fixed and Randomized Experimental Designs," *Sawtooth Software Conference Proceedings*, 225-32.
- Sawtooth Software, Inc. (1999), *CBC User Manual*, Sequim: Sawtooth Software.

Sawtooth Software, Inc. (1999), *The CBC/HB Module*, Sequim: Sawtooth Software.

SPSS (1997) *Trial Run*. Chicago: SPSS.

Steinberg, Dan and Phillip Colla (1998) *LOGIT: A Supplementary Module by Salford Systems*, San Diego: Salford Systems.

CUSTOMIZED CHOICE DESIGNS: INCORPORATING PRIOR KNOWLEDGE AND UTILITY BALANCE IN CHOICE EXPERIMENTS

Jon Pinnell

MarketVision Research, Inc.

ABSTRACT

Questions of design efficiency have long been a topic of interest to conjoint and choice researchers. These questions have focused both on tasks that researchers can construct as well as those that respondents are able to comprehend and reliably answer. With the recent gains in accessibility of Bayesian Methods, including commercial software recently introduced by Sawtooth Software, many of the same questions are being addressed in a new light. In this paper, we explore the role of utility balance in constructing choice tasks. We conclude that utility balance has a moderate positive impact, but that impact is frequently dwarfed by the improvements that Hierarchical Bayes (HB) can offer. As expected, the improvement is more striking in the face of greater respondent heterogeneity. We also present one additional finding as a cautionary tale when applying HB.

The question of design efficiency has recently received substantial attention. Researchers frequently are forced to strike a balance between design efficiency and respondents' ability to reliably answer particular choice tasks. Recent topics of design considerations have included: the benefit of asking second choices, the number of alternatives per task, the number of attributes per task, the number of tasks to ask, and how difficult those tasks should be. With the recent gain in popularity of Bayesian Methods, these design issues are likely to increase in importance and be evaluated in a potentially different light.

UTILITY BALANCE

Utility balance is one approach to making each choice a respondent makes more informative. The premise behind utility balance is that constructing choice tasks that equalize the choice probabilities of each alternative will eliminate dominated concepts. That is, we want to avoid tasks where we contrast 'products' that are good on every attribute to 'products' that are poor on every attribute. By eliminating these dominated concepts, every choice a respondent makes provides a greater insight into his or her utility structure. In essence, the effect of utility balance is to make the choices maximally difficult. Provided we, as researchers, can develop utility balanced tasks, the next obvious question is: can respondents deal with the added difficulty of the task?

The premise of utility balance is neither new nor unique to choice modeling. In fact, utility balance is included as a component of ACA (Johnson, 1987), in which pairs are constructed so that a respondent will be as nearly indifferent between them as possible. Huber and Hansen (1986) report that ACA produces better results when ACA presents "difficult" paired comparisons as opposed to "easy" (or dominated) pairs. Utility balance is frequently at-odds with other design criteria – orthogonality, level balance, and minimal overlap.

The most thorough discussion of utility balance as it relates to discrete choice designs is found in Huber and Zwerina (1995). The authors show that for fixed-design choice tasks, utility-balanced tasks can provide the same level of error around parameters with 10 to 50 percent fewer respondents. Their approach requires prior knowledge of the utilities. It has been suggested that even quite poorly specified utilities are better than assuming a null of all $\beta_s = 0$. Note that this assumption is different than saying no priors. In fact, specifying all $\beta_s = 0$ is probably a strong erroneous prior.

In addition to specifying priors, another difficulty is defining utility balance. It is common to specify utility balance as requiring the choice probabilities of all alternatives to be equal. This condition is illustrated in the following example:

Illustrative Examples of Utility Balance

		Concept			
	1	2	3	4	
Att1	25	5	10	15	
Att2	10	15	5	25	
Att3	5	10	20	15	
Att4	15	25	15	0	
Total	55	55	55	55	=220
Choice Prob.	.25	.25	.25	.25	

This hypothetical choice task clearly demonstrates utility balance. Each alternative is equally liked, that is, the maximum choice probability is near the inverse of the number of alternatives (1/4). We also believe that there is another way to specify utility balance. Imagine that the first two concepts each were improved by an equal amount and the last two alternatives were each diminished by some amount. As long as the first two alternatives were increased by an equal amount, they each would have equal expected choice probabilities. In this way, we have specified an alternative utility balance – the variance or difference of the two highest choice probabilities. In the previous example, the variance is small. The variance is small in the following example as well, but the maximum choice probability is quite different than the inverse of the number of alternatives.

Illustrative Examples of Utility Balance

		Concept			
	1	2	3	4	
A1	25	35	10	15	
A2	10	15	5	0	
A3	20	10	0	15	
A4	15	20	15	0	
Total	70	70	30	30	=200
Choice Prob.	.35	.35	.25	.25	

A case could be made that either of these conditions creates a more difficult task and, therefore, a more informative task. As both conditions fail to be met, the incremental value of the choice task is decreased, as illustrated in the following example.

Illustrative Examples of Utility Balance

		Concept			
	1	2	3	4	
A1	20	5	15	25	
A2	10	15	10	0	
A3	20	10	0	15	
A4	15	10	15	0	
Total	65	40	40	40	=185
Choice Prob.	.35	.22	.22	.22	

Throughout the paper, we define Utility Balance (UB) based on the difference in the choice probabilities of the two alternatives with the highest choice probabilities.

First, we wish to investigate the merits of Utility Balance without regard to the complexity of generating the prior utility estimates or of producing an appropriate design. As indicated above, we have defined utility balance based on the difference between the two most likely choices.

Empirical Data

Several datasets are explored in this paper. The first includes 650 personal computer purchase influencers. The data are explained in detail in Pinnell (1994). All respondents completed a ratings-based conjoint (ACA), and half followed that exercise with choice-based tasks. The eight choices included three alternatives and six attributes. In addition, complete rankings of three concepts in each of four holdout tasks were gathered before the ratings-based task and also after the choice-based task.

The second dataset, also experimental in nature, includes choices made by roughly 100 respondents who first completed twelve choice tasks made up of seven alternatives and then eight choice tasks of pairs. Seven attributes were included in this research. This dataset is fully described in Pinnell and Englert (1997).

Three additional commercial datasets also are included.

Findings

Among the respondents from the first dataset who received the choice section, we conducted a within-subject analysis controlling for utility balance. We divided each person's choices into those with the most and least utility balance based on aggregate logit utilities.

The eight tasks utilized randomized choice tasks, with no attempt to create utility balanced tasks. However, the following table shows that the random construction of tasks successfully produced differing levels of utility balance in various tasks.

Choice Probabilities
Task Sorted by Level of Utility Balance

Task Avg	Difference of Two Highest Choice Probabilities
1	0.08
2	0.17
3	0.27
4	0.36
5	0.47
6	0.57
7	0.70
8	0.82

The first three tasks were used for the utility-balance scenario. Tasks 5-7 were used for the unbalanced scenario. Task eight was used as a holdout task. Task four was excluded to equalize the number of observations for each set of utilities developed. Task eight, the most unbalanced, was retained as the hold-out in that it should be the most like tasks 5-7, the unbalanced scenario, making the test as conservative as possible.

Separate models were estimated for each set of tasks, pooling across respondents. By balancing within each respondent this way, the analysis can isolate the relative benefit from utility balance. Researchers sometimes express a concern that utility balance makes tasks too difficult, which causes respondents to take much longer to make their choices and more frequently ‘opt out’ of difficult tasks by using the ‘none’ alternative. Median time per task and none usage are shown in the following table.

Potential Negative Impacts of Utility Balance

	More UB	Less UB	Ratio
Median time per task	15 seconds	13 seconds	-13%
None usage	14.3%	9.4%	-52%

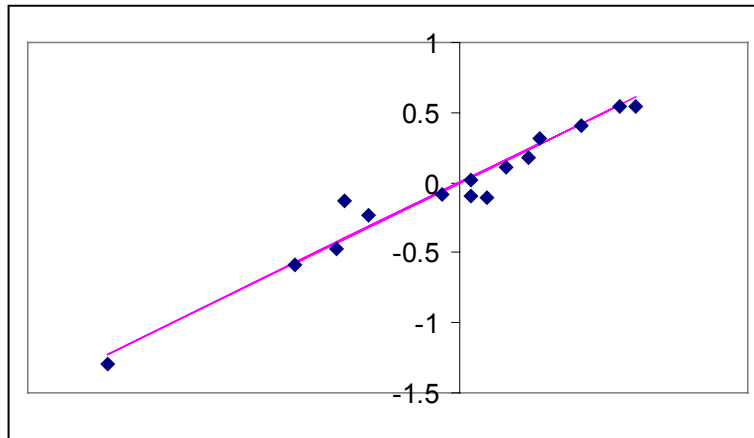
It does appear that respondents take longer to answer the utility-balanced tasks, but only marginally so. There appears to be an increase in the usage of a default alternative with the utility-balanced task. Two points must be noted. First, this dataset shows an increase in default usage (Huber and Pinnell, 1994a) but this finding was not replicated in other datasets (Huber and Pinnell, 1994b). Second, and more powerfully, Johnson and Orme (1996), after a thorough review of 15 datasets “lay to rest the conjecture that ‘None’ responses are often used as a way to avoid difficult choices.”

From this one dataset, we also see an improvement in a number of summary statistics related to the model fit in parameters. The following table shows the average of the standard errors and the average of the absolute value of the parameter t-ratios from the two models, controlling for the amount of utility balance.

Potential Positive Impacts of Utility Balance

	More UB	Less UB	Ratio
Average Standard Error	0.0699	0.0772	+10%
Average ABS(t-ratio)	4.895	4.329	+13%

The two sets of parameters are virtually indistinguishable, as shown in the following plot. The line shown on the plot represents $y=x$.



Based on the congruence, no formal tests of scale were conducted. The real question, though, is: does one provide better predictive ability than the other does? As indicated above, one task was held out for the purpose of evaluating hits. The hits under the two utility-balance scenarios are shown in the following table.

Predictive Validity Proportion Correctly Predicted By Amount of Utility Balance Experimental Study 1

With More Utility Balance	51.84%
With Less Utility Balance	50.61%
Difference	0.0123
Std. Error	0.0087
t-ratio	1.416

We see that the aggregate logit parameters estimated from the utility-balance scenario predict slightly better than those from the non-utility balance scenario do, but the difference is not significant.

A similar exercise was completed with the second experimental dataset discussed above. In these tasks, respondents were presented with choices made up of seven alternatives. The tasks were constructed using randomized choices with no attempt to create utility-balanced tasks. In total, respondents evaluated 12 tasks, with the six exhibiting the most utility balance being

separated from the six showing the least utility balance. Each set of parameters was used to predict choice from a set of pair-wise holdout choices that also were included for each respondent.

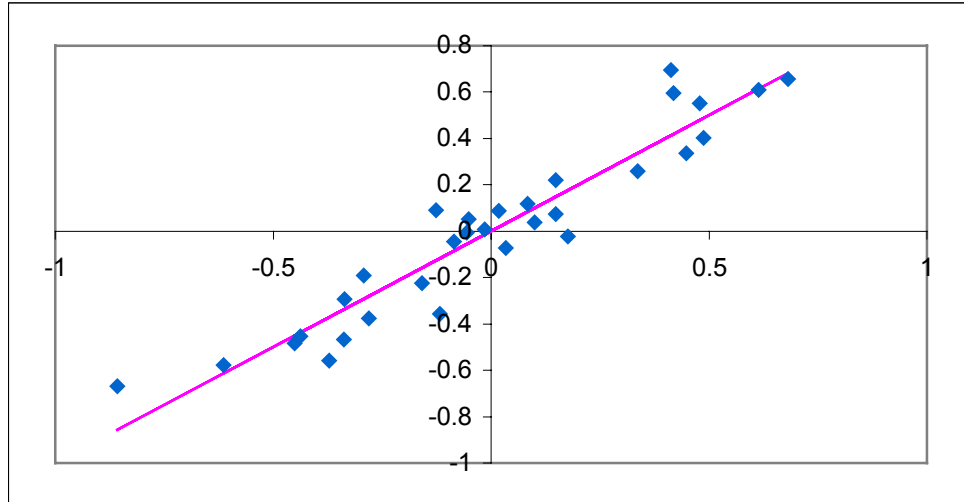
The findings are shown in the following table.

**Predictive Validity
Proportion Correctly Predicted
By Amount of Utility Balance
Experimental Study 2**

With More Utility Balance	61.16%
With Less Utility Balance	59.19%
Difference	0.0197
Std. Error	0.0092
t-ratio	2.142

Once again, we see that the aggregate logit parameters estimated from the utility balance scenario predict slightly better than those from the non-utility balance scenario do. In this case, the difference is significant with a t-ratio of 2.14.

The two sets of utilities are shown in the following plot. As in the previous plot, the line plotted represents $y=x$.



We completed a similar task two additional times with commercial studies.

In the first commercial study¹, respondents were presented with choice tasks made up of nine alternatives. Three of the choice tasks were held out for this analysis: the fourth, fifth, and sixth (based on the order presented to the respondent). The remaining six tasks were divided into those with the most and least utility balance. An independent set of logit utilities was developed from

¹ Due to anticipated heterogeneity, we had planned to use the choice tasks as a re-weighting mechanism only.

each set of tasks. Using each set of utilities, the three holdouts were predicted and the proportion of correctly predicted choices calculated. The results are shown in the following table:

**Predictive Validity
Proportion Correctly Predicted
By Amount of Utility Balance
Commercial Study 1**

With More Utility Balance	15.56%
With Less Utility Balance	13.90%
Difference	0.0166
Std. Error	0.0059
t-ratio	2.793

As expected (given the specific circumstances of the study), the hit rates are low, but this table replicates the finding that utility balance produces better predictions. The difference between the two hit rates, 0.017, is significant with a t-ratio of 2.8.

The same exercise was completed with a second commercial study, with the following results:

**Predictive Validity
Proportion Correctly Predicted
By Amount of Utility Balance
Commercial Study 2**

With More Utility Balance	75.29%
With Less Utility Balance	72.75%
Difference	0.0254
Std. Error	0.0132
t-ratio	1.921

Once again, we see that the aggregate logit parameters estimated from the utility balance scenario predict slightly better than those from the non-utility balance scenario do. This time the difference is not significant at the 95% level.

While only two of our four tests produced significant findings at 95%, we are impressed by the similarity of the improvements with utility balance. The following table summarizes the findings.

Study	Number of Attributes	Number of Parameters	Number of Alternatives	Hit Rate Improvement	Standard Error of Improvement	t-ratio
1 Experimental 1	6	15	3	0.0123	0.0087	1.42
2 Experimental 2	7	25	7	0.0197	0.0092	2.14
3 Commercial 1	4	20	9	0.0166	0.0059	2.79
4 Commercial 2	5	19	4	0.0254	0.0132	1.92

These findings are striking when evaluated by the number of alternatives per task, as shown in the following table:

Study	Number of Alternatives	t-ratio
3 Commercial 1	9	2.79
2 Experimental 2	7	2.14
4 Commercial 2	4	1.92
1 Experimental 1	3	1.42

The previous table shows a relationship between the number of alternatives presented in choice tasks and the benefits of utility balance. It has been shown that increasing the number of alternatives per task increases the efficiency² and that respondents can deal with the added complexity (Pinnell and Englert, 1997). A second benefit of including more alternatives per choice set (with randomized designs) is that utility balance is more likely to occur. This is shown in the following table of expected differences in choice probabilities between the concept with the highest choice probability and the concept with the next highest choice probability. This table was constructed based on simulating randomized choice tasks comprised of hypothetical utilities (normally distributed). Ten thousand simulations were conducted for each cell, and the average of each simulation is shown below.

Naturally Occurring Utility Balance

Average Difference of Two Highest Choice Probabilities
From Randomized Choice Designs

Number of Alternatives	Number of Attributes				
	4	5	6	7	8
2	.66	.69	.71	.73	.74
3	.54	.58	.60	.63	.65
4	.48	.51	.54	.57	.59
5	.44	.47	.51	.53	.55
6	.41	.44	.47	.50	.52
7	.38	.42	.45	.48	.50
8	.36	.40	.43	.46	.49
9	.35	.39	.42	.44	.47

The table above shows that adding *alternatives* is more likely to make the choice more difficult for respondents via utility balance. However, there is a negative impact on expected utility balance when *attributes* are added.

² Efficiency measured through utility neutral D-efficiency between alternatives with seven alternatives was more than twice that with two alternatives.

These analyses have relied on aggregate logit utilities to distinguish between the tasks that have the most and least utility-balance. When implementing utility balance, the researcher must decide between using aggregate or individual-level utilities to determine balance. Either case can produce substantial difficulties.

With aggregate utilities, it is normally necessary to conduct a small pre-test to develop the utilities that are then applied to create utility balance for the remainder of the study. This ignores any potential heterogeneity in the population of interest. The difficulty with heterogeneity is that priors are typically required for each person separately and then a customized fixed design is used for each person. This ordinarily requires two separate surveys and substantial design time in between.

However, an analysis that relies on aggregate logit utilities would appear a non sequitur to a design that takes account of individual-level heterogeneity. That is, why would one want to design an experimental treatment based on individual level differences, only to ignore them in the analysis. As such, individual-level UB should only make sense with a disaggregate analysis, such as Hierarchical Bayes.

DYNAMIC UTILITY BALANCE

In the past, we have implemented individual-level utility-balance in choice tasks when planning disaggregate analysis or, more frequently, for reweighting purposes.

However, we have implemented individual level utility balance using only one interview. For each respondent, we collect self-explication utilities, much like the priors section in ACA. Then, for each choice task that we wish to ask, we generate four randomly constructed choice sets and choose the one that best meets the UB criterion for the given respondent.

In one recent commercial study, we implemented such a design for half of the respondents and used purely random tasks for the other half. Using this design, we were able to reduce both the average maximum choice probability and the average difference between the two most preferred alternatives, as shown in the following table.

Reduction in Dominated Concepts With Dynamic Utility Balance

	Utility Balance Treatment	
	Yes	No
Average Difference of Top Two Alternatives	.33	.55
Average Maximum Choice Probability	.61	.74

Each choice task had three alternatives and an average³ of five attributes. Given this, and the table shown above based on simulation results, we would have expected a difference in choice probabilities between the two most preferred concepts to be 0.58. Our actual value from the control group of respondents (receiving no experimental treatment) was 0.55. With our

³ The tasks were partial-profile tasks and the number of attributes shown in each task was determined based on the amount of text (number of lines) required for each attribute. As such, it varied from task to task.

treatment, we were able to trim that average difference to 0.33. This represents a reduction of 40 percent.

We conducted another set of simulations to calculate the expected maximum choice probabilities. The results of those simulations are shown in the following table.

Naturally Occurring Utility Balance
 Expected Maximum Choice Probabilities
 From Randomized Choice Designs

Number of Alternatives	Number of Attributes				
	4	5	6	7	8
2	.83	.84	.86	.87	.87
3	.74	.77	.78	.79	.80
4	.68	.71	.73	.75	.77
5	.65	.67	.70	.72	.73
6	.61	.65	.67	.69	.71
7	.58	.62	.64	.67	.69
8	.56	.60	.63	.65	.67
9	.55	.58	.61	.63	.66

The expected value for the maximum choice probability was 0.77 with five attributes and three concepts per task. Our actual value from the control group was 0.74, while the value from the experimental group was 0.61. This represents a 31 percent reduction. It is important to recall that this reduction is based on a target maximum choice probability of 0.33 rather than 0.00, as illustrated below.

	Control	Experimental	Improvement
Observed	.74	.61	
Optimal	.33	.33	
Difference	.41	.28	.13 (.41 - .28)
Percent Improvement			31% (.13/.41)

As indicated above, utility balance is frequently at odds with other design criteria. Therefore, there is a concern that UB would diminish the efficiency of an experimental design. To explore this potential impact, we calculated utility neutral D-efficiency of the two designs: balanced (experimental) and unbalanced (control). Using this criterion, we calculate the balanced design to be less than 2% less efficient than the control design. While we did not calculate it directly, we believe that the increase in information from utility balance more than offsets this relatively minor decrease in design efficiency.

