PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

August 1997

1997 Sawtooth Software Conference Proceedings: Sequim, WA.

FOREWORD

It is our pleasure to present the Proceedings of the Sixth Sawtooth Software Conference, held in Seattle, Washington in August, 1997. It is hard to believe that it has been five years since our last conference! The past five years have seen a flood of important developments in interviewing technology and in marketing research. The overall quality of the papers in this volume seem to reflect both a seasoning of ideas and accumulation of advancements.

The focus of the conference was quantitative methods in marketing research. The presenters provided insights into the latest developments in computer interviewing, conjoint, choice, market segmentation, statistical modeling and perceptual mapping. Authors were charged to deliver presentations of value to both the most and least sophisticated members of the audience. We believe both the oral presentations and accompanying papers reflected this focus.

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

The papers and discussant comments are in the words of the authors, and very little copy editing was performed. We are in debt to the authors and discussants for making this conference a success and for advancing our collective knowledge in this exciting field.

Sawtooth Software
November, 1997

CONTENTS

SUMMARY OF FINDINGS

We've distilled some of the key points and findings from each presentation below. We apologize to the authors if our summaries do not reflect what they believe to be the high points of their presentations.

Overcoming the Problems of Special Interviews on Sensitive Topics: Computer Assisted Self-Interviewing Tailored for Young Children and Adolescents (Edith de Leeuw, Joop Hox, Sabina Kef, Marion Van Hattum): The authors presented results from two studies: the first examined bullying in elementary schools, the second surveyed blind adolescents and young adults. Key findings were:

- Respondents were more likely to share sensitive information under CASI.
- CASI resulted in fewer missing values and tighter standard deviations than paper.
- Counting all costs, CASI was significantly less expensive than the paper-based implementation.
- Interviewers and respondents alike were generally pleased and comfortable with computerized interviews.

Best Practices in Interviewing Via the Internet (Karlan Witt): The rapid growth of the Internet has opened up faster, less costly ways of collecting data. "The Internet brings with it a host of unique limitations that impact any research effort in this area," Karlan explained. According to Karlan, the incidence of Internet access in the U.S. stands at 16% as of Q2 1996. Karlan reported the relative incidence of browsers for Q4 1996: Microsoft Internet Explorer 6%, Netscape Navigator 37%, AOL browser 10%.

Karlan conveyed a lot of advice regarding the use of this new medium, which we unfortunately can't cover for space limitations. Two important points were:

- Ensure that potential respondents have access to the Internet, and be comfortable in navigating to the desired web site and using their browser to take the survey.
- Internet surveys must be tested under different platforms and browsers to ensure that the survey performs properly.

Karlan predicted that low barriers of entry will cause overuse of the Internet for conducting surveys. She warned that "Overuse and general abuse will likely lead to a backlash of potential respondents, similar to that currently seen in the telephone arena."

An Alternative Approach to Brand Price Trade-Off (Ray Poynter): While CBC is considered the tool of choice for pricing research in the U.S., Europe is still fond of the BPTO method. The traditional BPTO method focuses on just two attributes: brand and price. Respondents choose from a set of concepts (cards) with all brands starting at the lowest price. When a card is chosen, it is replaced by the same brand at a slightly higher price, while the non-chosen brands remain the same for the next task. Traditional BPTO has been faulted for encouraging patterned and unrealistic behavior.

Ray showed creative ways to break the patterned behavior by:

- Using more realistic starting prices for the first task,
- Increasing the price for the chosen concept and simultaneously reducing the price for the non-chosen items,
- Randomly removing brands from each choice set.

Ray programmed the improved BPTO task in Ci3, but comments that it takes a good Ci3 programmer. Ray described how to calculate PEP (Purchase Equilibrium Prices): dollar amounts that make a respondent indifferent between two brands, and how to incorporate this information for use in a first-choice simulator.

Creating End-User Value with Multi-Media Interviewing Systems (Dirk Huisman): Dirk showed examples of how multi-media technology can enhance the realism of surveys. Making interviews better reflect the real world may result in better data. Dirk reported the results of a split sample ACA interview, where some of the attributes were shown in multi-media. Interestingly enough, he found little difference between the utilities calculated from the text-based ACA versus the multi-media ACA.

A Comparison of Full- and Partial-Profile Best/Worst Conjoint Analysis (Keith Chrzan, Ritha Fellerman): Best-Worst is a questioning technique that displays a product described on multiple attributes and asks respondents to identify the features that make them most and least want to purchase the product. "The most unique strength of best/worst conjoint analysis," the authors stated, ". . . is that it eliminates the arbitrariness of the scale origins of the individual attributes."

The authors presented results from a study comparing full- and partial-profile best/worst experiments. They noted that full and partial profile best/worst models may result in different estimates of preference structure and concluded: "Apparently there is something specific to best/worst measurement that makes it not work with partial profiles."

Efficient Experimental Designs Using Computerized Searches (Warren Kuhfeld): Warren introduced the concept of design efficiency and argued that orthogonality in conjoint experiments is less necessary today. Orthogonality was important in days when computers were not widely available. If an orthogonal design was used, relatively simple formulas were available for hand or calculator ANOVA computations. Today, general linear models such as OLS do not require orthogonality for the unbiased estimation of effects.

Warren explained the principles of orthogonality and balance, introduced the measure of D-efficiency, and compared two computerized search routines for finding efficient experimental designs: SAS's PROC OPTEX procedure, and Sawtooth Software's CVA designer. For the size of designs commonly used in conjoint experiments, Warren found the CVA routine to find designs about 97% as efficient as OPTEX, but that CVA's designs tended to be more balanced. He also found CVA easier to use than OPTEX. Warren concluded: "For small problems like you would typically encounter in a full-profile conjoint study, CVA seems to do an excellent job. However, for larger and more difficult problems, it often fails to find more efficient designs that can be found with PROC OPTEX."

Practical Ways to Minimize the IIA-bias in Simulation Models (Rainer Paffrath): Many conjoint simulations suffer from IIA problems which can sometimes cause less-than-satisfactory results. Rainer reviewed the oft-cited red-bus/blue-bus example, which demonstrates how nearly-identical products together in a conjoint simulator can lead to net share inflation for like products. He pointed out weaknesses in Model 3 from the ACA, CBC and CVA simulators. Rainer contended that corrections for product similarity should be customized and usable at the individual level, and that each individual's importance structure should be taken into account.

The Number of Choice Alternatives in Discrete Choice Modeling (Jon Pinnell, Sherry Englert): Jon presented results from three choice studies in which the number of concepts (alternatives) was varied within and between respondents. He pointed out that choice tasks with just two alternatives (i.e. A vs. B) would lead to only one inferred inequality (if A is chosen, A>B); whereas a first-choice from a task with six concepts leads to five inequalities (if A is chosen, A>B, A>C, A>D, A>E, A>F). Jon showed that the additional time required to make choices from more complex tasks is comparatively less than the value of additional information gained.

After comparing part-worths from choice sets of 2, 4, and 7 alternatives, Jon concluded: ". . . our findings caution against the use of pairs. Our data show that pairs are processed differently, have lower predictive validity, are less stable, and don't save much time relative to larger tasks."

Extensions to the Analysis of Choice Studies (Tom Pilon): Tom presented some additional types of analysis that can be done using standard CBC data. He reported results from a beer study, and showed how cross-elasticities for brands could be calculated (by regressing the log of choice volume on the log of price) and incorporated into a market simulator.

Tom argued that the standard logit simulator which assumes constant cross-elasticity across brands was not entirely realistic for the beer market. A cross-elasticity simulator lets brands that compete closely (perceived as close substitutes) take relatively more share from one another as a result of price changes than from brands which are not perceived to be as substitutable. Tom also demonstrated how to convert a cross-elasticity matrix into a "brand similarities matrix" for use in an MDS perceptual map. Brands which competed closely with one another were situated close to one another on the map.

Respondents' Behavior in Complex Choice Tasks; A Segmentation-Based and Individual Approach (C.M. (Marco) Hoogerbrugge): Marco spoke of the superiority of individual-level models versus aggregate models. Even though choice modeling has received more attention than traditional conjoint as of late, most choice modeling still is done at the aggregate. Marco compared two methods for segmenting choice data: Latent Class and K-Logit.