In addition to the choice tasks, each respondent completed a set of partial profile paired comparisons. We developed two sets of aggregate utilities, separating those respondents who received the UB choice treatment from those who received the randomly generated choices. We used each set to predict respondents' hits for the hold out pairs. The findings are shown below.

**Predictive Validity
Proportion Correctly Predicted
By Amount of Utility Balance**

With More Utility Balance	0.7230
Standard error	(0.017)
With Less Utility Balance	0.7145
Standard error	(0.017)
Difference	0.0085
t-ratio	0.36

While the results are in the direction expected, the improvement is trivial.

HIERARCHICAL BAYES AND BAYESIAN METHODS

Over the last several years, a great amount of attention has been focused on disaggregating choice tasks, specifically through Markov Chain Monte Carlo methods like Hierarchical Bayes (HB).

In the following section, we compare the improvement in predictive ability by using HB. We use the two commercial studies referenced above under the UB section. These two commercial datasets were expected to exhibit substantial heterogeneity.

The findings are shown below:

Predicting Hold-out Tasks

	Aggregate Logit	Hierarchical Bayes	Improvement
Commercial Study One	75.75%	99.54%	23.79%
Commercial Study Two	24.79%	79.46%	54.67%

These improvements are striking. These two example datasets were selected in that we expected that they were the most likely to benefit from an HB analysis. Even using the one of the experimental datasets, which is expected to have less heterogeneity, we see improvements, though not nearly as substantial.

Predicting Hold-out Tasks

Aggregate Logit	60.54%
Hierarchical Bayes	62.62%
Difference	0.0208
Std. Error	0.0155
t-ratio	1.337

Given this apparent support for HB, we wanted to explore whether or not HB would help in the realm of individual-level utility balance.

UTILITY BALANCE WITH HB

Using the two experimental datasets described above, we divided each respondent's choice sets into those with the most and least utility balance based on individual utilities. We then conducted an HB analysis with each set and used those disaggregated utilities to predict hits in holdout choices. The results are shown below.

Predictive Validity First Experimental Study Proportion Correctly Predicted By Amount of Utility Balance

	Aggregate Utilities	HB Utilities
With More Utility Balance	51.84%	59.40%
With Less Utility Balance	50.61%	59.19%
Difference (Using HB Utilities)		0.0021
Std. Error		0.0174
t-ratio		0.120

Predictive Validity Second Experimental Study Proportion Correctly Predicted By Amount of Utility Balance

	Aggregate Utilities	HB Utilities
With More Utility Balance	61.16%	79.27%
With Less Utility Balance	59.19%	78.80%
Difference (Using HB Utilities)		0.0047
Std. Error		0.6610
t-ratio		0.711

In both instances, we show that using HB utilities provides better predictions than using aggregate utilities. The benefit of utility balance remains slight, whether utilities are estimated by aggregate logit or HB.

Given this directional improvement with HB, we wished to explore the potential improvements of using HB with the Dynamic Utility Balance from the commercial study discussed above.

**Predictive Validity
By Amount of Utility Balance
With Aggregate Utilities and HB Utilities**

	Aggregate Utilities		HB Utilities		Difference in Number Predicted
	Number Predicted	Hit Rate	Number Predicted	Hit Rate	
With More Utility Balance	5.7841	.7230	5.3296	.6662	-.4545
With Less Utility Balance	5.7159	.7145	5.5681	.6960	-.1478

We were surprised by these findings. The HB findings are significantly inferior to the aggregate logit utilities (the t-ratio of the difference is -3.09 with more balance and -1.10 with less balance).

It is unclear why the HB utilities under-perform relative to the aggregate logit utilities. The study was a moderately large study, with 39 levels from a total of 12 attributes. Each respondent completed 10 tasks, which might have been too few for the HB analysis to be successful given the partial profile nature of the tasks.

Other potential explanations, though still conjecture on our part, include the following:

- There wasn't enough heterogeneity among the respondents that would allow disaggregation of respondents to be beneficial.
- The holdout tasks were dominated, which would limit their resolving power.
- The process of imposing Utility Balance degrades the quality of the design such that the power of model is reduced.

For each, we provide findings to eliminate them as possible explanations.

Heterogeneity

While it might be the case that there is not enough variability among respondents to benefit from an individual level analysis, there does appear to be heterogeneity. Using only respondents' prior estimates of utilities, we identified three segments of respondents. Deterministically assigning each respondent to one of the three segments, we estimated three independent logit models. Even accounting for scale differences between the three groups (which were trivial), we conclude that the parameters are different ($p < .01$), and we see moderately improved predictions with disaggregation (t-ratio > 1).

Dominance

Even in the face of respondent heterogeneity, it might be the case that the holdout tasks were too dominated to allow what might be slightly noisier individual estimates to better predict holdouts. Recall that there were eight pairwise holdout tasks. These tasks varied in their level of dominance. The average choice proportion was roughly 70/30, but the two alternatives with the least dominated pairs had choice proportions that were nearly equal, and the most dominated pair had choice proportions of approximately 90/10. The level of domination of the pairs affects the degradation in hit rates only slightly. If we explore the two holdouts with nearly equally preferred alternatives (itself a potential sign of heterogeneity), we would draw the same basic

conclusion as we would with all eight holdouts (t-ratio of HB hits relative to aggregate logit hits approx. = -1).

Utility Balance

While the process of creating utility balance is likely to degrade the design some, as reported above, the efficiency of the utility balanced design, relative to the control (non-balanced) design, had a loss of D-efficiency of less than 2%. We believe this is unlikely to be the cause of our finding.

Overall, this finding intrigues us, mainly because we don't fully understand it. The benefits of HB, including several documented in this work, appear promising. However, this finding should reinforce that HB is not a panacea and should not be applied indiscriminately.

SUMMARY

Randomly generated choice tasks frequently produce a large number of dominating concepts. These dominating concepts are intuitively less useful than more balanced concepts would be. If they are selected, the researcher learns very little about a respondent's underlying utility structure.

The premise of utility balance is to reduce the occurrence of dominating concepts so that each choice becomes maximally informative. There is a difficulty in constructing these UB tasks and there is a concern about respondents' ability to process and reliably answer these more difficult tasks.

We show that Utility Balanced tasks can produce utilities that yield more accurate predictions of choice. The finding is stronger, apparently, when derived from choices with more alternatives.

We also explored whether HB can strengthen our UB findings. While we document substantial improvements in overall predictive ability using HB over aggregate logit utilities, HB fails to improve upon our UB findings, and in one case is actually detrimental.

The improvements of HB alone are larger than the magnitude of the UB improvement.

REFERENCES

- Huber, Joel and David Hansen (1986), "Testing the Impact of Dimensional Complexity and Affective Differences of Paired Concepts in Adaptive Conjoint Analysis." *Advances in Consumer Research*, Vol. 14, M. Wallendorf and P. Anderson, Eds. Provo, UT: Association for Consumer Research, 159-63.
- Huber, Joel and Jon Pinnell (1994a), "The Impact of Set Quality and Decision Difficulty on the Decision to Defer Purchase." Working Paper, Fuqua School of Business, Duke University.
- Huber, Joel and Jon Pinnell (1994b), "The Meaning of a 'NONE' Response in Commercial Studies Using Choice-Based Conjoint." Presented at Association for Consumer Research Conference, Boston, MA.
- Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs." *Journal of Marketing Research*.
- Johnson, Richard M. (1987), "Adaptive Conjoint Analysis." *Proceedings of the Sawtooth Software Conference*. Ketchum, ID: Sawtooth Software, 253-65.
- Johnson, Richard M. and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint?" Presentation at AMA Advanced Research Techniques Forum, Beaver Creek, CO.
- Pinnell, Jon (1994), "Multistage Conjoint Methods to Measure Price Sensitivity." Presentation at AMA Advanced Research Techniques Forum, Beaver Creek, CO.
- Pinnell, Jon and Sherry Englert (1997), "The Number of Choice Alternatives in Discrete Choice Modeling." *Proceedings of the Sawtooth Software Conference*. Sequim, WA: Sawtooth Software, 121-53.

UNDERSTANDING HB: AN INTUITIVE APPROACH

Richard M. Johnson
Sawtooth Software, Inc.

INTRODUCTION

Hierarchical Bayes Analysis (HB) is sweeping the marketing science field. It offers valuable benefits, but it's so different from conventional methods that it takes some getting used to. The purpose of this talk is to convey an intuitive understanding of what HB is, how it works, and why it's valuable.

We'll start with a simple example to show how Bayesian analysis is fundamentally different from what we usually do. Then I'll describe hierarchical models useful for estimating conjoint part worths and the algorithm used for computation. I'll finish with evidence that HB does a better job than conventional analysis, and describe some of the questions that remain to be answered.

THE BASIC IDEA OF BAYESIAN ANALYSIS

In a small air charter service there was a new pilot named Jones. Jones was an excellent pilot in every way but one – he had trouble making smooth landings. The company did customer satisfaction research, and one of the questions asked of passengers was whether the landing had been smooth. Fully 20% of Jones's landings were judged to be rough. All the other pilots had better scores, and the average percentage of rough landings for the rest of them was only 10%.

We could summarize that information in a table like this, which gives probabilities of rough or smooth landings for Jones and all other pilots combined:

**Probabilities of Rough or Smooth Landings,
Conditioned on Pilots**

	Rough	Smooth	Total
Jones	.20	.80	1.00
Other	.10	.90	1.00

One day a passenger called the president of the company and complained about a rough landing she had experienced. The president's first reaction was to think "Probably Jones again." But then he wondered if that was reasonable. What do you think?

These data do give us a probability: If Jones is the pilot, the probability of a rough landing is .2. This is the conditional probability of a rough landing, given that the pilot is Jones. But the president was considering a different probability: If it was a rough landing, what was the probability that the pilot was Jones?, which is the conditional probability of Jones, given it was a rough landing.

In conventional statistics we are accustomed to assuming a hypothesis to be true, and then asking how the data would be expected to look, given that hypothesis. An example of a

conventional research question might be, “Given that Jones is the pilot, what’s the probability of a rough landing?” And by a conventional analysis we can learn the answer, which is “Twenty Percent.”

With Bayesian analysis we can turn things around and ask the other question: “Given that the landing was rough, what’s the probability that Jones was the pilot?”

We can’t answer that question from these data, because we don’t know anything about how often Jones flies. For example, if there were a million pilots, all flying equally often, the answer to this question would obviously be different than if there were only two.

To answer the president’s question we must provide one further kind of information, which Bayesians call “Prior Probabilities.” In this case that means “Irrespective of this particular flight, what is the probability that Jones will be the pilot on a typical flight.” Suppose this company has 5 pilots, and they are all equally likely to have drawn this flight on this particular day. Then a reasonable estimate of the prior probability that Jones would have been the pilot would be one fifth, or .2.

Suppose we multiply each row of the previous table by its prior probability, multiplying Jones’s row by .2, and everyone else’s row by .8, to get the following table:

Joint Probabilities of Rough or Smooth Landings,

	Rough	Smooth	Total
Jones	.04	.16	.20
Other	.08	.72	.80
Total	.12	.88	1.00

This table gives the probabilities of all combinations of pilots and outcomes, saying that out of 100 flights with this airline, we should expect 4 rough landings by Jones. Probabilities like these are called “Joint Probabilities” because they describe combinations of both variables of interest. Notice that they sum to unity. The row sums give the overall probability of each type of pilot flying, and the column sums give the overall probability of each kind of landing.

Once we have joint probabilities, we are able to answer the president’s question. Given that the flight was rough (.12), the probability that Jones was the pilot was only $.04 / .12$, or one third. So the president was not justified in assuming that Jones had been the pilot.

Three kinds of probabilities are illustrated by this example.

Prior Probabilities are the probabilities we would assign *before we see the data*. We assign a prior probability of .2 for Jones being the pilot, because we know there are 5 pilots and any of them is equally likely to be at the controls for any particular flight.

Likelihood is the usual name for the *probability of the data, given a particular hypothesis or model*. This is the kind of probability we’re used to: the likelihood of a rough landing, given that the pilot is Jones, is .2.

Posterior Probabilities are the probabilities we would assign *after we have seen data*. Posterior probabilities are based on the priors as well as information in the data. Combining

the priors and the data, our posterior probability that the pilot of a rough landing was Jones is one third. After learning that flight had a rough landing, the probability that Jones was its pilot was updated from .2 to .333.

BAYES' RULE:

To discuss probabilities, we need some notation. For any two events, X and Y, we define:

$P(X)$ = the marginal probability of X (e.g., without respect to Y)

$P(X,Y)$ = the joint probability of X and Y (e.g. probability of both X and Y)

$P(X | Y)$ = the conditional probability of X given Y

The definition of conditional probability is

$$P(X | Y) = \frac{P(X, Y)}{P(Y)} \quad (1)$$

Starting from equation (1) we can derive Bayes' Rule by simple algebra. Multiply both sides of equation (1) by P(Y) to get

$$P(X | Y) * P(Y) = P(X, Y) \quad (2)$$

Since X and Y could be any events, we can write equation (3) just by exchanging the roles of X and Y in equation (2):

$$P(Y | X) * P(X) = P(Y, X) \quad (3)$$

Noting P(X, Y) is the same thing as P(Y, X), we can equate the left hand sides of equations (2) and (3), getting

$$P(X | Y) * P(Y) = P(Y | X) * P(X) \quad (4)$$

Finally, dividing both sides by P(Y), we get what is known as "Bayes' Rule,"

$$P(X | Y) = \frac{P(Y | X) * P(X)}{P(Y)} \quad (5)$$

Bayes' Rule gives us a way of computing the conditional probability of X given Y, if we know the conditional probability of Y given X and the two marginal probabilities. Let's apply Bayes rule to our example. Let X be the event "Jones was the pilot" and Y be the event "Rough Landing." Then Bayes' Rule says:

$$P(\text{Jones} | \text{Rough}) = \frac{P(\text{Rough} | \text{Jones}) * P(\text{Jones})}{P(\text{Rough})} = \frac{.2 * .2}{.12} = 1/3 \quad (6)$$

This is the same computation that we performed before. Bayes' rule gives us a way to answer the question posed by the president. This same relationship between probabilities underlies the entire field of Bayesian analysis, including HB.

In real-life Bayesian applications, the probability in the denominator in equation (5) is often hard to compute. Also, since it often depends on arbitrary factors like the way measurements are made and data are coded, it is seldom of much interest. For example, the president wasn't wondering about the proportion of flights that had rough landings, which can be determined easily by other means. Our question was: *Given* a rough landing, what was the probability that Jones was the pilot?

Therefore, in practical applications the denominator is often regarded as a constant, and Bayes rule is expressed as:

$$P(X | Y) \propto P(Y | X) * P(X) \quad (7)$$

where the symbol \propto means "is proportional to." The main thing to remember is the relationship indicated in equation (7), which may be stated: ***Posterior probabilities are proportional to likelihoods times priors.*** If you can remember this slogan, you are well on the way to understanding Bayesian analysis.

SOME CHARACTERISTICS OF BAYESIAN ANALYSIS

We've used this simple example to introduce Bayes' rule, and to show how to produce posterior probabilities by updating prior probabilities with likelihoods obtained from data. Bayesian analysis differs from conventional analysis in several ways:

Bayesian analysis is sometimes said to involve "subjective probabilities" because it requires specification of priors. The priors in the example were obtained by assuming each pilot was equally likely to have any flight, but such sensible priors are not always available. Fortunately, there is enough data in large-scale applications using HB that the priors have very little effect on the posterior estimates.

Bayesian analysis is sometimes said to involve "inverse probabilities." In conventional analysis, we regard parameters as fixed and the data as variable. In Bayesian analysis things are reversed. After the data are in hand we regard them as fixed, and we regard the parameters as random variables, with distributions that we try to estimate.

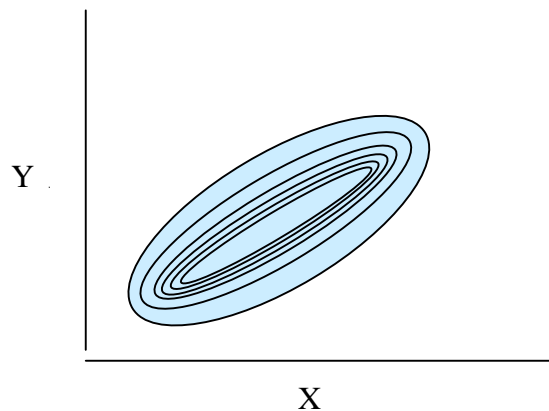
Although many Bayesian models are simple in concept, their actual estimation is often difficult, and requires computer simulations. Those simulations can take a long time – perhaps many hours rather than just a few seconds or minutes.

In exchange for these complexities, Bayesian methods offer important benefits. HB can produce better estimates of individual values. For conjoint analysis, we can get equivalent accuracy using shorter questionnaires. We can also get useful individual estimates where before we might have had to settle for aggregate estimates. And this is true not only for conjoint analysis, but also for customer satisfaction research and scanner data.

MARKOV CHAIN MONTE CARLO ANALYSIS

Bayesian analysis has benefited enormously from recent increases in computer speed. A group of methods which have been especially useful for HB estimation are known as MCMC or “Monte Carlo Markov Chain” methods. One MCMC method particularly useful for HB is called the “Gibbs Sampler.” The basic idea behind the Gibbs Sampler can be demonstrated by a small example.

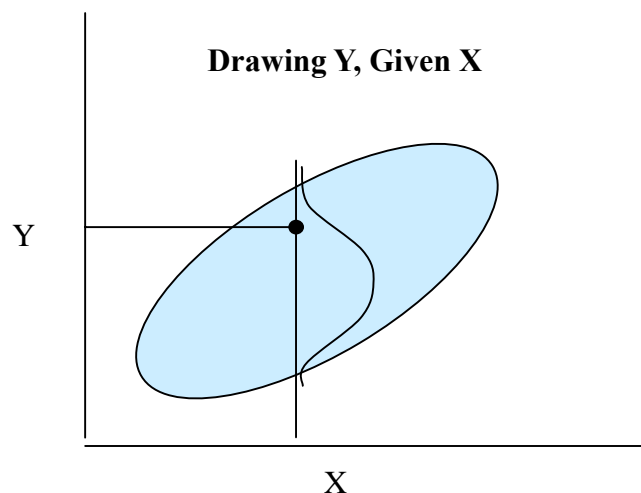
Suppose X and Y are two normally distributed variables with similar variability. The joint distribution of two variables in a sample is often shown with a contour map or a scatter-plot. If the variables are uncorrelated, the scatter-plot is approximately circular in shape. If the variables are correlated, the swarm of points is elongated.



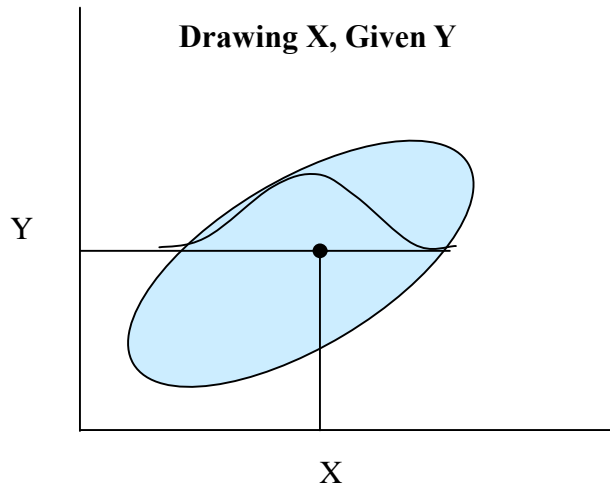
Suppose we don't know the joint distribution of X and Y , but wish we did. Suppose, however, that we know both conditional distributions: given a value for X , we know the distribution of Y conditional on that X ; and given a value for Y , we know the distribution of X conditional on that value for Y . This information permits us to simulate the joint distribution of X and Y using the following algorithm:

Use any random value of X to start.

Step (1) Given X , draw a value of Y from the distribution of Y conditional on X .



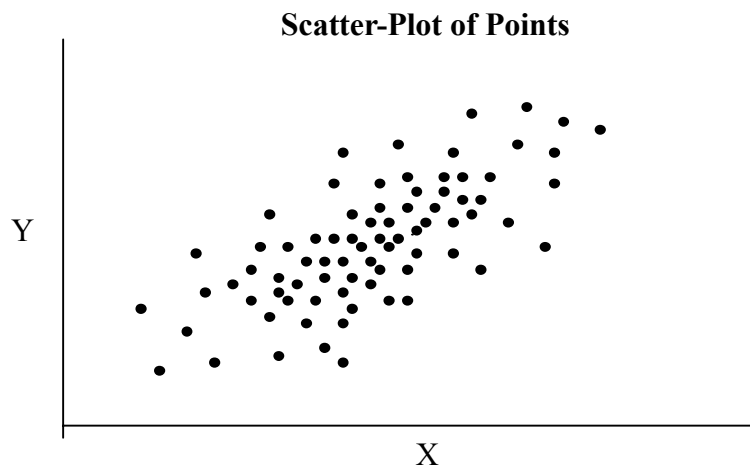
Step (2) Given Y, draw a value of X from the distribution of X conditional on Y.



Repeat steps 1 and 2 many times, labeling successive draws X_1, X_2, X_3, \dots

And Y_1, Y_2, Y_3, \dots . Plot each pair of points X_i, Y_i .

As the number of steps (iterations) increases, the scatter-plot of the successive pairs of X, Y approximates the joint distribution of X and Y more and more closely.



This principle is valuable for HB, because the *joint* distribution of many variables can horrendously complicated and impossible to evaluate explicitly. But the statistical properties of normal distributions permit us to estimate *conditional* distributions much more easily. We use a similar iterative process where we repetitively select each variable and estimate all the others conditionally upon it. This process is a Markov chain because the results for each iteration depend only on the previous iteration, and are governed by a constant set of transition probabilities.

HIERARCHICAL MODELS FOR CONJOINT ANALYSIS

In this presentation we skip technical details in an attempt to give an intuitive idea of how HB works. Explanatory papers are on the sawtoothsoftware.com web site providing further details.

Suppose many survey respondents have each answered several choice questions. We want to estimate part-worths for each individual contained in a vector \mathbf{b} , the mean for the population of individuals contained in the vector \mathbf{a} , and the variances and covariances for the population contained in the matrix \mathbf{C} .

The model is called “hierarchical” because it has two levels. At the *upper level*, we assume that individuals’ vectors of part-worths are drawn from a multivariate normal distribution:

$$\text{Upper Level Model: } \mathbf{b} \sim N(\mathbf{a}, \mathbf{C})$$

At the *lower level*, we assume a logit model for each individual, where the utility of each alternative is the sum of the part-worths of its attribute levels, and the respondent’s probability of choosing each alternative is equal to its utility divided by the sum of utilities for the alternatives in that choice set.

$$\text{Lower Level Model: } \mathbf{u} = \sum \mathbf{b}_i$$

$$\mathbf{p} = \exp(\mathbf{u}) / \sum \exp(\mathbf{u}_j)$$

One starts with initial estimates for of \mathbf{a} , the \mathbf{b} ’s, and \mathbf{C} . There is great latitude in choosing these estimates. Our estimates of \mathbf{b} for each individual are the numbers of times each attribute level is in the chosen alternatives, divided by the number of times each attribute level is present in all alternatives. Our initial estimate of \mathbf{a} has all elements equal to zero, and for \mathbf{C} we set initial variances at unity and covariances at zero.

The algorithm repeats these steps in each iteration.

Step(1) Given current estimates of the \mathbf{b} ’s and \mathbf{C} , estimate the vector \mathbf{a} of means of the distribution.

Step(2) Given current estimates of the \mathbf{b} ’s, and \mathbf{a} , estimate the matrix \mathbf{C} of variances and covariances.

Step(3) Given current estimates of \mathbf{a} , and \mathbf{C} , estimate a new \mathbf{b} vector for each respondent.

The process is continued for many thousands of iterations. The iterations are divided into two groups.

The first several thousand iterations are used to achieve convergence, with successive iterations fitting the data better and better. These are called “preliminary,” “burn-in,” or “transitory” iterations.

The last several thousand iterations are saved for later analysis, to produce estimates of the \mathbf{b} ’s, \mathbf{a} , and \mathbf{C} . Unlike conventional statistical analysis, successive iterations do not converge to a single “point-estimate” for each parameter. Even after convergence, estimates from successive iterations bounce around randomly, reflecting the amount of uncertainty that exists in those estimates. Usually we want a point-estimate of part-worths for each respondent, and this is

obtained simply by averaging estimates of that individual's \mathbf{b} 's for the last several thousand iterations.

During iterations, successive values of \mathbf{a} and \mathbf{C} are estimated by straight-forward statistical procedures that we shall not consider. However, successive values of the \mathbf{b} 's are estimated by a "Metropolis-Hastings" algorithm which illustrates the Bayesian nature of the process, and which we shall now describe.

The following is done for each individual at each iteration:

Define the individual's previous estimate of \mathbf{b} as \mathbf{b}_{old} . Construct a candidate for a new estimate for that individual by adding a small random perturbation to each element of \mathbf{b}_{old} , calling the resulting vector \mathbf{b}_{new} . Using the data and the logit model, we compute the likelihood of seeing that respondent's set of choices, given each of those \mathbf{b} vectors. Each likelihood is just the product of the predicted probabilities of all choices made by that respondent, given that estimate of \mathbf{b} . We compute the ratio of those likelihoods, $\mathbf{l}_{new} / \mathbf{l}_{old}$.

Recall that the hierarchical model regards the individuals' \mathbf{b} vectors to have been drawn from a multivariate normal distribution with mean vector \mathbf{a} and covariance matrix \mathbf{C} . We can use standard statistical formulas to compute the relative probabilities that \mathbf{b}_{new} and \mathbf{b}_{old} would have been drawn from that distribution, indicated by the height of the distribution's graph at each point. We compute the ratio of those probabilities, $\mathbf{p}_{new} / \mathbf{p}_{old}$.

Finally, we compute the product of ratios,

$$\mathbf{r} = (\mathbf{l}_{new} / \mathbf{l}_{old}) * (\mathbf{p}_{new} / \mathbf{p}_{old}) = \frac{\mathbf{l}_{new} * \mathbf{p}_{new}}{\mathbf{l}_{old} * \mathbf{p}_{old}} \quad (8)$$

Recall that *posterior probabilities are proportional to likelihoods times priors*. The \mathbf{p} 's may be regarded as priors, since they represent the probabilities of drawing each vector from the population distribution. Therefore, \mathbf{r} is the ratio of posterior probabilities of \mathbf{b}_{new} and \mathbf{b}_{old} .

If \mathbf{r} is greater than unity, the new estimate has a higher posterior probability than the previous one, and we accept \mathbf{b}_{new} . If \mathbf{r} is less than unity we accept \mathbf{b}_{new} with probability equal to \mathbf{r} .

Over the first several thousands of iterations, the \mathbf{b} 's gradually converge to a set of estimates that fit the data while also conforming to a multinormal distribution.

If a respondent's choices are fitted well, his estimated \mathbf{b} depends mostly on his own data and is influenced less by the population distribution. But if his choices are poorly fitted, then his estimated \mathbf{b} depends more on the population distribution, and is influenced less by his data. In this way, HB makes use of every respondent's data in producing estimates for each individual. This "borrowing" of information is what gives HB the ability to produce reasonable estimates for each respondent even when the amount of data available for each may be inadequate for individual analysis.

TYPICAL HB RESULTS FOR CHOICE DATA

Huber, Arora & Johnson (1998) described a data set in which 352 respondents answered choice questions about TV preferences. There were 6 conjoint attributes with a total of 17 levels. Each respondent answered 18 customized choice questions with 5 alternatives, plus 9 further holdout choice questions, also with 5 alternatives.

We examine HB's ability to predict holdout choices using part-worths estimated from small numbers of choice tasks. Part-worths were estimated based on 18, 9, 6, and 4 choices per respondent. Point estimates of each individual's part-worths were obtained by averaging 100 random draws, and those estimates were used to measure hit rates for holdout choices. The random draws were also used in 100 first-choice simulations for each respondent, with predicted choices aggregated over respondents, to measure Mean Absolute Error (MAE) in predicting choice shares. Here are the resulting Hit Rate and MAE statistics for part-worths based on different numbers of choice tasks:

TV Choice Data Holdout Prediction With Subsets Of Tasks

# Tasks	Hit Rate	MAE
18	.660	3.22
9	.602	3.62
6	.556	3.51
4	.518	4.23

Hit rates are decreased by about 10% when dropping from 18 choice tasks per respondent to 9, and by about 10% again when dropping from 9 tasks to 4. Similarly, mean absolute error in predicting choice shares increases by about 10% each time the number of tasks per respondent is halved. Even with as few as four questions per respondent, hit rates are much higher than the 20% that would be expected due to chance.

TYPICAL HB RESULTS FOR ACA DATA

This data set was reported by Orme, Alpert, and Christensen (1997) in which 80 MBA students considered personal computers, using 9 attributes, each with two or three levels. Several kinds of data were collected in that study, but we now consider only ACA data plus first choices from five holdout tasks, each of which contained three concepts.

ACA provides information about each respondent's "self-explicated" part-worths, as well as answers to paired-comparison tradeoff questions. The HB user has the option of fitting just the paired comparison data, or the paired comparison data plus the self-explicated information. Also, there is an option of constraining the part-worths to obey order relations corresponding to the self-explicated information. Here are hit rates for several methods of estimating part-worths:

	MBA Data Hit Rate %
ACA Version 4 ("optimal weights")	66.25
Pairs Only With Constraints	71.50
Pairs + Self-Explicated With Constraints	70.75
Pairs + Self-Explicated, No Constraints	69.35
Pairs Only, No Constraints	69.00

It is noteworthy that all four of the HB combinations are more successful than the conventional method of estimation offered by ACA. Similar results have been seen in several other comparisons of HB with conventional ACA estimation methods.

TYPICAL HB RESULTS FOR REGRESSION DATA

In the Orme et al. study each respondent also did a full-profile card-sort in which 22 cards were sorted into four piles based on preference, and then rated using a 100 point scale. Those ratings were converted to logits, which were used as the dependent variable, both for OLS regression and also by HB. In these regressions each respondent contributed 22 observations and a total of 16 parameters were estimated for each, including an intercept. Here are hit rates for predicting holdout choices:

MBA Data

Ordinary Least Squares	72.00%
HB	73.50%

HB has a 1.5% margin of superiority. This is typical of results seen when HB has been compared to individual least squares estimation.

REMAINING QUESTIONS

Although HB users have seemed pleased with their results, there are two important problems yet to be solved. The first problem is that of enforcing **order constraints**. Conjoint analysis is usually more successful if part-worths are constrained so that values for more desirable levels are greater than values for less desirable levels. Not only is this true for attributes where everyone would agree that “more is better”, but it also appears to be true in ACA when each respondent’s part-worths are constrained to have the same order relations as his self-explicated information.

Our HB module for ACA does permit enforcing such constraints, but to accomplish this we have had to employ a procedure which is not strictly correct. We are working on this problem, and hope soon to include a statistically-correct capability of enforcing order constraints in each of our HB software products.

The second problem is that of knowing **how many iterations to specify**. After the iterations have been completed, it is not difficult to look at their history and decide whether the process appeared to converge during the “burn-in” period. But not much information has been available about how many iterations are likely to be required. To be on the safe side, we have suggested between 10 and 20 thousand preliminary iterations. But if fewer are really required, it may be possible to cut run times dramatically. The authors of the next paper have examined this question, and I look forward to hearing what they have to report.

SHOULD YOU USE HB?

There is no denying that HB takes longer than other methods of individual estimation. We have seen run times ranging from a few minutes to a few days. Here are some timing examples for HB-Reg, the package for general regression applications such as customer satisfaction or scanner data. The time required is approximately proportional to number of respondents, the

average number of answers per respondent, and (for large numbers of variables) the square of the number of variables. Using a computer with a 500 mhz Pentium III processor, we have observed these times for 20,000 iterations with 300 respondents and 50 answers per respondent:

Representative Times for Computation

Number of Variables	Time for 20,000 Iterations
10	1 hour
40	3 hours
80	14 hours

To summarize our experience with HB, part-worths estimated by HB have usually been better and almost never worse at predicting holdout choices than part-worths estimated by previously available methods.

HB also has the valuable property of being able to produce useful individual estimates even when few questions are asked of each respondent, in situations where previously the researcher had to settle for aggregate estimates.

I have not said much about customer satisfaction research, but there is also evidence that HB produces better estimates of attribute importances in the face of colinearity, such as often found in customer satisfaction data.

Given our favorable experience with HB estimation, we recommend using HB whenever the project schedule permits doing so.

REFERENCES

Huber, J., Arora, N., and Johnson, R. (1998) "Capturing Heterogeneity in Consumer Choices," *ART Forum*, American Marketing Association.

Orme, B. K., Alpert, M. I. & Christensen, E. (1997) "Assessing the Validity of Conjoint Analysis – Continued," *Sawtooth Software Conference Proceedings*, Sawtooth Software, Sequim.

Sawtooth Software (1998) "The CBC/HB Module for Hierarchical Bayes Estimation," Technical Paper accessible from sawtoothsoftware.com web site.

Sawtooth Software (1999) "The ACA/HB Module for Hierarchical Bayes Estimation," Technical Paper accessible from sawtoothsoftware.com web site.

Sawtooth Software (1999) "HB-Reg for Hierarchical Bayes Regression," Technical Paper accessible from sawtoothsoftware.com web site.

HB PLUGGING AND CHUGGING: HOW MUCH IS ENOUGH?

*Keith Sentis and Lihua Li*¹
Pathfinder Strategies

INTRODUCTION

Over the past five years, data analytic techniques that use a Hierarchical Bayes (HB) approach have burst on the marketing research scene. A practical overview by Gelman, Carlin, Stern and Rubin (1995) and papers by Allenby and Ginter (1995) and Lenk, DeSarbo, Green and Young (1996) are landmarks in this flurry of activity.

The evidence is now quite clear that estimates of conjoint utilities based on HB algorithms require less data than other estimation procedures and yet yield choice models with greater predictive accuracy. Our particular interest in HB methods stems from this seemingly magical property, namely, that HB methods give you “more from less”. Since our clients constantly ask us to give them more for less, this new procedure held out considerable promise.

Our interest was motivated also by the fact that most of our client projects are centred around individual estimates of response functions yet our datasets tend to be minimalist in that we rarely have the luxury of collecting two dozen choice tasks. Given the rather sparse nature of our datasets, our experience with other individual estimation procedures such as ICE has been less than inspiring. When the HB process that offered “more for less” at the individual respondent level was made accessible by the Sawtooth Software modules, it really caught our attention. In some preliminary explorations with a beta version of the HB-CBC module, the HB utilities looked very promising. So when the HB module for CBC was released, we were eager to learn about its behaviour.

We all know that there is no such thing as a free lunch, and sure enough, the magical property of HB that was so compelling comes at a cost. Namely, the huge amount of processing time that is required to perform the magic. We were appalled at the amount of time required to run these HB analyses. Our reaction was reminiscent of how shocked we were when the Sawtooth CBC Latent Class module was released and segmentations took several hours to finish. After years of working with KMeans clustering and becoming accustomed to the relatively quick computing times required by these procedures, the amount of computing time that the Latent Class algorithm took was off the scale. Well, compared to HB, Latent Class analyses finish in an eyeblink. When Lihua and I set up the first HB-CBC run, we were eagerly watching the screen to see what was happening. After a bit of churning, this little message appeared to the effect that we should come back later to see the results – 56 hours later! With the wonderful set of clients we have, coming back in 56 hours is not usually an option. So Lihua and I set out to explore how this magical HB module really behaves.

¹ The authors wish to thank Rich Johnson for an extended series of stimulating discussions throughout the course of this project. We also thank Carol Georgakis for contributing her graphics expertise and Maria Ioia for logistical support.

FIRST EXPERIMENTAL DESIGN

One would expect that all this plugging and chugging would have some influence on the quality of the estimates that are produced. HB procedures grind away on two distinct stages of iterations. The first stage involves a set of initial iterations (the burn-in) that are intended to allow the process to settle down. The second stage involves saving sets of estimates for each respondent.

We began our exploration of how the HB-CBC module behaves by looking at the effects of increases in the plugging and chugging that is done in the initial stage of the process. We then looked at what happens when you vary the number of saves during the second stage of the process. The initial results were interesting, but also somewhat puzzling and we decided that there would be something to learn by systematically varying these two independent stages of plugging and chugging. It seemed reasonable to use the Sawtooth recommendation on these two stages as the starting point in an experimental design.

In our first experimental design, the first factor was INITIAL ITERATIONS in which we examined what we expected to be a reasonable range of initial iterations. In particular, we wanted to know if grinding away at the initial stage of the HB process to a point beyond the Sawtooth default would yield anything useful. So we set the levels of this first factor at between 5,000 and 25,000 initial iterations, using 5,000 iterations as the increment.

The second factor in this design was the number of draws that are saved for each respondent. Given the nature of the estimation procedures, these draws are not independent, so it is useful to iterate a few times and skip these before saving the next draw. We used a skip factor of 10 for this experiment and for the rest of the work that I'll describe today. Our SAVES factor had four levels as shown below. We examined two levels below the Sawtooth default and one level above.

		Initial Iterations				
		5,000	10,000	15,000	20,000	25,000
Draws Saved Per Respondent	100					
	500					
	1,000				Sawtooth default	
	2,000					

This first experimental design had 20 cells, one of which corresponds to the Sawtooth default.

With this design, we could examine the extent to which each stage of the HB process influenced the quality of the estimates:

- does it help to do more initial iterations that allow the algorithm to settle down?
- do you gain more precision in the estimates by saving more draws for each respondent?
- are the two stages of the process compensatory in some way such that an interaction of the two factors will emerge?

The first dataset we analysed was from a food product study with a sample size of 839. Respondents went through 20 choice tasks. Each task had 6 concepts and NONE. We were estimating 24 parameters in this study.

To develop a criterion measure of the quality of the utilities, we held out tasks 5, 10 and 15 from each respondent. Since the tasks were generated randomly in Ci3, each respondent had a different set of hold outs.

We then did a series of HB-CBC runs on the 17 tasks. We ran the HB module in each of the 20 cells in our experimental design. The first HB run used 5,000 initial iterations with 100 saves. The second run used 10,000 initial iterations with 100 saves and so forth. Thus, we produced 20 sets of HB utilities, one set in each cell of the design.

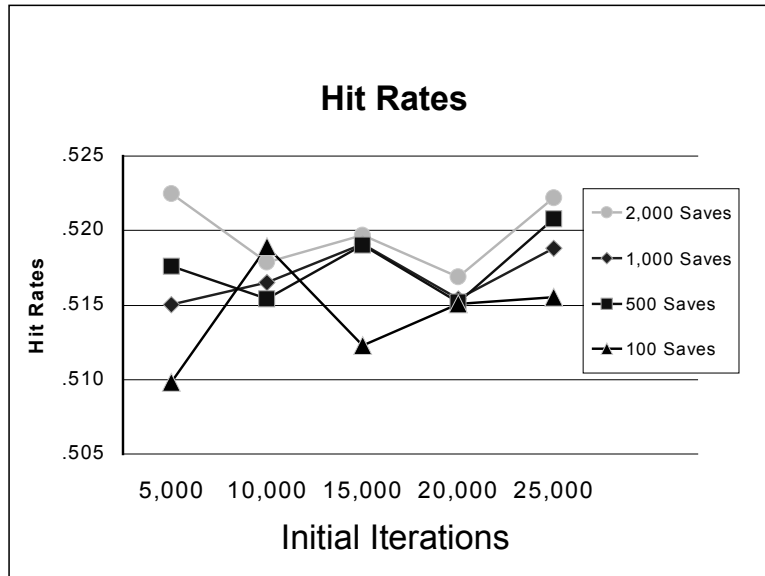
We used these utilities to calculate the hit rates for the hold out tasks. Since the utilities from each cell of the design are likely to be different, so are the hit rates for each cell in the design. These hit rates provided the dependent measure in our experimental design. In other words, the hit rates on the hold outs in each cell should vary as a function of the amount of initial iterating and of the number of saves per respondent.

We then conducted the identical type of analysis on a second dataset with 442 respondents who made 20 choices among six concepts plus NONE. In this study, we were estimating 19 parameters.