Latent Class segments the data based on choices and respondents have a probability of membership in each group. K-Logit is much like cluster analysis in that it finds segments and assigns each respondent to one—and only one—segment. Marco reported that K-Logit is much faster than Latent Class, but that the results are less robust.

Individual Utilities from Choice Data: A New Method (Rich Johnson): Rich presented a new method for calculating individual-level utilities from CBC data. He explained that Latent Class assumes each individual belongs to one group or another, with probabilities of membership summing to 100%. In the past, some researchers have calculated individual utilities by multiplying probabilities of membership by class utilities. Rich graphically demonstrated that probability weighting assumes all respondents lie between the Lclass groups. Such solutions may fit average respondents well, but may improperly represent most cases. By recognizing that individual-level utilities can be calculated using a linear combination of group utilities where weights can be both positive *and* negative, his method captures more heterogeneity and better reflects individuals' positions.

Rich compared results from Monte Carlo simulations and real data sets. His new method performed better than probability weighting in terms of R-squared with known utilities and hit rates for holdout choices. Rich commented that Hierarchical Bayes methods are probably the best overall approach for representing individual utilities from choice, but pointed out that computers are still too slow to make these useful in practice. Rich's method computes much more quickly and can be a practical solution for now.

He concluded, "One of the problems . . . with choice data, is that of predicting the market's response to complex combinations of interactions, differential cross effects, and varying similarities among products. It seems likely that all of these problems will be diminished when modeled at the individual level."

Assessing the Validity of Conjoint Methods—Continued (Bryan Orme, Mark Alpert, Ethan Christensen): Bryan pointed out that despite over 20 years of conjoint research, very little actual evidence has been published about conjoint's ability to predict real world decisions. He suggested that holdout tasks commonly used in survey research may not be realistic, and that they may better gauge respondent consistency than validity.

Bryan presented the results of a pilot study where respondents received both regular holdout choice tasks, and a more intensive 10-minute exercise which he termed the "Super Holdout Task." The attribute importances did not appear to differ between the two types of holdout tasks. Importances for traditional full-profile conjoint and CBC were shown to be more extreme than ACA. Bryan speculated that the Super Holdout Task might not have been realistic enough to accurately reflect the real world, and challenged the attendees to publish validation studies with actual purchase data.

Solving the Number-of-Attribute-Levels Problem in Conjoint Analysis (Dick Wittink, Bill McLauchlan, P.B. Seetharaman): Dick is credited as one of the first to document the number of attribute levels effect in conjoint, and is probably the leading expert on that topic. In his presentation, he demonstrated that researchers can dramatically increase the importance that attributes receive by simply increasing the number of levels on which attributes are described, and provided quantitative evidence on how dramatic this effect can be. He commented that ACA is less susceptible to the number of levels effect than full-profile. Dick argued that the source of the effect in ACA can be largely attributed to two factors:

- the lack of perfect utility balance in the pairs design,
- the propensity of respondents to answer toward the middle of the scale even though the predicted response would be more extreme.

When respondents "split the difference" between the predicted value and the midpoint in paired comparison ratings, it tends to increase the importance of the attribute defined on more levels.

A customized version of ACA was developed to achieve better utility balance by expanding the number of levels for more important attributes. A split-sample study was conducted using ACA vs. the customized version. Holdout hit rates were higher for the customized version.

What We Have Learned from 20 Years of Conjoint Research: When to Use Self-Explicated, Graded Pairs, Full Profiles or Choice Experiments (Joel Huber): Joel pointed out that respondents adopt different strategies for answering different types of conjoint questions. Researchers should understand these simplification strategies and match the right method to the context of actual marketplace decisions. He summarized the strengths of the methods as follows:

- ***Self-explicated models*** are best in the case of many attributes, where expectations about levels and associations among attributes are stable. They work better in predicting decisions about independent alternatives than for competitive contexts.
- ***Paired comparisons*** are most appropriate for modeling markets in which alternatives are explicitly compared with one another, approximating a deeper search of a broad range of attributes, and where within-attribute value steps are smooth and approximately linear.
- ***Full-Profile*** works best when it is desirable to abstract from short run beliefs, when market choices reflect simplification toward the most important variables, and the decision focus is more within alternative rather than explicitly made using side-by-side comparisons between options.
- ***Choice*** is most appropriate for simulating immediate response to competitive offerings, when decisions are made based on relatively few attributes with substantial aversion to the worst levels of each attribute, and when consumers make decisions based on comparative differences among attributes.