RESULTS FROM FIRST EXPERIMENTAL DESIGN

When we completed the 20 HB runs on each of these datasets, we had our first inkling of an answer to the question “how much plugging and chugging is enough?” The results would indicate whether more hours spent on the initial iteration stage returned anything of value. The results also would indicate whether more hours spent saving additional draws returned anything of value.

In the chart below, we plot the hit rates on the ordinate and use a different line on the graph for each row of cells in our experimental design.



Let's consider the last row of cells on our experimental design – the cells in which we saved 2,000 draws for each respondent. The hit rates are pretty flat as we increased the number of initial saves.

Looking at the cells with 1,000 saves, there is no real change as a function of increases in initial iterations. You can see that this set of cells includes the Sawtooth default which produces hit rates that are in the middle of the pack.

A similar pattern obtains in the cells with 500 saves. We can see that varying the amount of plugging and chugging results in hit rates that range from 0.515 to 0.523.

Suffice it to say that the pattern in these hit rates, if there was one, was not at all clear to us. With 100 saves, there is a bit of movement but not much. The hit rates across the entire 20 cells ranged from .510 to .523.

Based on our understanding of the theory that underpins the HB process, these results were quite unexpected. Aside from a tiny effect for number of saves, the short answer to the questions that we posed at the beginning of this exploration is “NO” and “PRACTICALLY, NO”. That is, additional initial iterations have no effect on the quality of the utilities and the number of saves has practically no effect on the quality of the utilities.

SECOND EXPERIMENTAL DESIGN

Based on our preliminary explorations with these two datasets, it seemed that the initial estimates used by the HB-CBC module are quite good and therefore, very little initial iterating is required before the process settles down. It also seemed that little was to be gained by saving a large number of draws for each respondent. Therefore, we saw the possibility of saving HB users a lot of valuable time by cutting back on both stages of iterations.

To test these ideas, we shifted our focus in the first experimental design upward and to the left. We wanted to examine the effects of fewer initial iterations as well as fewer saved draws per respondent. In our second design, we set the levels of the initial iterations at either 10 or 100 or 1,000 or 10,000. We set the saves per respondent at either 1 or 10 or 100 or 1,000. We used the same skip factor of 10 for these saves. In our second design, we had 16 cells formed by crossing four levels of initial iterations with four levels of saves. Also, for purposes of comparison, we continued to use the Sawtooth default of 20,000 initial iterations and 1,000 saves.

Second Experimental Design

		Initial Iterations				20,000
		10	100	1,000	10,000	
Draws Saved	1					Sawtooth default
	10					
	100					
	1,000					

Since the hit rates from our first experimental design seemed to hop around a little, we decided to run a few replications in each cell to smooth out any random jitters in the hit rates. We ran 5 separate HB-CBC runs in each of the 16 cells. As I mentioned, we used the Sawtooth default for a comparison point. We ran two separate HB-CBC runs in the comparison cell. We also ran five replications in two other comparison cells that I'll describe a bit later.

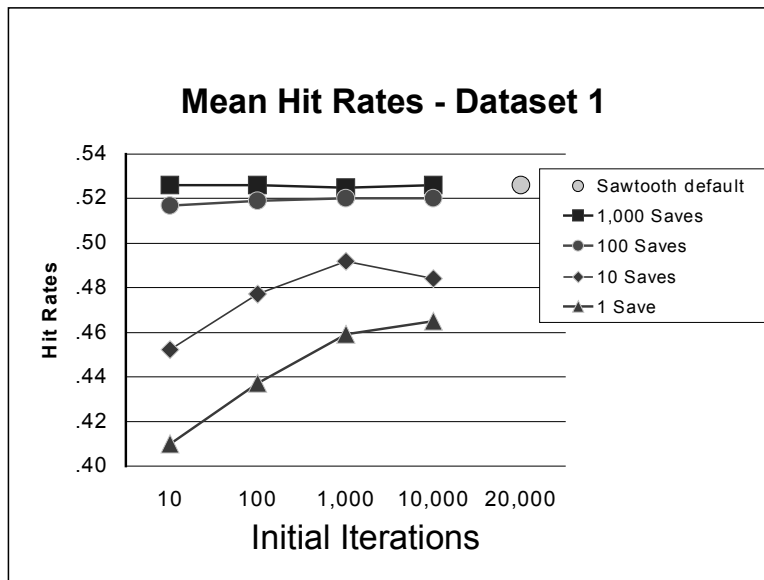
This generated 92 sets of utilities – 5 replications in each of the 16 cells and 12 replications in the comparison cells.

Since we were after a robust answer to our questions, we conducted these analyses on 22 datasets. In case you're not counting, that's more than 2,000 HB-CBC runs that required more than 10,000 hours of processing. We were plugging and chugging for several months on this project.

RESULTS FROM SECOND EXPERIMENTAL DESIGN

The first data set that we examined in the second experimental design was from another FMCG study with 957 respondents who made 14 choices among 3 concepts or NONE. We were estimating 16 parameters.

We conducted the same type of analysis on this dataset that we had done previously by using three of the choice tasks as hold outs.



Recall that we have five sets of utilities in each cell so the hit rates shown on the chart are the means from the five sets of utilities.

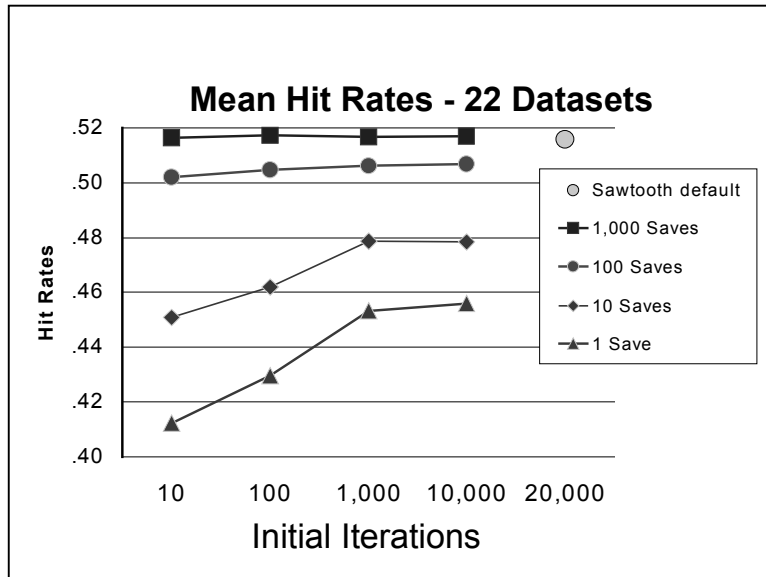
In the first row in our design, we examine the effect on the hit rates of more initial iterations when we saved only one draw per respondent. As you can see, the mean hit rate improves markedly with more initial iterations. It looks like this improvement begins to taper off after 1,000 initial iterations.

We also see that saving 10 draws gives us a bump in the quality of the utilities. With only 10 initial iterations, the hit rate performance with 10 saves almost reaches the maximum of the 1 save cells. The combination of 10 saves and 1,000 initial iterations yields better hit rates than any of the 1 save combinations. Once again, the tapering off beyond the 1,000 initial iteration mark is evident.

Looking at the cells with 100 saves shows a startling result. There is virtually no effect for number of initial iterations. The same flat line obtains for the 1,000 saves with only a very small improvement in the hit rates compared to 100 saves. The hit rates from the cell with the Sawtooth default values are shown for comparison.

The maximum hit rate of slightly more than .52 is obtained with 1,000 saves. This same hit rate is achieved irrespective of the number in initial iterations. We certainly did not expect this result. However, the pattern that emerged from each of the 22 datasets was very similar to the one that obtained in this first dataset.

Across a set of 22 datasets that include both packaged goods and services, having samples from 100 to 900, with between 14 and 20 choice tasks and between 10 and 42 parameters, the pattern of hit rates in our second experimental design is remarkably similar.



More initial iterations are helpful when saving only 1 draw per respondent. Saving 10 draws improves the quality of the utilities overall and the effect of increased initial iterations is still evident. Saving 100 draws improves the hit rates further and the positive effect of more initial iterating is nearly wiped out.

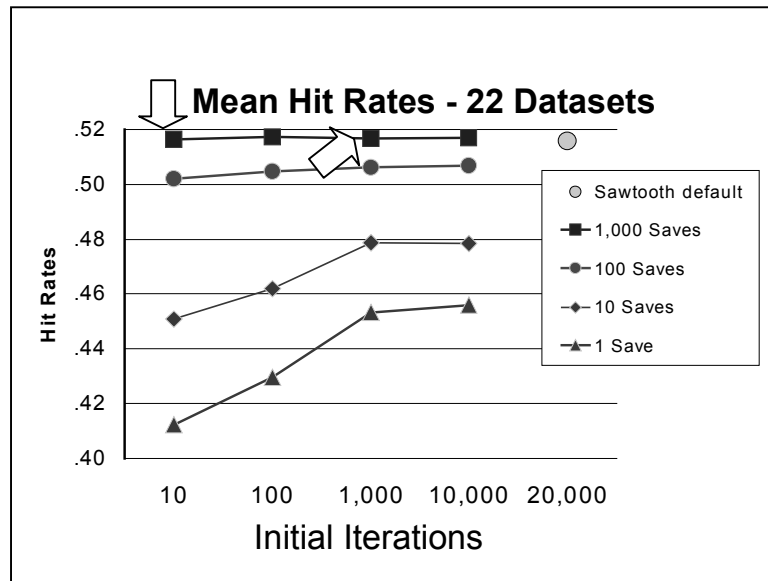
Saving 1,000 draws is better still and there is no advantage of more initial iterations. The cell with the Sawtooth defaults yields hit rates that are the same as others with 1,000 saves.

So 10,000 hours of computing later, we have answers to our questions. You will achieve the best quality utilities in the shortest amount of time by stopping the initial iterating right away and by saving 1,000 draws.

You may recall that we had looked at 2,000 saves in our first experimental design. We wanted to know if there was anything to be gained by pushing the SAVES factor further along. So we developed another set of comparison cells to address this question. On our 22 datasets, we ran the five replications in two cells with 2,000 saves:

- 10 initial iterations and 2,000 saves
- 1,000 initial iterations and 2,000 saves

The mean hit rates for the comparison cells with 2,000 saves are flagged with arrows in the chart below.

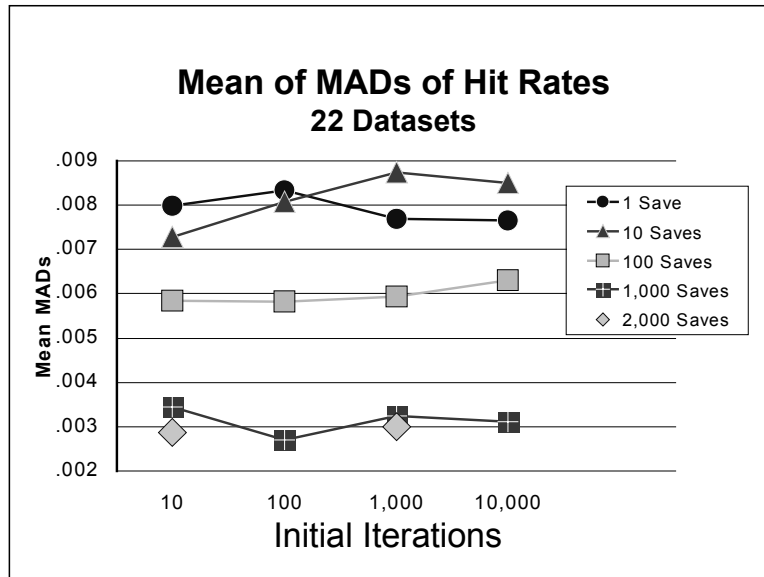


There is no improvement in the hit rates gained by saving 2,000 draws.

STABILITY IN THE QUALITY OF THE ESTIMATES

So we're nearing the end of this saga. Saving more draws has a positive effect on the quality of the HB utilities. This improvement stops after 1,000 saves. With 1,000 saves, there is no practical effect of increased initial iterations. That said, we wondered about the stability of the quality of the HB utilities. Perhaps the additional plugging and chugging would be beneficial by reducing the variability in the estimates.

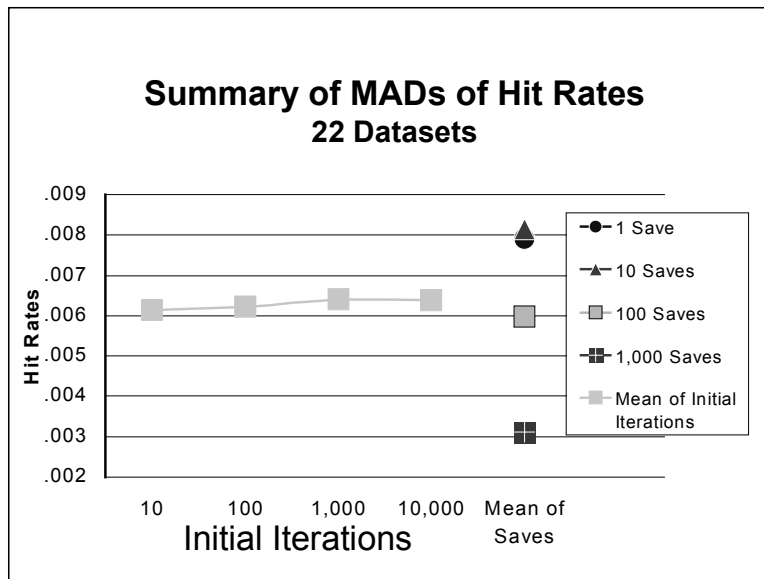
As just described, we had calculated the mean hit rate across the five sets of utilities in the 16 cells for 22 datasets. To address the variability issue, we calculated the absolute deviations from the cell mean and averaged these absolute deviations. This measure of variability is the Mean Absolute Deviation and is abbreviated MAD. We averaged these MAD scores across the 22 datasets. This is the mean of the MADs and the results are shown in the next chart.



At the bottom of this graph, the cells with 2,000 saves and with 1,000 saves have the lowest variability as manifest in the lowest Mean MADs. The cells with 100 saves have the next lowest variability. Saving 10 draws or one draw yields comparable levels of variability that are higher than 100 saves. This analysis shows that the stability of the utilities does not vary as a function of the amount of initial iterating.

This next chart shows the main effects that I just described. Saving additional draws reduces the variability in the quality of the HB utilities. The findings here are clear:

- if you save 1,000 draws, you don't need to worry about being unlucky because your random number generator is having a "bad-hair day"
- doing more initial iterations does not reduce the variability in the estimates.



SUMMARY OF RESULTS FROM SECOND EXPERIMENTAL DESIGN

In our second experimental design, the cell with 10 initial iterations and 1,000 saves performs as well as any of the 16 cells in the design both in terms of absolute hit rate and variability in the hit rates. There is a theoretical rationale for doing more initial iterations that relates to the “junk” in the initial estimates. Since there is very little additional computing time required to go from 10 initial iterations to 1,000 initial iterations, it doesn’t hurt to do so.

So here are the numbers that will most efficiently yield the HB magic:

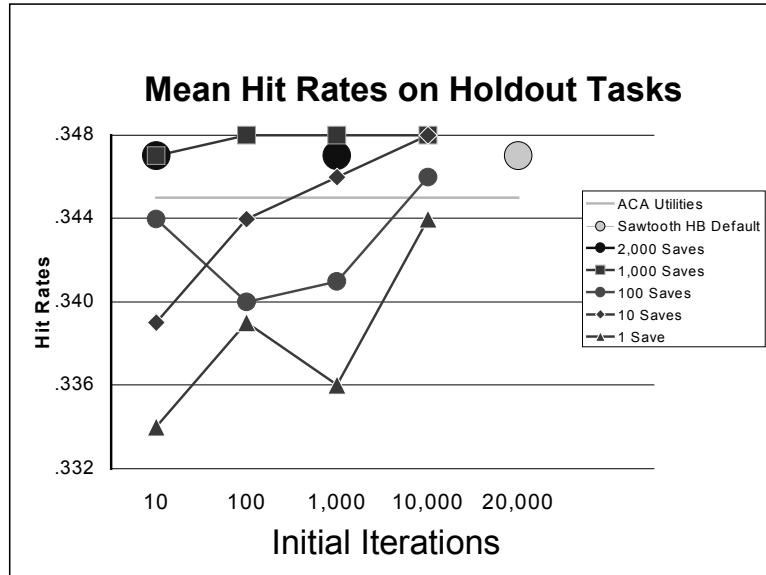
- 1,000 initial iterations
- 1,000 saves

The answers to our initial questions are that 1,000 and 1,000 are the magic HB numbers.

DO THESE RESULTS GENERALISE?

We wondered if the results from the HB-CBC module would generalise to the other two Sawtooth HB modules. To have a look at this issue, we analysed an ACA dataset using the same 16 cell design and the comparison cells. We used the HB-ACA estimation model with pairs only and constraints. We also ran the normal ACA utilities as another comparison point.

Again we had the 92 sets of utilities plus the normal ACA utilities. We used these to calculate the hit rates for five holdout choice tasks that had been administered after the ACA portion of the interview.



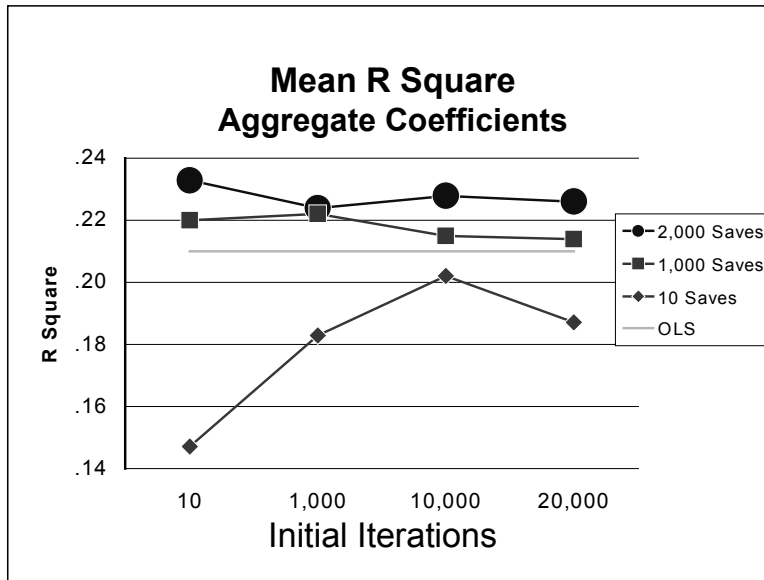
The results shown in the chart above are similar to those we found with the HB-CBC module:

- with small number of saves, increased initial iterating improves the hit rate
- when saving 1,000 draws, the results are just as good with few initial iterations as they are with many initial iterations
- the HB utilities with 1,000 saves are better than the normal ACA utilities
- saving 2,000 draws offers no advantage

Note, however, that these differences are rather small.

Next we looked at the Sawtooth HB-Regression module. We analysed a dataset consisting of about 1,300 cases with an average of approximately 7 cases per respondent. We held out 15% of the cases that were selected at random. For this analysis, we selected six of the cells from our second design as well as the comparison cells with 2,000 saves. Unlike the HB-CBC and HB-ACA modules, the HB-REG module cannot base its initial estimates on the data. Therefore, we had a look at 20,000 initial iterations also to see if additional “burn-in” would be of benefit.

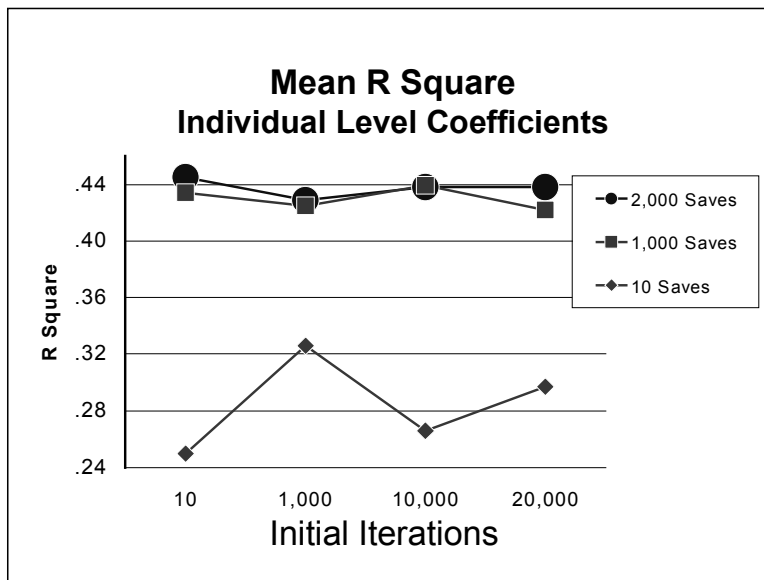
We ran five replications of the HB-REG routine in these 12 cells. Each of these 60 runs produced a set of aggregate coefficients and a set of individual level coefficients. We used these coefficients to predict the value of the dependent variable in the hold out cases. We then correlated the predicted value with the observed value in the holdout cases and squared this to yield an accuracy score for each set of coefficients. This r square measure is a rough analog to the hit rate that we have examined with the other 22 CBC datasets and the ACA dataset. We also computed the OLS coefficients as a comparison.



Looking at the aggregate coefficients, the pattern is similar to what we have seen before:

- with 10 saves, increases in initial iterations are helpful
- the HB coefficients with 1,000 and 2,000 saves are slightly better than the OLS estimates

The next chart shows the individual level results from this same dataset.



Of course, the fits are much better than those from the aggregate coefficients. Here the pattern in the HB cells is the same as the HB-CBC results, namely that once you save 1,000 it doesn't matter how much you initially iterate.

So our initial explorations suggest that our HB-CBC findings will generalise to the HB-ACA and HB-REG modules.

CONCLUSION

In conclusion, after nearly 19 million iterations, we can state with a fair degree of confidence this answer to the question that we posed at the beginning of our project:

- you do need to save 1,000 draws, but good results do not require the 20,000 initial iterations that are suggested as defaults

REFERENCES

- Allenby, G. M. and Ginter, J. L. (1995) "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, (November) 392-403.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) "Bayesian Data Analysis," Chapman & Hall, Suffolk.
- Lenk, P. J., DeSarbo, W. S., Green, P. E. and Young, M. R. (1996) "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15, 173-191.

PREDICTIVE VALIDATION OF CONJOINT ANALYSIS

Dick R. Wittink
Yale School of Management

ABSTRACT

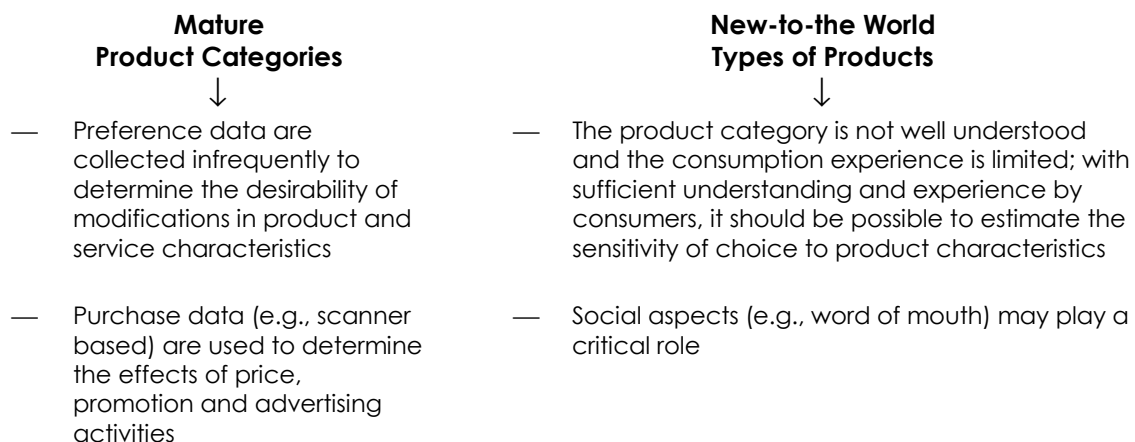
We describe conditions under which conjoint should predict real-world behavior. However, practical complexities make it very difficult for researchers to obtain incontrovertible evidence about the external validity of conjoint results. Since published studies typically rely on holdout tasks to compare the predictive validities of alternative conjoint procedures, we describe the characteristics of such tasks, and discuss the linkages to conjoint data and marketplace choices. We describe arguments relevant to enhancing the forecast accuracy of conjoint results, and we provide empirical support for these arguments.

INTRODUCTION

Conjoint is traditionally used to help managers forecast the demand (preference share), conditional upon a product category purchase, for continuous innovations. The preference share is the predicted share for a product configuration, given availability and awareness of the alternatives included in a respondent's consideration set. A continuous innovation refers to relatively minor changes in existing products or services. Specifically, if consumers do not have to make fundamental changes in their behavior after they adopt the new product, we classify a new product as a continuous innovation. The reason conjoint is applicable to continuous innovations is that respondents can easily imagine their liking for possible product configurations. By contrast, for discontinuous innovations the analyst should use more elaborate data-collection procedures. Urban, Weinberg and Hauser (1996) discuss enhancements to a conjoint exercise that facilitate forecasting of really new products (discontinuous innovations).

To understand conjoint's role in marketing, consider the dichotomy in Figure 1.

Figure 1
A Dichotomy of Product Types



As suggested in the left column of Figure 1, conjoint is highly applicable to mature products for which new features may enhance their attractiveness to customers. Conjoint-based surveys can provide insight into how customer behavior will change if existing products are modified or new items are introduced in a category. Also, if price has not varied much in the marketplace, conjoint can be used to help management understand consumers' price sensitivities by varying prices in hypothetical product descriptions. Purchase data may be used to understand how price, advertising and promotion affect the demand.

In the right column of Figure 1, the applicability of conjoint is low. Yet for discontinuous innovations (products that are new to the world, not just variations on existing products) managers often have the greatest need for information about demand sensitivities. Analysts can still use conjoint if, prior to the conjoint task, advertising and other communication media are used to educate respondents about the category, and respondents have an opportunity to familiarize themselves with the product in a consumption context. For new-to-the-world types of new products, social influence processes must also be accommodated in respondent selection and in market simulations. In addition, the interest for new-to-the world types is primarily in product category demand, whereas conjoint is designed to facilitate the prediction of shares, conditional upon a product category purchase. Thus, the applicability of conjoint is greatest for mature product categories.

FORECASTING MARKETPLACE BEHAVIOR

Conjoint results should be predictive of marketplace behavior, for the target market they represent, if the following conditions hold:

- respondents are representative of marketplace decision makers in the product category, for example in the sense of comprising a probability sample of the target market;
- the set of attributes and the levels is complete in the sense that relevant characteristics of current and future products can be accommodated in market simulations;
- the estimated preference model is a valid representation of how consumers make trade-offs among product attributes in the marketplace, and the conjoint exercise induces respondents to process information as they do in the marketplace;
- respondents make their marketplace choices independently or, if not, word-of-mouth and other social effects commensurate with marketplace behavior are accommodated;
- respondents' predicted choices are weighted by their purchase intensities;
- the set of alternatives for which the analyst makes predictions is the set the respondent considers, and it reflects expected consumer awareness and expected availability of the alternatives at the future time period for which predictions are made.

Conjoint-based marketplace forecasts are conditional upon the set of attributes, consumers' awareness of alternatives, the availability of alternatives to consumers, etc. However, one can accommodate additional complexities to produce forecasts that approximate actual conditions. For example, one can ask respondents to describe their purchase frequency or purchase amount during a specified time period and the brands they consider in the product category. With this information, respondents' predicted choices can be weighted and made conditional upon

availability and awareness. Without such adjustments, one cannot expect forecasts of marketplace behavior to be accurate.

Specifically, suppose that a conjoint exercise provides predictions of choices among a set of alternatives. To determine the accuracy of such predictions, we can specify the set of alternatives available in the marketplace at a given time. Let the predicted share for alternative i that belongs to this set be \hat{S}_i . This predicted share can and should reflect differences in purchase frequencies and/or purchase amounts between respondents. Even then, however, the predicted share can still be systematically different from the observed share S_i . To be able to predict actual marketplace shares we also need to take into account differences between the alternatives in availability and awareness (both also weighted by purchase frequency and/or amount) as well as differences in other relevant factors, such as promotions, that are outside the conjoint design. For each alternative, the relevant comparison is conceptually the following. We want to determine how the predicted conjoint-based share \hat{S}_i^a (adjusted for availability and awareness) compares with the actual share S_i^p (adjusted for promotions and other marketplace factors outside the conjoint exercise).

Srinivasan and deMaCarty (2000) provide an interesting variation on this approach. They propose that the various elements that complicate the validation exercise can be eliminated under certain conditions. Specifically, if one conjoint study leads to the introduction by a firm of (at least) two items in the product category studied, at about the same time and with comparable marketing support, then the ratio of preference shares predicted from the conjoint results should correspond closely to the ratio of market shares for the two products. The use of ratios of shares for the two products eliminates the effects of variables excluded from the exercise if, in addition to the conditions stated above: (1) the demand function is properly characterized by a multiplicative model and (2) the sales effects of the marketing activities are about the same for the two products. The use of ratios of shares generated by a single conjoint exercise also eliminates systematic effects that could be attributed to, for example: (1) the type of method used (Srinivasan and deMaCarty used self-explicated data which can produce similar results as conjoint studies), (2) the set of attributes used, (3) the selection of respondents, and (4) the time of data collection. For these claims to hold there must also be no dependencies between the effects of such factors and the two products.

PREDICTIVE VALIDITY

Conjoint analysis can provide accurate predictions of marketplace behavior.

Conjoint analysis is popular in many organizations. Surveys of the commercial use of the method indicate extensive and growing numbers of applications in virtually all consumer and industrial markets, products and services, in multiple continents (for North American applications, see Cattin and Wittink (1982) and Wittink and Cattin (1989); for European, see Wittink, Vriens and Burhenne (1994)). One likely reason for the method's popularity is that it provides management with information about (potential) customers that differs from management beliefs. For example, the tradeoffs between product attributes that can be inferred from conjoint results often differ dramatically from what management believes them to be.

Unfortunately there is not much hard evidence that *future* market outcomes are predictable. Benbenisty (1983) compared the market share predicted by conjoint analysis with the result achieved in the marketplace for AT&T's entry into the data-terminal market. The conjoint model predicted a market share of eight percent for AT&T four years after launch. The actual share was just under eight percent. However, it is not clear how various complexities were handled. Nor is it obvious how the timing of a predicted share is handled.

In a study of commuter modes (auto, public transit, and car pool), Srinivasan et al. (1981) forecast travel mode shifts, if gasoline prices increase, that turned out to be consistent with actual changes in market shares. Kopel and Kever (1991) mention that the Iowa lottery commissioned a study to identify new-product opportunities after it experienced a decline in revenues. It used conjoint results to create an option within an existing lottery game that increased sales for the game by 50 percent.

Srinivasan and deMaCarty (2000) report that Hewlett Packard (HP) conducted separate self-explicated studies on four categories: portable personal computers, tabletop personal computers, calculators, and universal frequency counters. Following each study HP introduced two products. For each pair of products, the product predicted to have the greater share did obtain the greater share ($p < .10$). And the ratio of market shares was within two standard errors of the ratio of preference shares for three of the four predicted ratios. These results provide strong support for the validity of self-explicated models in predicting marketplace choice behavior.

A few studies show that *current* market conditions can be reproduced (e.g., Parker and Srinivasan (1976); Page and Rosenbaum (1987); Robinson (1980)). A Harvard Business School case on the Clark Material Handling Group concerns the application of conjoint analysis to product-line and pricing changes in hydraulic-lift trucks. The prediction of market share for the study's sponsor appears to have been fairly accurate (Clarke 1987). Louviere (1988) focuses on the validity of aggregate conjoint choice models and concludes that well-designed studies can predict marketplace behavior.

One of only a few published studies that predict *future* marketplace decisions at the *individual* level concerns MBA job choices at Stanford University (Wittink and Montgomery 1979). In this study, MBA students evaluated many partial job profiles, each profile defined on two out of eight attributes manipulated in the study. About four months later the students provided evidence on the same attributes for all the job offers they received and they indicated which job they had chosen. Wittink and Montgomery report 63-percent accuracy (percent hits) in predicting the jobs students chose out of those offered to them, compared to a 26-percent expected hit rate if the students had chosen randomly (they averaged almost four job offers).

In this study, the hit rate is far from perfect for the following reasons: (1) job choice is a function of many job characteristics, out of which only eight attributes with levels common to all respondents were used; (2) the job offers varied continuously on many attributes, whereas only a few discrete levels were used in the study; (3) the preference judgments were provided by the MBA students, and no allowance was made for the influence of spouses, parents or friends; (4) the preference judgments were provided prior to many recruiter presentations and students' visits to corporate locations; (5) the preference model assumed only main effects for the eight attributes; and (6) the part worths were estimated for each student based on a modest number of judgments.

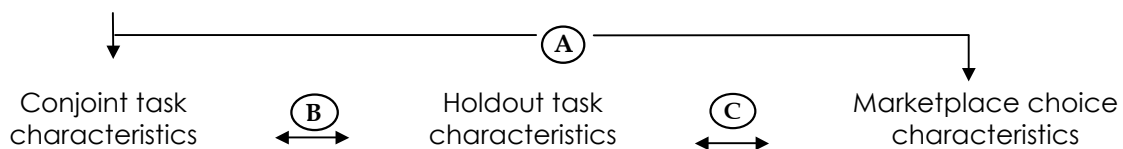
On balance, the published results of forecast accuracy are very supportive of the value of conjoint results. One should keep in mind that positive results (conjoint analysis providing accurate forecasts) are favored over negative results for publication. Nevertheless, the evidence suggests that marketplace forecasts have validity.

HOLDOUT TESTS

Since one must measure and control for many additional variables if one is to use marketplace behavior to validate conjoint results, much of the extant research relies primarily on holdout data. A typical holdout task consists of two or more alternatives, and respondents indicate which one they would choose. Respondents may face multiple holdout choice tasks, in which case the analyst has several opportunities to determine the predictive accuracy of the conjoint results. For the holdout choices to provide useful information, it is important that the characteristics of the tasks resemble marketplace choices as much as possible. At the same time the holdout task differs from the marketplace in that it eliminates the influence of many other factors, such as awareness and availability of alternatives. In addition, the holdout choices can be collected immediately whereas it may take some time before respondents make marketplace choices on which the conjoint results can be tested. Still, the holdout task may resemble the conjoint task more than it resembles marketplace choices, and since respondents provide holdout choices immediately after the conjoint task, predictive accuracy may be high by definition. Thus, the *absolute* amount of predictive accuracy observed in holdout choices will not generalize to marketplace choices. However, differences in how alternative conjoint methods perform in holdout choices should persist in actual marketplace choices, under certain conditions. So the *relative* predictive accuracies for alternative methods are expected to be applicable. We discuss below the characteristics of holdout tests in more detail.

Consider the diagram in Figure 2, in which we differentiate between the characteristics of the conjoint task, those of the holdout task, and those of marketplace choices. The question we address is the similarity in characteristics between the three types of data. Intuition might suggest that the conjoint task characteristics should resemble marketplace characteristics as much as possible, i.e. linkage (A) in Figure 2 should be very strong. However, this need not be the case.

Figure 2
Validation of Conjoint Results



To understand the dilemma for linkage (A), suppose that for a given product category marketplace choices are actually based on information on, say, 25 separate attributes. If the purchase decision concerns a costly item, such as an automobile, customers may spend days inspecting, deliberating about, and reflecting on the options. Yet respondents typically spend no more than about 15 minutes evaluating hypothetical products in a conjoint task. The question is what will respondents do in a conjoint task to resolve complex tradeoffs they would spend days

resolving in the real world? The conjoint analyst wants to obtain a valid understanding of these tradeoffs in a very short span of time. If the hypothetical products are described in a similar manner as marketplace alternatives, the time constraint will force respondents to simplify the task, for example by ignoring all data except for the two or three most critical attributes. To avoid this, the analyst can simplify the conjoint task so that detailed insights about a larger number of attributes are obtained from the respondents.

To accomplish this, the conjoint analyst may use a procedure that *forces* respondents to consider tradeoffs among all 25 attributes. The manner in which this is done varies. One possibility is to force the respondents to compare objects described on only a few attributes at a time. By varying the attributes across the preference questions, the analyst can obtain information about trade-offs among all 25 attributes. Thus, it is possible that the simplification in the conjoint task which appears to reduce the similarity in characteristics between conjoint and marketplace may in fact enhance the predictive validity of conjoint results to marketplace choices. This occurs if the conjoint task can be structured so as to facilitate the respondents' performing compensatory processing for all the attributes between which purchasers make trade-offs in marketplace choices.

To complete the discussion, we also need to consider differences c.q. similarities between the characteristics of the conjoint and holdout tasks, and those of the holdout tasks and marketplace choices. The dilemma with *holdout* tasks is that one may argue their characteristics should be as similar as possible to those of the marketplace choice situation (linkage ©). Yet respondents may also simplify their approach to the holdout task. Thus, even if the holdout characteristics do resemble marketplace characteristics, the holdout choices may not resemble marketplace choices. It follows that how alternative conjoint methods perform in holdout tasks may not predict their marketplace performance. Essentially, the characteristics of the holdout task must still facilitate compensatory processing by respondents if the part of marketplace choices we want to predict is subject to compensatory processing. To the extent that this is the case, we expect that *differences* in performance among alternative procedures observed in holdout tasks generalize to the real world (external validity). This should be more so for holdout-choice tasks than for holdout-rating or ranking tasks, since choices more directly mirror marketplace decisions.

Finally, if the holdout task is to provide more than a measure of reliability, linkage (B) should not be strong either. To minimize the similarity in characteristics, the analyst can vary the description of attributes between the conjoint and holdout tasks. For example, the conjoint task may elicit preference judgments for individual objects defined on attributes described according to one order, while the holdout task may elicit choices from two or more alternatives defined on the same attributes but described in a different order.

VALIDATION MEASURES

Two measures of the validity of holdout results are commonly used. One measure is defined at the level of the *individual* respondent. It assesses how well the conjoint results can predict each individual's holdout choices. The common summary measure for this is the *proportion of hits*, where a hit is a choice correctly predicted. The result obtained on this measure is usually compared against what would be expected in the absence of information (random choices) and against the maximum possible which is a function of the reliability of the holdout choices (Huber

et al. 1993). This measure, proportion of hits, is especially relevant if management wants to predict marketplace choices of individual decision makers (e.g., in business-to-business markets where the number of customers is small).

The other measure is defined at the *aggregate* level. The argument in favor of an aggregate-level measure is that managers often need to predict (market) shares and not which members of the target market will purchase a specific alternative. In this case, we compare the proportion of choices for each holdout alternative with the proportion of predicted choices. The measure of forecast error is the *deviation between holdout shares and predicted shares*. To determine the quality of aggregate forecasts in holdout tasks, we should compare the results against the expected result based on random choices (the minimum) and against the result based on the maximum possible accuracy which depends on the holdout share reliabilities (Huber et al. 1993).

Although these two summary measures are positively related, they can conflict. The prediction of each respondent's holdout choices is based on that person's estimated preference function. This preference function may be misspecified, and the data-collection method may introduce additional biases. Such elements tend to reduce the *validity* of conjoint results and such reductions affect both summary measures. The *reliability* is determined by the error variance in each respondent's estimated function. Importantly, more complex preference functions can increase the validity but reduce the reliability of the results, relative to simple models. However, if we aggregate the predictions, errors due to unreliability tend to cancel while errors due to invalidity (bias) remain. The prediction of shares is therefore less sensitive to unreliability than is true for the prediction of individual choices. For this reason the two measures can disagree.

Hagerty (1986) introduced a formula that shows how the accuracy of a multiple-regression prediction depends on reliability and validity. For forecasts of holdout choices at the individual level, the accuracy can depend as much on the reliability as on the lack of bias. Thus, simple models, which often have high reliability, may outperform more complex models even if the complex models have less bias, if the focus is at the individual level.