In contrast to what is becoming popular agreement regarding the superiority of choices, Joel cautioned that choices may not always work better than more traditional approaches.

Current Practices in Perceptual Mapping (Tom Wittenschlager, John Fiedler): John rehearsed why he prefers using discriminant analysis (DA) based perceptual mapping versus Correspondence Analysis. Among many reasons, he lists:

- There are more interpretable relationships between attribute vectors and product points than Correspondence Analysis.
- DA is more efficient at cramming a lot of information into a low dimensional space.

John showed a perceptual map he created, and the various refinements it underwent to reflect the market in the most meaningful way for his client. He argued that the APM software package is elegant in its approach, but that the software is outdated. He provided SPSS code and steps for creating DA maps using APM's method.

John recommended that respondents not just rate brands on the stated most important attributes. "Restricting ratings to 'most important' attributes may overlook attributes critical to marketplace differentiation," he argued. To maximize the value of each respondent's contribution toward a meaningful and discriminating map, John recommended that each respondent rate more products at the expense of attributes. He maintained that "It is a waste to have a respondent rate only one or two brands on dozens of attributes when he or she could rate five or six brands on seven or eight attributes."

Obtaining Product-Market Maps from Preference Data (Terry Elrod): Terry applied a different mapping technique to the same data set used by John Fiedler. John's map had been based entirely on ratings of brands on attributes, whereas Terry's map was based entirely on brand preferences. Terry's technique is a maximum likelihood method that assumes a continuous distribution of individual preferences, and finds the brand locations and the preference distribution that together best fit the data. Terry noted that his map appeared to be similar to John's, but that John had required several re-computations to incorporate client reactions, in contrast to his which was based on the data alone. Terry's method is so computationally intensive that it was not possible until recently, but may become a more useful approach as computer speeds continue to improve.

Integrated Choice Likelihood (ICL) Model (Carl T. Finkbeiner): Carl noted that several different kinds of data are useful in studying respondent preferences although current methods usually employ only one type of data at a time. However, there may be benefit in being able to combine data of several types to estimate part worths for each respondent. Carl described an "Integrated Choice Likelihood Model" which does this by integrating self-explicated ratings of attributes, full-(or partial-) profile conjoint choice likelihood ratings, and choice or constant sum ratings. The model can also be used to estimate choice probabilities for new products, and is not subject to IIA difficulties.

Neural Networks and Statistical Models (Tony Babinec): Tony discussed neural networks, stating that he preferred to regard them as a flexible form of regression or discriminant analysis rather than "simulated biological intelligence." He observed that "In reality, as in conventional statistical modeling, one must invest a lot of 'sweat equity' and think through one's problem when applying neural nets." He recommended their use when:

- The functional form relating input variables to the response variable is not known or well understood, but is not thought to be linear.
- There is a large sample of data.
- A premium exists for better prediction that makes it worth the added effort to fit a well-tuned neural network.

OVERCOMING THE PROBLEMS OF SPECIAL INTERVIEWS ON SENSITIVE TOPICS: COMPUTER ASSISTED SELF-INTERVIEWING TAILORED FOR YOUNG CHILDREN AND ADOLESCENTS

Edith De Leeuw, Joop Hox,
Sabina Kef and Marion Van Hattum [1]
Department of Education
University of Amsterdam

ABSTRACT

Self-administered questionnaires have many advantages, especially when sensitive questions are asked. However, paper self-administered questionnaires have a serious draw-back: only relatively simple questionnaires can be used. Computer Assisted Self-Interviewing (CASI) can overcome these problems, and make it possible to use very complex self-administered questionnaires.