At the aggregate level, a true model has (asymptotically) zero error, while for an incorrect model the error is attributable to the systematic difference in predicted and true values (see Krishnamurthi and Wittink (1991)). This suggests that everything that *enhances the validity* of the model should be included to maximize aggregate-level predictive validity. The unreliability of parameter estimates tends to increase with model complexity, but this unreliability has very little impact at the aggregate level. For aggregate-level forecasts, unreliability approaches zero as the number of respondents increases. Thus the model that has the smallest bias tends to provide the best *share* forecasts. Note that the model can still be estimated separately for each respondent. The aggregation does not involve the model, only the predicted and actual values of the criterion variable are aggregated.

Complex models provide better share forecasts than simple models do.

One can increase the complexity (validity) of a preference model by:

- a) including attribute interaction terms in addition to the main-effect variables;
- b) accommodating heterogeneity in attributes, levels and model parameters across respondents;

- c) allowing for maximum flexibility in the functional form that expresses how preferences depend on each attribute (e.g. using indicator variables).

Hagerty (1986) shows analytically and empirically that in predicting preference share, a complex model is likely to be more accurate than a simple model. For example, he shows in several conjoint applications that a model with attribute interaction terms (allowing the effect of changes in one attribute to depend on the level of another attribute) has better aggregate-level predictions than a model without these terms. A full factorial design according to which a conjoint analyst constructs hypothetical objects will of course allow for the estimation of all main- and interaction effects (especially first-order interactions) in a preference model. Importantly, if attributes interact, managers who contemplate making changes in the characteristics of an existing product will find that interacting attributes should change together. One implication is that alternatives available in the marketplace should also exhibit attribute correlations that are consistent with estimated attribute interaction effects (at least alternatives belonging to the same consideration set). However, under these conditions it is undesirable to ask respondents to evaluate objects with uncorrelated attributes, since this affects the ecological validity (Cooksey, 1996). Thus, the frequent use of partial factorial designs, which generate uncorrelated attributes but typically do not allow for the estimation of interaction effects, seems misguided, not only because of missing interactions but also because a design with uncorrelated attributes tends to create unrealistic objects.

For holdout tasks to create responses that show the superiority of a model with attribute interaction effects (over one without), it is important that the holdout stimuli have the characteristics that allow for such superiority to show. If the conjoint study is limited to the set of attributes on which *existing* products differ, it should be sufficient to use holdout choice alternatives that resemble existing products. However, it is likely that the study involves new features and attribute ranges that differ from the current marketplace. The challenge then is for the researcher to anticipate attribute interaction effects and allow for those effects not only in the conjoint design but also in the holdout task. For example, the stimuli used in the holdout task should then also represent the attribute correlations implied by the (expected) attribute interaction effects.

With regard to parameter heterogeneity, Moore (1980) shows that models which accommodate parameter heterogeneity produce superior predictions at the aggregate level over models that do not. Krishnamurthi and Wittink (1991) find that, at the aggregate level, the part-worth model (a preference function estimated with indicator variables representing all of the attributes) is empirically almost always superior to models that assume continuous functions for one or more attributes.

Although most researchers insist that respondents are heterogeneous, and that the parameters (e.g. part worths) should accommodate this heterogeneity, surprisingly little is done to allow the attributes and their levels to be heterogeneous across respondents. However, if it is beneficial to allow the part worths to differ across respondents, it should also be beneficial to let the set of attributes and levels be respondent-specific. Zwerina, Huber and Arora (1999) have pursued this issue in the construction of customized choice sets. They report superior forecast accuracy for their proposed customized choice design relative to a traditional design. The improvement is especially strong at the aggregate level. For example, they find that 9 choice sets based on a traditional design provide the same hit rate as 5 choice sets based on a customized design. Since

the customized design is stronger on validity (due to customization) but weaker on reliability (fewer choice sets), it is favored on the error in forecasting shares, as they observed in forecasting accuracy of shares.

Simple models may provide better forecasts of individual choices than complex models.

If we want to predict the choices of individual consumers accurately (i.e. achieve a high percent of correctly predicted choices for the responses in holdout tasks), we have to seek a balance between bias and unreliability. We can minimize bias (in predictions) through the use of complex models. However, as models become more complex, the number of parameters estimated tends to increase as well. For a given number of preference judgments per respondent (opinions about the “optimal” number vary considerably), the more parameters, the greater their unreliability or statistical uncertainty. And the more unreliable the parameter estimates are, the more unreliable the predictions of individual choices are.

Simple models have an advantage over complex models in the sense that parameter estimates tend to have lower variance. If the ratio of the number of data points over the number of parameters is small, as is more likely to occur for complex models, parameter estimates may fall outside plausible ranges due to statistical uncertainty. Importantly, a simple model will outperform a complex model at the *individual* level if the loss due to bias inherent in the estimated simple model is smaller than the gain due to lower statistical uncertainty.

Hagerty (1986) finds that the same models with interaction terms that improve aggregate-level predictions make individual-level predictions worse. Green (1984) also finds that the inclusion of interaction terms often reduces models’ predictive validity at the individual level. For functional form, Krishnamurthi and Wittink (1991) show that the part-worth model can be outperformed by a model with fewer parameters at the individual level. However, with regard to parameter heterogeneity, Wittink and Montgomery (1979) obtain superior predictions at the individual level with models that fully accommodate parameter heterogeneity. The percent of hits is highest when respondent-specific parameters are estimated, and lowest when common parameters are used for all respondents. These results suggest that the improvement in validity is often (much) greater when models accommodate respondent heterogeneity in parameters than when the functional form for the main effects of continuous attributes is completely flexible or interactions between attributes are accommodated.

Constraints imposed on estimated parameters (e.g. part worths) reduce share forecast accuracy but improve forecasts of individual choices.

Srinivasan, Jain and Malhotra (1983) show that one can increase the percent of choices correctly predicted by imposing constraints on parameters based on a priori knowledge of the preference ordering for different levels of an attribute. Sawtooth Software (1997) refers to various studies that show parameter constraints improve the hit rates of full-profile (hypothetical object described on all the manipulated attributes) conjoint utilities, with an average improvement of nine absolute percentage points. However, for ACA (Johnson 1987), the average improvement is only two absolute percentage points. The explanation for this difference between full profile and ACA is that the ACA solution is partly based on self-explicated data which reflect the parameter constraints. Sawtooth concludes for any conjoint method it is useful to

impose constraints on the part worths of attributes with strong a priori preference orderings of the levels, if the accuracy of individual-level forecasts is to be maximized.

For many attributes it is possible to state expectations about a respondent's preference order for alternative levels. For example, we expect that respondents prefer lower prices over higher prices, everything else being equal. Yet the part worths for a given respondent may be inconsistent with this expectation due to imprecision (statistical uncertainty) in the part worths. Removal of the inconsistency will then improve the forecast accuracy at the individual level. However, since statistical uncertainty of individual part worths cancels out at the aggregate level, the use of constrained estimation will not improve aggregate predictions. In fact, since the inconsistency may also reflect an unusual way of thinking, leaving inconsistencies alone is the best approach for aggregate-level forecasts. Indeed, we expect aggregate-level forecast accuracy to decrease because the removal of inconsistencies is usually done in a one-directional manner.

To illustrate, suppose that all respondents are indifferent between two alternative colors, green and red, for a product. Even then, the part worths will not be zero due to sampling error. However, on average the part worths for both colors will approach zero as the number of respondents increases.

Now suppose that we have reason to believe that no respondent can logically prefer red over green. We would then constrain the color green's part worth to be no less than red's part worth, for every respondent. This may improve our ability to predict (holdout) choices for the respondents whose part worths are so constrained. However, if truly all respondents are indifferent between red and green, imposing the constraint that the color red's part worths are no less than green's would be equally helpful. Importantly, if we only impose the former constraint and not the latter, we will have average part worths that favor green. It is this imbalance that causes the constrained parameter estimation (as it is typically practiced in conjoint) to reduce the forecast accuracy at the aggregate level. For example, Johnson (1999) examines the forecast accuracy of the ACA/HB (Hierarchical Bayes) module. Based on a holdout task with six choices from three alternatives each, he finds that the imposition of constraints improves the hit rate but it also increases the mean absolute error in share predictions. If, however, "balanced" constraints were imposed in the sense that each constraint in one direction is offset by a corresponding constraint in the opposite direction, the effect on the aggregate measure of forecast accuracy will be neutral.

Combining the results from different methods provides better forecasts than single methods do.

Marketplace choices are influenced by many factors. It is inconceivable that one method for collecting preferences about hypothetical options can tap into all relevant elements. However, each method of preference measurement can provide both insights common to all methods and unique insights obtainable only from that method. If each conjoint method captures only a subset of the real-world complexities, and the methods differ in the types of complexities they capture, then a combination of output from different approaches can provide better forecasts than any single method.

The different methods of data collection have unique strengths and weaknesses. The full-profile method is realistic in that each profile shows information on all attributes included in the study (similarity in task and marketplace choice characteristics). However, when filling out

surveys, respondents try to reduce the information processing burden. Thus, they can be expected to use simplification strategies in the conjoint task that they might not use in the marketplace. In ACA, the assumption is that respondents will limit their attention to a small number of attributes at a time. In practice, analysts using ACA usually pair objects defined on just two attributes, under the assumption that subjects are more likely to use compensatory processing when objects are defined on just a few attributes. In related research, Payne (1976) shows that compensatory processing is more evident when respondents choose between two alternatives than when they choose between larger numbers of alternatives. Thus, Johnson (1987) expects that ACA output has external validity, even if the task characteristics in ACA differ from real-world-choice characteristics. By using results from different methods, analysts should be able to combine the strengths that differ between methods.

ACA already combines data from different methods, as this method collects self-explicated data (respondents rate the importance of the difference between the best and worst levels, separately for each attribute) as well as preference intensity judgments for paired partial profiles. An attractive aspect of ACA is that it customizes the partial-profile characteristics based on each respondent's self-explicated data. It obtains the final preference function coefficients by pooling the two types of data.

Huber et al. (1993) observe that the final ACA solution provides holdout choice predictions, at the individual and at the aggregate level, that are superior to those of the initial ACA solution (which is based only on self-explicated data). They obtain even better predictions by combining full-profile results with ACA output, based on a weighting of the results from the different conjoint methods that optimizes predicting holdout choices (weighting the results from different methods equally would also have given them more accurate forecasts than any single method). Cattin, Gelfand and Danes (1983) also report achieving superior predictions by adding self-explicated data to conjoint results. However, Srinivasan and Park (1997) fail to improve the predictions of job choices from self-explicated data when they combine these with the results of full-profile conjoint. That is, the best predictions of individual choices were obtained by giving zero weight to the full-profile conjoint results.

Motivating respondents improves forecast accuracy.

Wright and Kriewall (1980) used experimental manipulations to test whether model-based predictions of college choice by high school students improve when an imminent commitment ("act as if tomorrow is the deadline for applying") was invoked. Relative to the control group, the college choice predictions for students confronted with the commitment were indeed more predictive of actual college choices. Wright and Kriewall also found that sending materials relevant to the conjoint survey in advance (and urging respondents to practice choice strategies) improved predictive accuracy. This strategy for eliciting responses is similar to political pollsters asking "If the election were held today, who would you vote for?" Essentially, these manipulations heighten respondents' involvement in the task in the sense that the questions posed or the preferences elicited become more relevant to respondents.

Wittink and Montgomery (1979) report the accuracy of job choice predictions for Stanford MBA's. At the first presentation of these results in a research seminar, one person asked if the predictions might be biased upward due to some respondents having chosen a job prior to the conjoint survey. For example, respondents may desire to minimize cognitive dissonance, and provide preference judgments for hypothetical jobs in such a way that the judgments would be

consistent with the actual job chosen. Wittink and Montgomery knew when each respondent accepted a job offer, and they were able to compare the percent of job choices correctly predicted between those accepting early (before the conjoint survey) and those accepting late. The difference in the percent correctly predicted was actually in favor of students who had not yet chosen a job at the time the survey was conducted. These students also reported taking longer to complete the survey and being more confident about their preference judgments. All these differences are consistent with the notion that the students who had not yet chosen a job were more motivated. Indeed, several students commented that they found the conjoint exercise a very useful means for them to confront tradeoffs between job characteristics. Thus, respondents who plan to make a decision in the near future, relevant to the topic of a conjoint study, should be more motivated to provide valid judgments than respondents who have no such plans. Further, by invoking imminent commitment, the quality of respondent judgments can be increased further.

If the holdout task is properly constructed, then a method designed to avoid a specific bias will have superior forecasts over other methods.

Conjoint analysis, like most survey-based methods, has limitations. One of these limitations is that the substantive results can be influenced by the *number* of levels the analyst chooses for an attribute in designing the conjoint study. Specifically, increasing the number of intermediate levels tends to increase the distance between the part worths for the best and worst levels. For example, suppose that in a conjoint study the lowest price is \$5 and the highest is \$7. Then, holding all other things constant, the inclusion of \$6 as an intermediate level will tend to enhance the importance of price, relative to a conjoint design restricted to \$5 and \$7. And including \$5.50 and \$6.50 will imply that price is even more important.

Researchers disagree about what produces this effect. One possibility is that it is caused by weaknesses in the measurement scale. Wittink, Krishnamurthy and Reibstein (1989) provide three dilemmas that show how the number-of-levels effect can be derived from rank-order preferences. They show, for example, that the ratio of the maximum possible weights (relative importances) for two attributes, one defined on three levels, the other on two, is 1.33. They also report experimental results that show that the magnitude of the number-of-levels effect is similar for ranks and for preference ratings. This suggests that ratings have rank-order-like characteristics. Indeed, Steenkamp and Wittink (1994) find that magnitude estimation, which should obtain strong (at least interval-scaled) preference measures, generates results with a reduced number-of-levels effect for respondents whose judgments satisfy the criteria for strong (metric) measurement, relative to other respondents.

Another possibility is that the effect emanates from a psychological or behavioral phenomenon (Poulton 1989). Respondents may pay more attention to an attribute as the amount of its variation (the number of levels) increases. Green and Srinivasan (1990, p. 7) favor this interpretation. Johnson (1991) provides evidence for a behavioral explanation. He describes an experiment in which respondents were told they could purchase a 17-inch TV with monophonic sound for \$200. They were asked about the value to them of improvements in both of the non-price attributes. Half the respondents were asked to provide the monetary value for a TV with a 21-inch screen and monophonic sound, and to also state their values for a 17-inch TV first with good-, then with excellent stereo sound. The other half of the respondents were similarly asked to give a value for excellent stereo sound (skipping the good sound), and to give values for 19-

followed by 21-inch screens. Across the experimental conditions, the ratio of average incremental values for the best option on sound (three levels versus two) was 1.31, while for screen size the ratio was 1.33. In both cases this ratio would be expected to be 1.0 in the absence of a number-of-levels effect.

These ratios are very similar to the ratio of maximum possible relative importances (three-versus two-level attributes) for rank order preferences reported by Wittink, Krishnamurthi and Reibstein (1989, p. 117). One possible explanation of Johnson's result is that the incremental dollar values have properties that resemble rank order data. Importantly, and independent of the reason for the number-of-levels effect, the literature on conjoint analysis focuses on the consequence of the number-of-levels effect on derived attribute importances of attributes. However, predictions of preference shares (and, hence, the results of market simulations) may also be affected by the number-of-levels effect.

Wittink, McLauchlan and Seethuraman (1997) use a modified ACA method that is designed to reduce the number-of-levels effect. In this method, the number of (intermediate) levels for a respondent depends on the self-explicated importance that the respondent assigns to each attribute. That is, the self-explicated importances obtained in ACA are used to customize the numbers-of-levels for the attributes in the conjoint design. The authors compare the predictive validity of the modified ACA method to that of the traditional ACA method to demonstrate the modified method's superiority. To accomplish this, they use a design for the holdout objects that is sensitive to the number-of-levels effect.

Wittink et al. assigned 600 respondents randomly to one of three conditions. They administered the modified ACA method to those in condition A. Condition B respondents saw the extreme levels and one intermediate level for all (five) attributes. The use of the same number of levels for all attributes in this condition is based on the idea that the number-of-levels effect is psychological in origin. That is, an attribute may become more important as it varies more frequently across the objects. Condition- C respondents saw the extreme levels plus two intermediate levels for two attributes, no intermediate levels for two other attributes, and one intermediate level for the final attribute. The number-of-levels effect is traditionally detected by comparing results between conditions B and C. That is, the distance between the part worths for the extreme levels of a four-level attribute (in condition C) should be greater than the distance between the part worths for the same extreme levels of a three-level attribute in condition B. Similarly, it should be smaller for a two-level attribute in C than it is for the same attribute with three levels in B.

To demonstrate a number-of-levels effect on predicted shares, Wittink et al. defined all objects in the holdout sets for all respondents on the extreme attribute levels (the only levels that all respondents would necessarily see in all three conditions). To understand how predicted choices can be sensitive to the effect, suppose a product is defined on only two attributes. In condition C, respondents are asked to choose between one alternative that has the best level of a four-level attribute and the worst level of a two-level attribute, and another that has the worst level of the four-level attribute and the best level of the two-level attribute. In condition B, respondents see exactly the same objects in the holdout task, but in the conjoint task the best- and worst levels represent attributes defined on three levels. In this example, the number-of-levels effect predicts that the object with the best level of a four-level attribute (and the worst level of a two-level attribute) will garner a higher percent of *predicted* choices in condition C than the

same object (which has the best level of the corresponding three-level attribute and worst level of the other three-level attribute) in B. This object will be favored more strongly in C because of a higher increase in the predicted preference due to the four-level attribute on which the object is favored, and a smaller decrease in the predicted preference due to the two-level attribute on which it is disfavored.

Wittink et al. constructed 10 unique holdout sets that differed on at least two attributes (each difference involving the best and worst levels). Every holdout set showed a difference in predicted shares between conditions B and C consistent with expectations. On average, the products had a predicted share of 46 percent in condition B but 57 percent in condition C, revealing a large number-of-levels effect on predicted shares.

To assess how much the modified conjoint version (condition A) can improve forecast accuracy, they employ a statistic that takes into account the predictive validity from ACA's self-explicated data and the unreliability of the holdout choices (since neither of these can be assumed to be equal across the experimental treatments). The modified ACA version (condition A) showed that the conjoint data improved the forecast accuracy (actual minus predicted share) relative to the maximum possible by 82 percent. This compared with 68 percent of the maximum possible improvement for the version with the same number of levels for all attributes (condition B), and 42 percent for the version in which the number of levels varied from two to four (condition C). These results show that a reduction in bias improves forecast accuracy at the aggregate level, if the holdout task is designed to be sensitive to the effect of the bias.

CONCLUSIONS

Conjoint analysis is an attractive method, used by managers in virtually all industries to quantify customer preferences for multiattribute alternatives. Its popularity suggests that the results have external validity. Published reports of the predictive accuracy of conjoint results to current and future marketplace choices are positive.

We have provided arguments that can help managers design conjoint studies such that they obtain accurate forecasts. For predictions at the aggregate level, they should use arguments that enhance the *validity* of conjoint results. On the other hand, for predictions of individual behavior, they must also consider the impact on *reliability*.

The quality of data collection may improve once we obtain a better understanding of the processes consumers use in making choices in the marketplace. For example, they may go through multiple stages in making decisions. In a first stage, they may use a noncompensatory process to eliminate many alternatives. Then, in a second stage they may use a compensatory process to evaluate the remaining alternatives. An implicit assumption in the typical conjoint study is that respondents' preferences pertain to such a second stage.

Conjoint results have been shown to have limitations. The number-of-attribute levels effect is one such limitation. Ongoing research should give us a better understanding of the source(s) for this effect. The following three scenarios indicate the importance of this research. One possibility is that real-world choices are also subject to a number-of-levels effect. For example, it is conceivable that the more alternatives under consideration vary on an attribute, the more consumers' attention will focus on this attribute. If this is true, then the conjoint analyst should first learn the characteristics of the alternatives each consumer actively considers in the

marketplace, so that the analyst can customize the number of levels in the conjoint task based on this information. Under this scenario, whatever context effects exist in the marketplace should be captured in the conjoint task.

A second possibility is that the effect occurs only in the conjoint task. If this effect stems from respondents becoming more sensitive to variation in attributes as the number of levels increases, then the analyst should use the same number of levels for each attribute in a conjoint study design. A third possibility is that the effect occurs because of other limitations as Wittink, McLauchlan and Seethuraman (1997) propose. In that case, analysts should customize the conjoint design or use enhanced estimation methods as done in ACA 4.0 (see also Wittink and Seethuraman (1999)).

Given the popularity of conjoint analysis, researchers should address the issues that currently limit its effectiveness. One interesting opportunity lies in using conjoint for continuous market feedback (Wittink and Keil 2000). For example, managers may discount ad hoc study results because they do not understand the method well enough, because the results are inconsistent with their beliefs or because they are rewarded for attending primarily to continuous monitoring systems (such as the information in market status reports for their brands). As interest in the use of customized marketing programs grows and as managers need frequent updates on customer preferences, researchers should determine in what manner and how frequently to update conjoint results efficiently.

REFERENCES

- Benbenisty, R. L. (1983), "Attitude Research, Conjoint Analysis Guided Ma Bell's Entry into Data Terminal Market," *Marketing News*, (May 13), 12.
- Cattin, P. & D. R. Wittink (1982), "Commercial Use of Conjoint Analysis: A Survey," *Journal of Marketing*, (Summer), 44-53.
- _____, A. Gelfand, & J. Danes (1983), "A Simple Bayesian Procedure for Estimation in a Conjoint Model," *Journal of Marketing Research*, 20 (February), 29-35.
- Clarke, D. G. (1987), *Marketing Analysis & Decision Making*. Redwood City, CA: The Scientific Press, 180-92.
- Cooksey, R. W. (1996), *Judgment Analysis: Theory, Methods and Applications*, San Diego: Academic Press.
- Green, P. E. (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, 21 (May), 155-9.
- _____, & V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54 (October), 3-19.
- Hagerty, M. R. (1986), "The Cost of Simplifying Preference Models," *Marketing Science*, 5 (Fall), 298-319.
- Huber, J. C., D. R. Wittink, J. A. Fiedler, & R. L. Miller (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, 30 (February), 105-114.

- Johnson, R. M. (1987), "Adaptive Conjoint Analysis," *1987 Sawtooth Software Conference Proceedings*, Sawtooth Software Inc., Sequim, WA, 253-66.
- ___ (1991), "Comment on 'Attribute Level Effects Revisited' ...", *Second Annual Advanced Research Techniques Forum*, R. Mora ed., Chicago, American Marketing Association, 62-4.
- ___(1999), "The ACA/HB Module," Sawtooth Software.
- Kopel, P. S. & D. Kever (1991), "Using Adaptive Conjoint Analysis for the Development of Lottery Games - An Iowa Lottery Case Study," *1991 Sawtooth Software Conference Proceedings*, 143-54.
- Krishnamurthi, L. & D. R. Wittink (1991), "The Value of Idiosyncratic Functional Forms in Conjoint Analysis," *International Journal of Research in Marketing*, Vol. 8, No. 4 (November), 301-13.
- Louviere, J. J. (1988), "Conjoint Analysis Modeling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity," *Journal of Transport Economics and Policy*, 22, 93-119.
- Moore, W. L. (1980), "Levels of Aggregation in Conjoint Analysis: An Empirical Comparison," *Journal of Marketing Research*, 17 (November), 516-23.
- Page, A. L. & H. F. Rosenbaum (1987), "Redesigning Product Lines with Conjoint Analysis: How Sunbeam Does It," *Journal of Product Innovation Management*, 4, 120-37.
- Parker, B. R. & V. Srinivasan (1976), "A Consumer Preference Approach to the Planning of Rural Primary Health Care Facilities," *Operations Research*, 24, 991-1025.
- Payne, J. W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis," *Organizational Behavior and Human Performance*, 16, 366-387.
- Poulton, E.C. (1989), *Bias in Quantifying Judgments*, Hillsdale: L. Erlbaum Associates.
- Robinson, P. J. (1980), "Application of Conjoint Analysis to Pricing Problems," in *Proceedings of the First ORSA/TIMS Special Interest Conference on Market Measurement and Analysis*, D.B. Montgomery and D.R Wittink eds. Cambridge, MA: Marketing Science Institute, 193-205.
- Sawtooth Software (1997), "Using Utility Constraints to Improve the Predictability of Conjoint Analysis," *Sawtooth Software News*, 3-4.
- Srinivasan V. & C. S. Park (1997), "Surprising Robustness of the Self-explicated Approach to Customer Preference Structure Measurement," *Journal of Marketing Research*, 34 (May), 286-91.
- ___ & P. deMaCarty (2000), "An Alternative Approach to the Predictive Validation of Conjoint Models," *Marketing Research*, forthcoming.
- ___, A. K. Jain, & N. K. Malhotra (1983), "Improving Predictive Power of Conjoint Analysis by Constrained Parameter Estimation," *Journal of Marketing Research*, 20 (November), 433-8.

- _____, P. G. Flaschbart, J. S. Dajani, & R. G. Hartley (1981), "Forecasting the Effectiveness of Work-trip Gasoline Conservation Policies through Conjoint Analysis," *Journal of Marketing*, 45 (Summer), 157-72.
- Steenkamp, J.B. and D.R. Wittink (1994), "The Metric Quality of Full Profile Judgments and the Number-of-Levels Effect in Conjoint Analysis," *International Journal of Research Marketing*, Vol 11, No. 3 (June), 275-86.
- Urban, G. L., B. D. Weinberg, & J. R. Hauser (1996), "Pre-market Forecasting of Really-new Products," *Journal of Marketing*, 60 (January), 47-60.
- Wittink, D. R. & D. B. Montgomery (1979), "Predictive Validity of Trade-off Analysis for Alternative Segmentation Schemes," in *Educators' Conference Proceedings*, Series 44, N. Beckwith et al., eds. Chicago: American Marketing Association, 69-73.
- _____ & P. Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53 (July), 91-6.
- _____ & P.B. Seetharaman (1999), "A Comparison of Alternative Solutions to the Number-of-Levels Effect," *1999 Sawtooth Software Conference Proceedings*, 269-81.
- _____ & S. K. Keil (2000), "Continuous Conjoint Analysis," in: A. Gustafsson, A. Herman and F. Huber (eds.) *Conjoint Measurement: Methods and Applications*, Springer-Verlag, 411-34.
- _____, L. Krishnamurthi, & D. J. Reibstein (1989), "The Effect of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters*, 1 (Number 2), 113-23.
- _____, M. Vriens, & W. Burhenne, (1994), "Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections," *International Journal of Research in Marketing*, Vol. 11, No. 1, (January), 41-52.
- _____, W. G. McLauchlan, & P.B. Seethuraman, (1997), "Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis," *1997 Sawtooth Software Conference Proceedings*, 227-40.
- Wright, P. & M. A. Kriewall (1980), "State of Mind Effects on the Accuracy with which Utility Functions Predict Marketplace Choice," *Journal of Marketing Research*, 17 (August), 277-93.
- Zwerina, K., J. Huber and N. Arora (1999), "Improving Predictive Performance by Customizing Choice Designs," working paper, Fuqua School of Business, Duke University.

COMPARING HIERARCHICAL BAYES DRAWS AND RANDOMIZED FIRST CHOICE FOR CONJOINT SIMULATIONS

Bryan Orme and Gary Baker
Sawtooth Software, Inc.

INTRODUCTION

Conducting market simulations is one of the most valuable uses of conjoint analysis data. Market simulations transform raw conjoint part-worths (which to a non-researcher can seem quite esoteric) to the more managerially satisfying model of predicting buyer choices for specific market scenarios.

The last few years have seen important new developments for estimating conjoint part-worths and simulating shares of preference. Foremost among the advances in part-worth estimation in our opinion is the use of hierarchical Bayes (HB) to estimate individual-level part-worths from conjoint or choice data. Another recent advancement has been the introduction of Randomized First Choice for conducting market simulations (Orme 1998, Huber *et al.* 1999) and its general availability within Sawtooth Software's conjoint analysis systems.

The typical application of conjoint analysis has involved estimating many independent parameters (attribute level part-worths) from only marginally more observations (questions/tasks). Despite this, the results (especially for predicting aggregate shares of preference) typically have been quite useful.

And then HB became available. HB is a very effective "data borrowing" technique that stabilizes part-worth estimates for each individual using information from not only that respondent, but others within the same data set. HB generates multiple estimates of the part-worths for each respondent, called *draws*. These multiple draws can be averaged to create a single vector of part-worths for each respondent (*point estimates* of the part-worths). One can also use those draws to estimate the variances and covariances of part-worths within each respondent. Another potential application is to use the draws themselves, rather than point estimates, in market simulations. Reviewing the theory and mathematics behind HB are beyond the scope of this paper. For the interested reader, we suggest the CBC/HB Technical Paper (Sawtooth Software, 1999).

Over the years, two main simulation models have been applied to part-worth data: First Choice and Share of Preference (logit). The First Choice model, while immune to IIA, is typically too extreme (and not tunable for scale). The Share of Preference model, while tunable, suffers from IIA. Randomized First Choice (RFC) adds back random variation to the point estimates of the part-worths during simulations. RFC is appropriate for aggregate (logit) or disaggregate part-worths (e.g. ACA, traditional full-profile conjoint (CVA), Latent Class or even self-explicated utilities). Each respondent's (or group's) point estimates are sampled multiple times during simulations, with different random variance added each time. The utility of alternatives is computed at each iteration (draw) and choices are assigned applying the First Choice rule. On the surface, this approach resembles using HB draws in simulations. Both

techniques reflect uncertainty (error distributions) about the part-worths and simulate multiple choices per respondent (or group).

In this paper, we compare the use of HB draws and RFC. Our findings show that using HB draws in simulations seems to work well, but applying RFC to individual-level point estimates of part-worths works even better. We also discuss two potential biases when using HB draws in simulations: a *reverse number of levels effect*, and an *excluded level effect*. These biases may in part explain why simulating using draws was not as successful as applying RFC to point estimates for our data set.

RANDOM EFFECTS THEORY

Before we continue, we should review the basic random effects model. The random effects model (McFadden 1973) takes the following form:

$$U_i = X_i (\beta) + \varepsilon_i$$

where U_i = utility of alternative i
 X_i = row vector of independent variables (attribute level codes) associated with alternative i
 β = vector of part-worths
 ε_i = an error term

In fact, if ε_i is distributed as *Gumbel and the First Choice rule is applied in simulations, the expectation (given a very large number of draws) is identical to the logit simulation model. Adding larger error variance to the utility of each alternative is equivalent to applying a smaller “scale factor” and share predictions are “flattened.” Adding less error makes the share predictions “steeper.” Adding zero error to the utility of each alternative is equivalent to the First Choice model. During the remainder of this paper, the reader should keep in mind this inverse relationship between error and the resulting scale of the predicted shares.

The next three sections are an introduction to Randomized First Choice and are taken (with a few minor modifications and additions) from the CBC v2.0 manual (Sawtooth Software, 1999). Those familiar with RFC may choose to skip these sections.

PREVIOUS SIMULATION METHODS

The First Choice model (maximum utility rule) has a long history in conjoint analysis for simulating shares of preference among competitive product concepts. It is intuitively easy to understand and is immune from IIA (Red Bus/Blue Bus) difficulties. However, it also often does not work very well in practice. The share estimates usually tend to be too “steep” relative to shares in the real world. The standard errors of the simulated shares are much greater than with logit (Share of Preference) simulations, since product preference is applied as an “all or nothing” 0/1 rule. Moreover, the notion that a respondent who is “on the fence” with respect to two alternatives will make a sure choice is simplistic and counter-factual (Huber *et al.* 1999).

* The Gumbel (extreme value) distribution is a double negative exponential distribution, drawn by taking ($Y = -\ln(-\ln x)$), where x is a rectangularly distributed random variable ($0 < x < 1$).

Main Point #1: First Choice share predictions are usually too extreme and are not tunable. But, they avoid IIA problems.

The Share of Preference (logit) simulation model offers a way to tune the resulting shares to the desired scaling. It also captures relative information about the value of *all* product alternatives rather than just the best one, thereby increasing the precision of the simulated shares. However, the model is subject to IIA (Red-Bus/Blue-Bus) problems. Within the unit of analysis, cross-elasticities and substitution rates among products are assumed to be constant. This drawback can be quite damaging—especially for aggregate models (i.e. aggregate logit or Latent Class).

Main Point #2: Share of Preference share predictions can be tuned and have greater precision than First Choice shares. But, they have IIA problems.

RANDOMIZED FIRST CHOICE

The Randomized First Choice (RFC) method combines many of the desirable elements of the First Choice and Share of Preference models. As the name implies, the method is based on the First Choice rule, and helps significantly resolve IIA difficulties. As with the Share of Preference model, the overall scaling (flatness or steepness) of the shares can be tuned.

Most of the theory and mathematics behind the RFC model are nothing new. However, to the best of our knowledge, those principles had never been synthesized into a generalized conjoint/choice market simulation model. RFC, suggested by Orme (Orme 1998) and later refined by Huber, Orme and Miller (Huber *et al.* 1999), was shown to outperform all other Sawtooth Software simulation models in predicting holdout choice shares for a data set they examined. The holdout choice sets for that study were designed specifically to include identical or near-identical alternatives.

Rather than use the part-worths as point estimates of preference, RFC recognizes that there is some degree of error around these points. The RFC model adds unique random error (variation) to the part-worths and computes shares of preference using the First Choice rule. Each respondent is sampled many times to stabilize the share estimates. The RFC model results in an automatic correction for product similarity due to correlated sums of errors among product alternatives defined on many of the same attributes. To illustrate RFC and how correlated errors added to product utilities can adjust for product similarity, consider the following example:

Assume two products: A and B. Further assume that A and B are unique. Consider the following product utilities for a given respondent:

	<u>Avg. Product Utilities</u>
A	10
B	30

If we conduct a First Choice simulation, product B captures 100% of the share:

	<u>Avg. Product Utilities</u>	<u>Share of Choice</u>
A	10	0%
B	30	100%

However, let's assume that random forces come to bear on the decision for this respondent. Perhaps he is in a hurry one day and doesn't take the time to make the decision that optimizes his utility. Or, perhaps product B is temporarily out-of-stock. Many random factors in the real world can keep our respondent from always choosing B.

We can simulate those random forces by adding random values to A and B. If we choose large enough random numbers so that it becomes possible for the utility of A sometimes to exceed the utility of B, and simulate this respondent's choice a great many times (choosing new random numbers for each choice replication), we might observe a distribution of choices as follows:

	<u>Avg. Product Utilities</u>	<u>Share of Choice</u>
A	10	25.00%
B	30	75.00%

(Note: the simulation results in this section are for illustration, to provide an intuitive example of RFC modeling. For this purpose, we assume shares of preference are proportional to product utilities.)

Next, assume that we add a new product to the mix (A'), identical in every way to A. We again add random variation to the product utilities so that it is possible for A and A' to be sometimes chosen over B, given repeated simulations of product choice for our given respondent. We might observe shares of preference for the three-product scenario as follows:

	<u>Avg. Product Utilities</u>	<u>Share of Choice</u>
A	10	20.0%
A'	10	20.0% (A + A' = 40.0%)
B	30	60.0%

Because unique (uncorrelated) random values are added to each product, A and A' have a much greater chance of being preferred to B than either one alone would have had. (When a low random error value is added to A, A' often compensates with a high random error value). As a simple analogy, you are more likely to win the lottery with two tickets than with one.

Given what we know about consumer behavior, it doesn't make sense that A alone captures 25.0% of the market, but that adding an identical product to the competitive scenario should increase the net share for A and A' from 25.0% to 40.0% (the classic Red Bus/Blue Bus problem). It doesn't seem right that the identical products A and A' should compete as strongly with one another as with B.

If, rather than adding uncorrelated random error to A and A' within each choice replication, we add the same (correlated) error term to both A and A', but add a unique (uncorrelated) error term to B, the shares computed under the First Choice rule would be as follows:

	<u>Avg. Product Utilities</u>	<u>Share of Choice</u>
A	10	12.5%
A'	10	12.5% (A + A' = 25.0%)
B	30	75.0%

(We have randomly broken the ties between A and A' when accumulating shares of choice). Since the same random value is added to both A and A' in each repeated simulation of purchase

choice, A and A' have less opportunity of being chosen over B as compared to the previous case when each received a unique error component (i.e. one lottery ticket vs. two). The final utility (utility estimate plus error) for A and A' is always identical within each repeated First Choice simulation, and the inclusion of an identical copy of A therefore has no impact on the simulation result. The correlated error terms added to the product utilities have resulted in a correction for product similarity.

Let's assume that each of the products in this example was described by five attributes. Consider two new products (C and C') that are not identical, but are very similar—defined in the same way on four out of five attributes. If we add random variation to the part-worths (at the attribute level), four-fifths of the accumulated error between C and C' is the same, and only one-fifth is unique. Those two products in an RFC simulation model would compete very strongly against one another relative to other less similar products included in the same simulation. When C received a particularly large positive error term added to its utility, chances are very good that C' would also have received a large positive error term (since four-fifths of the error is identical) and large overall utility.

RFC MODEL DEFINED

We can add random variation at both the attribute *and* product level to simulate any similarity correction between the IIA model and a model that splits shares for identical products:

$$U_i = X_i (\beta + E_a) + E_p$$

where:

- U_i = Utility of alternative i for an individual or homogenous segment at a moment in time
- X_i = Row of design matrix associated with product i
- β = Vector of part-worths
- E_a = Variability added to the part-worths (same for all products in the set)
- E_p = Variability (i.i.d Gumbel) added to product i (unique for each product in the set)

Repeated draws are made to achieve stability in share estimates, computed under the First Choice rule. We used E_a error distributed as Gumbel, but a normal distribution could be used as well.

(Note that when the attribute variability is zero, the equation above is identical to the random effects model presented earlier, which is identical to the logit rule.)

In RFC, the more variation added to the part-worths, the flatter the simulations become. The less variation added to part-worths, the more steep the simulations become. Under every possible amount of attribute variability (and no product variability), net shares are split in half for identical products, resulting in no “inflation” of net share. However, there may be many market scenarios in which some share inflation is justified for similar products. A second unique variation term (E_p , distributed as Gumbel) added to each product utility sum can tune the amount of share inflation, and also has an impact on the flatness or steepness of the overall share results. It can be shown that adding only product variability (distributed as Gumbel) within the RFC

model is identical to the familiar logit model (Share of Preference Model). Therefore, any degree of scaling or pattern of correction for product similarity ranging between the First Choice model and Share of Preference can be specified with an RFC model by tuning the *relative* contribution of the attribute and product variation.

The obvious question for the researcher is how much share inflation/correction for product similarity is justified for any given modeling situation. To answer this, holdout choice tasks that include some alternatives that are very similar alongside others that are quite unique should be included in the study and used for tuning the RFC model.

USING HB DRAWS IN SIMULATIONS

As introduced earlier, hierarchical Bayes can estimate individual-level part-worth utilities for choice or conjoint experiments. HB is a computationally-intensive iterative process. After a period of “burn-in” iterations, convergence is assumed and the results of subsequent iterations are saved. One usually saves many iterations (draws) for each respondent. Point estimates of part-worths are computed for each individual as the average of the saved draws.

The analyst has the choice of using the point estimates or the multiple draws per respondent in market simulations. Indeed, using HB draws rather than point estimates has a great deal in common with RFC. The draws reflect error distributions around the average parameters for each individual. However, the variances and covariances of the parameters in HB are empirically estimated. In contrast, RFC assumes (within each respondent or unit of analysis) that the variances of the part-worths are equal and the covariances are zero.

Main Point #3: Two simulation methods show promise for reducing the IIA problem while at the same time being tunable: using HB draws and RFC.

Before comparing HB draws and RFC, we were faced with the question of how many draws to use for HB simulations and how many sampling replications to employ with RFC. The answer would affect the computational time required to perform the simulations for this paper. More importantly, we recognize that researchers face the same decision when analyzing real-world studies—and they are usually under greater time constraints than we were.

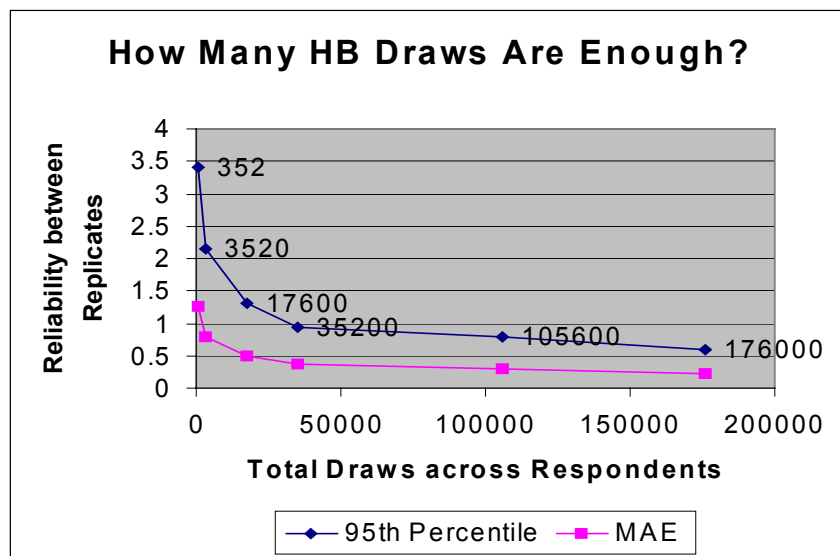
As background, the data set we used was collected by Huber, Orme and Miller in 1997. Shoppers were intercepted at shopping malls and asked to participate in a CBC study dealing with television sets. Three-hundred fifty-two respondents saw 18 randomly designed choice tasks followed by nine fixed holdout choice tasks. The design had six total attributes with 17 total levels. The holdouts were very carefully designed to have near utility balance. Also, each holdout choice task included five product concepts, two of which were either identical, or nearly identical. There were four versions of the fixed holdout choice tasks resulting in (4 versions)(9 tasks)(5 concepts per task) = 180 product concepts for which we could compare actual versus simulated shares of preference.

We ran HB for 35,000 burn-in iterations, then saved 1000 draws (skipping every tenth draw) per respondent, for a total of (352 respondents)(1000 draws) = 352,000 sets of 17 part-worths, resulting in a 57 Megabyte data file. To test the stability of simulated shares of preference given different numbers of draws per respondent, we simulated shares (First Choice rule) for the 180 holdout products using only the first draw for 352 respondents compared to the shares computed using only the 1000th draw. The difference was measured in terms of MAE (Mean Absolute

Error). We repeated the analysis multiple times to stabilize the results (draw #2 versus draw #999, draw #3 versus draw #998, etc). We also sorted the MAE figures from worst (largest) to best (smallest) and recorded the 95th percentile MAE. We repeated that same analysis for 10 draws at a time per respondent, 50, 100, 300 and 500. The results are displayed in the table and graph below. For example, when using just one draw per respondent in simulations, the simulated shares for holdout concepts differed on average by 1.25 share points between replicates, but 5% of the shares differed by 3.41 points or more.

Table 1

Draws per Respondent	Total Draws across All Respondents (n=352)	Mean Absolute Error between Replicates	95 th Percentile Mean Absolute Error
1	352	1.25	3.41
10	3,520	0.80	2.16
50	17,600	0.49	1.31
100	35,200	0.38	0.95
300	105,600	0.30	0.78
500	176,000	0.22	0.59



With 35,200 draws, 95% of the shares between replicates differ by less than a share point. We think the data suggest that using more than about 50,000 total draws does not offer *practical* benefit in terms of stability of aggregate share estimates. Until computers become much faster, if the researcher is only interested in aggregate share predictions, we suggest using at least 30,000 but not much more than about 50,000 total draws (across all respondents). Much larger than that and the files become difficult to manage and processing time too much to swallow for “in-the-trenches” research. However, the constraint of file size assumes that the researcher wants to use HB draws during simulations, which we’ll argue may not be the suggested approach.

COMPARING HB DRAWS AND RFC

We compared simulations from HB draws and RFC on point estimates by examining the predicted versus actual holdout shares. MAE (Mean Absolute Error) quantifies the difference between actual and predicted shares. For example, suppose we have three products with actual share percentages of 10, 30, and 60. If the predicted shares are respectively 15, 20, and 65, the MAE is $(|10-15|+|30-20|+|60-65|)/3 = 6.67$. For reporting purposes, we scale the MAE results as a percentage of test/retest reliability. For this study, the test/retest reliability was 3.5. A result of 113% means the predictions were 13% worse than test/retest reliability.

Table 2
HB Draws Performance

Simulation Method	Error Relative to Test/Retest Reliability
First Choice	113%
Share of Preference (tuned)	109%

Table 3
HB Point Estimates Performance

Simulation Method	Error Relative to Test/Retest Reliability
First Choice	116%
Share of Preference (tuned)	110%
RFC (tuned, E _A only)	108%
RFC (tuned, E _A and E _P)	107%

For simulations using HB draws, we used 50,000 total draws and used both First Choice and Share of Preference models (see Table 2). The First Choice model resulted in an MAE of 3.95, 13% less accurate overall than test/retest reliability. The First Choice shares were a bit too extreme. We found that we could improve matters by using a Share of Preference model. The Share of Preference model, after tuning the exponent, resulted in an MAE of 3.80—9% worse than test/retest reliability.

We then turned to simulations using HB point estimates for the First Choice and Share of Preference models (see Table 3). These were 16% and 10% worse (respectively) than the test/retest reliability. These results were slightly worse than the HB Draw results.

How well did the RFC model predict shares using the point estimates? We used 50,000 replications across respondents. Adding only attribute error, we achieved results of 108%. When product error was added and the relative contribution of attribute and product error tuned, we achieved results of 107%. These predictions are slightly better than the simulations using the HB draws.

The most important finding is that using HB draws in simulations is not better than using RFC on point estimates, despite RFC's simplifying assumptions regarding the error distributions. By using RFC with point estimates, one also avoids having to deal with very large draw files.

Main Point #4: RFC has two important advantages: it produces better predictions than using HB draws, and it avoids dealing with enormous data files.

There are two interesting effects that may help explain why using RFC with point estimates is more accurate than using draws in simulations for our data set: a *reverse number of levels* effect and an *excluded levels* effect.

“REVERSE” NUMBER OF LEVELS EFFECT

The Number of Levels Effect (NOL) is well-documented and prevalent in varying degrees in all types of conjoint analysis studies. The NOL effect as described in the literature is as follows: holding the range of variation constant, if an attribute is defined on more rather than fewer levels, it tends to get more importance, as measured by the range of part-worths for that attribute.

Many researchers have tried to prove an algorithmic explanation to the NOL effect using synthetic, computer-generated data. With the exception of rankings-based card-sort conjoint or pairwise matrices, we are unaware of a previous synthetic data set that has successfully demonstrated a NOL effect. Our research finds when using HB draws a consistent NOL effect using computer-generated data. But the consistent finding is for a “reverse NOL” effect. We use the term “reverse” because it works in the opposite way that we are used to thinking about the NOL effect. Rather than attributes with more levels being biased toward more importance (*ceteris paribus*), those attributes with more levels have *less* importance.

Most treatments of the NOL effect have focused on a measure of importance equal to the range in part-worths for each attribute divided by the sum of the ranges of the part-worths across all attributes. In contrast, this research focuses on the impact an attribute has on what most practitioners use conjoint data for: simulated shares of preference.

HB results in multiple replicates (draws) for each respondent. From those draws, one can estimate within-respondent variances for individual part-worths. It was while studying these variances that we noticed that attributes with more levels tended to have larger variances around their part-worths and attributes with fewer levels had smaller variances. To illustrate this, we generated a synthetic CBC data set with 6 attributes with known utilities ranging from 1 to 0 within each attribute (equal importances). Half of the attributes had 2 levels (part-worths of 1, 0) and the other half had 4 levels (part-worths of 1, 0.66, 0.33, 0). Three-hundred simulated respondents completed 20 tasks with 3 concepts each. Random heterogeneity (between respondents) was added to the part-worths. Five-hundred draws were saved for each respondent. The average part-worths and error variances are presented in Table 4:

Table 4

Part-Worths and Variances
for Synthetic Data Set 1

Attribute	Level#	Avg. Part- Worth	Avg. Within- Person Variance
1	1	2.44	2.86
	2	0.94	3.11
	3	-1.03	3.63
	4	-2.35	3.82
2	1	2.64	1.45
	2	-2.64	1.45
3	1	2.69	3.00
	2	1.23	3.00
	3	-1.12	3.35
	4	-2.79	3.69
4	1	2.97	1.49
	2	-2.97	1.49
5	1	3.16	3.06
	2	0.57	2.69
	3	-0.95	3.38
	4	-2.78	3.33
6	1	2.53	1.44
	2	-2.53	1.44

Importance of attributes 1+3+5: 49.9%
Importance of attributes 2+4+6: 50.1%

Even though the importance of the 4-level attributes (as computed from the aggregate part-worths above) was the same as the 2-level attributes (within 2/10 of a percentage point), the within-respondent variance is much greater around the part-worth estimates than for the 2-level attributes.

If the variances of part-worths are influenced by the number of levels, and the variance of part-worths is directly related to the “scale” of the predictive market choice model, it follows that these differences might lead to systematic biases for simulated shares of preference. To test this hypothesis, we conducted sensitivity simulations.

Shares of choice (First Choice rule) were simulated for each of the (500 draws)(300 respondents), for a total of 150,000 cases. Our approach was one of sensitivity analysis, to test the maximum impact of each attribute on choice versus a constant alternative (with utility of 0, representing an average desirability). For example, starting with attribute one, one enters a product concept made up of levels 1 through 4 (holding all other attributes constant) in four separate simulation steps.

We'll define the "simulated importance" of an attribute as the maximum range of share impact from sensitivity simulations. For example, the shares of choice for attribute one (versus a constant alternative) at each of its four levels was: 0.80, 0.61, 0.36 and 0.22, for a simulated importance of $0.80 - 0.22 = 0.58$.

The simulated importances for all attributes are shown in Table 5:

Table 5

Simulated Importances for
Synthetic Data Set 1

Attribute	#levels	Simulated Importance
1	4	0.58
2	2	0.74
3	4	0.64
4	2	0.82
5	4	0.68
6	2	0.75

Avg. Simulated Importance for 4-level Attributes: 0.63
Avg. Simulated Importance for 2-level Attributes: 0.77

This example demonstrates a consistent NOL effect. The two-level attributes on average have 22% *more* impact in simulations than the four level attributes.

WHAT CAUSES THE "REVERSE" NOL EFFECT?

The variances among HB estimates reflect our uncertainty about parameter values. Consider a balanced design that has some attributes with two levels and others with four. The ratio of observations to parameters to be estimated (within each attribute) is much greater for attributes with fewer levels relative to attributes with more. For the attributes with two levels, its levels occur twice as often within the design relative to attributes with four levels. Therefore, there is more information available to stabilize the part-worths for the two-level attributes.

WHICH IS BIGGER: "REVERSE" OR "REGULAR" NOL EFFECT?

The practical question is how big are these effects? Can conducting simulations with HB draws significantly reverse the negative consequences of the usual NOL effect? If it did, would that be a prudent action?

Dick Wittink is the most-published researcher with regard to NOL. To paraphrase his findings over the years, NOL is a significant problem for traditional full-profile conjoint and choice-based conjoint and much less of a problem for ACA.

In the 1997 Sawtooth Software proceedings, Wittink published findings for an experimental study among human respondents rather than "computer" respondents (Wittink 1997). A split-sample study was conducted in which the range of variation was held constant for attributes, but some respondents saw more levels than others. Two cells in the design (B and C) had the

following numbers of levels per attribute (four in total) and resulting importances (as computed by examining the ranges of the aggregate part-worths):

Table 6

(B) Number of Levels	Importance	(C) Number of Levels	Importance
3	25%	2	16%
3	16%	4	24%
3	32%	2	29%
3	28%	4	31%

The percentage gain in importance when increasing the number of levels from 3 to 4 is $1 - [(24+31)/(16+28)] = +25\%$. The net loss in importance when decreasing the number of levels from 3 to 2 is $1 - [(16+29)/(25+32)] = -21\%$. The relative gain for 4-level attributes relative to 2-level attributes is then $(1+0.25) / (1-0.21) = +58\%$.

In the 1999 Sawtooth Software Proceedings, Wittink reported findings from a study by (Shifferstein *et al.* 1998) for another full-profile experiment. Again, human respondents were randomly divided into different groups receiving different versions of the conjoint design. Between cells A and B, the range of variation was held constant, but the number of levels used to describe the attributes differed.

Table 7

(A) Number of Levels	Importance	(B) Number of Levels	Importance
4	0.21	2	0.13
2	0.17	4	0.26
2	0.07	2	0.07

For attribute 1 (the first row), increasing the number of levels from 2 to 4 resulted in a $1 - (0.21/0.13) = 62\%$ increase in importance. For attribute two, the same doubling in levels resulted in a 53% increase in importance. We should note, however, that the changes in importance for these two attributes are not independent. From version A to version B, losses in importance (by reducing levels from 4 to 2) for attribute 1 are enhanced by gains in importance (by increasing levels from 2 to 4) for attribute 2. So the net gain in importance (*ceteris paribus*) one should expect from doubling the number of attributes from 2 to 4 for this data set is something less than either 53% or 62%.

Summarizing the findings from these two full-profile studies (and taking some liberties with the data), doubling the number of levels from 2 to 4 levels (but holding the range of variation for an attribute constant) results in roughly a 50% artificial increase in importance (measured by examining ranges of part-worths).

We noted above that using HB draws in simulations resulted in a reverse NOL effect of about 22% for 4-level relative to 2-level attributes. Again, applying loose math, the “usual” NOL effect

is about 2 to 3 times as strong as the reverse NOL effect detected earlier. We might conclude that if we simulated results using HB draws for the two full-profile data sets above, we could cut the NOL effect by about one-third to one-half.

“REVERSE” NOL EFFECT: GOOD NEWS OR BAD NEWS?

If the analyst accepts the usual NOL effect as bad news, anything that counters that effect should be considered good. A counter argument (Wittink 1999b) states that *if* the psychological explanation to NOL holds, there is also likely a NOL effect in the real world for attributes that naturally have different numbers of levels to define available products. This would suggest that our designs should reflect the natural number of level differences and our results should reflect a NOL effect in the part-worths consistent with real world preferences. If methods such as the use of HB draws in simulations consistently reduce that true effect, this then would *not* be a welcomed outcome.

Indeed the argument surrounding NOL is complex. In any case, we aren't comfortable with the temptation to simulate shares from draws to reduce a NOL bias that results from an unbalanced design. Two wrongs don't make a right. We'd prefer to see researchers take appropriate steps to minimize the NOL problem and then simulate shares using RFC and point estimates. Using RFC with point estimates does not reflect the strong reverse NOL bias displayed by HB draws simulations.

“EXCLUDED LEVEL” EFFECT

After examining a number of HB runs from artificial data sets with known utilities, we noticed that the variance of the last level of each attribute tended to be greater than the variance of the other levels. The reader may notice that the variances for the attribute part-worths presented in Table 4 hint at this effect. It turns out that the greater the number of levels, the more pronounced the difference in variances becomes.

We generated a synthetic data set with two attributes at 12 levels each, with known utilities of zero for all levels (random responses).

Table 8

Attribute	Level#	Within-Respondent Variance
1	1	0.383
	2	0.364
	3	0.417
	4	0.416
	5	0.470
	6	0.385
	7	0.404
	8	0.359
	9	0.359
	10	0.374
	11	0.404
	12	0.856
2	1	0.407
	2	0.371
	3	0.350
	4	0.341
	5	0.430
	6	0.309
	7	0.372
	8	0.427
	9	0.327
	10	0.428
	11	0.315
	12	0.848

Though the expected variances should be equal, the variance of the twelfth level of each attribute is more than double the size of the other levels. Recall that the more variance added to the utility for product concepts, the “flatter” the share of preference in simulations. Therefore, the last levels of each attribute bias the shares for products in which they are included, making them tend toward 50%.

This *excluded level effect* is an artifact resulting from the effects-coding procedure used in coding the independent variable matrix. Effects-coding constrains part-worths to be zero-centered. The last level is “excluded” from the design and solved later as negative the sum of the effects of the other levels within the same attribute.

We are indebted to Rich Johnson for providing this explanation regarding why the excluded level has a much higher variance:

“The interesting curiosity is that the final (excluded) level has variance much greater than that of the other (included) levels, and the discrepancy increases as the number of levels in the attribute. Of course the reason for this is that the variance of a sum is equal to the sum of the variances and covariances. If the covariances among levels were zero, then the variance of the excluded level would be $(n - 1)$ times as large as the included levels, where n is the number of attributes. Since the covariances are for the most part negative, the actual effect is smaller than that, but still sizeable.

“One doesn’t see that effect with logit or other aggregate methods because in that case the expansion is done on point estimates, which have small variances. But when we do it on individual draws, the effect looms large.

“As one might expect, a similar but opposite thing occurs with dummy-variable coding. In that case the excluded level is estimated by taking the negative of the mean of the remaining levels, so one would expect its variance to be smaller. That turns out to be the case. There appears to be no way around this problem, which may limit the usefulness of HB draws in first choice simulations.”

The *excluded level effect* does not exist for two-level attributes, and is very minor for attributes with only a few more levels than that. For attributes with many levels, it could conceivably lead to significant biases in market simulations when using the HB draws.

Main Point #5: Two reasons why RFC may work better are that HB draws have a reverse NOL effect and an excluded level effect.

SUMMARY AND CONCLUSIONS

We have reviewed the two most widely used methods for simulating choices from conjoint or choice part-worths, namely the First Choice and Share of Preference (logit) rules. The First Choice model is immune to IIA difficulties, but is often too steep and is not tunable. The logit rule is tunable, but suffers from IIA. Randomized First Choice (RFC) combines benefits of both models and can improve the predictive accuracy of market simulations.

Like RFC simulations, hierarchical Bayes draws reflect uncertainty about the point estimates for part-worths. But, within the unit of analysis, RFC assumes a covariance matrix for part-worths with equal variances along the diagonal and zeroes in the off-diagonal elements. HB makes neither of these assumptions: the variances of the part-worths can differ, and covariances are not assumed to be zero. We compared the predictive accuracy of RFC simulations on point estimates versus conducting simulations using HB draws. The RFC simulations were slightly more accurate for our data set, and they avoided having to use the huge draw files.

We also demonstrated that using HB draws in simulations is subject to two biases: a *reverse number of levels* effect, and an *excluded level* effect. These biases have the potential to significantly degrade the predictive accuracy of market simulations.

A number of sophisticated approaches have been suggested for circumventing IIA and improving the predictive validity of market simulations. These techniques have included Mother Logit, Multinomial Probit and Nested Logit. We have not attempted to test or expound on those techniques here. In our opinion, for “in-the-trenches” practical research, a well-tuned RFC model operating on well-developed individual-level point estimates from HB estimation is hard to beat.

REFERENCES

- Huber, Joel, Orme, Bryan K. and Richard Miller (1999), "Dealing with Product Similarity in Conjoint Simulations," Sawtooth Software Conference Proceedings, pp 253-66.
- McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.) *Frontiers in Econometrics*. New York: Academic Press.
- Orme, Bryan (1998), "Reducing the IIA Problem with a Randomized First Choice Model," Working Paper, Sawtooth Software, Sequim, WA.
- Sawtooth Software (1999), "The CBC System, v2.0," Sequim, WA.
- Sawtooth Software (1999), "The CBC/HB Technical Paper," <http://www.sawtoothsoftware.com/TechPap.htm>.
- Shifferstein, Hendrik N. J., Peter W. J. Verlegh and Dick R. Wittink (1998), "Range and Number-of-Levels Effects in Derived and Stated Attribute Importances," Working Paper.
- Wittink, Dick R. (1997), "Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis," Sawtooth Software Conference Proceedings, pp 227-40.
- Wittink, Dick R. (1999a), "A Comparison of Alternative Solutions to the Number-of-Levels Effect," Sawtooth Software Conference Proceedings, pp 269-84.
- Wittink, Dick R. (1999b) "Comment on McCullough," Sawtooth Software Conference Proceedings, p 117.