CASI can take several forms, for instance, it can be a part of a personal (CAPI) interview where the interviewer hands over the computer to the respondent for specific questions. Another form is a computerized version of the mail survey: Disk-by-Mail. We have used both forms in an application for very special populations. In the first study we implemented a Disk-by-Mail survey on bullying in primary schools; the respondents were 6428 pupils aged 8-12 years. The second study was a survey on personal networks, dating, and well-being of adolescents and young adults with a visual handicap (aged 14-24). This study was a mixed-mode CAPI and CASI survey.

This paper presents a literature review of data quality in CASI-surveys, describes the general logistic of both surveys and the special adaptations we had to make, and presents empirical findings on data quality and general recommendations for the adaptation of computer assisted (self) interviewing for special populations.

Key words: sensitive questions, special groups, disk by mail, CAPI, self-interviewing

1. INTRODUCTION

Self-administered questionnaires have many advantages, especially when sensitive questions are asked. Self-administered procedures evoke a greater sense of privacy, which leads to more self-disclosure (Sudman & Bradburn, 1974; Tourangeau & Smith, 1996). Empirical research has shown that self-administered questionnaires when compared to interviews, produce more valid reports of sensitive behavior and less social desirable answers in general (for a comprehensive review see De Leeuw, 1992).

[1] Authors are listed in alphabetical order

Furthermore, in self-administered procedures the respondent is the locus of control, who determines the pacing of the question-answer process. The more leisurely pace of the self-administered procedure gives the respondent more time to understand the meaning of the question, and retrieve and compose an answer, which improves the quality of answers (cf. Schwarz, Strack, Hippler & Bishop, 1991). This is especially important when surveying special populations, such as children, adolescents and elderly (De Leeuw & Collins, 1997). Additional advantages of mail surveys are low costs and minimum requirements of resources.

However, paper self-administered questionnaires have a serious draw-back: only relatively simple questionnaires can be used. Complicated skip and branch patterns or adjustments of the order in which the questions are posed, threaten both the data quality and the motivation of the respondent to complete the questionnaire. Computer Assisted Self-Interviewing (CASI) overcomes these problems, and makes it possible to use very complex self-administered questionnaires successfully. In CASI the interview program handles the questionnaire logic and question flow. Respondents simply read each question from the screen, type in an answer, and are no longer burdened with complex routing instructions. In the case of very sensitive questions, the use of a computer may further enhance the feeling of privacy of the self-administered form. After an answer is given, it disappears from the screen, while an answer written down remains on the paper for everyone to see. Therefore CASI is especially suited for special population surveys on sensitive topics.

CASI can take several forms, for instance, it can be a part of a personal (CAPI) interview where the interviewer hands over the computer to the respondent for specific questions. Another form is a computerized version of the mail survey: Disk-by-Mail. We have used both forms in applications for special populations. In the first study we implemented a Disk-by-Mail survey on bullying in primary schools; the respondents were pupils aged 8-12 years. The second study was a survey on personal networks, social support, and well-being of adolescents and young adults with a visual impairment (aged 14-24). This study was a mixed-mode CAPI and CASI survey.

In this paper we start with a literature review of data quality in CASI-surveys, and we then describe the general logistic of both surveys and the special adaptations we had to make. We present empirical findings on data quality and end with general recommendations for the adaptation of computer assisted (self) interviewing for special populations.

2. DATA QUALITY IN COMPUTER ASSISTED SELF INTERVIEWING (CASI).

In this section we review the literature on acceptability of CASI for respondents and the impact of CASI on data quality.

2.1. Acceptability for respondents

Respondents generally like CASI; they find it interesting, easy to use, and amusing (Zandan & Frost, 1989; Witt & Bernstein, 1992). Beckenbach (1992, 1995) reports that more than 80% of the respondents had no problem at all using the computer and the interviewing program, and that few respondents complained about physical problems such as eye-strain. Furthermore, respondents tend to underestimate the time spent answering a computer assisted questionnaire (Higgins, Dimnik & Greenwood, 1987).

The general positive appreciation of CASI also shows in the relative high response rate with Disk By Mail (DBM) surveys. DBM response rates vary between 25% and 70%, and it is not unusual to have response ratio's of 40 to 50 percent without using any reminders (Saltzman, 1992). Assuming that DBM is typically used with a special population interested in the research topic, a comparable, well conducted, paper mail survey using no reminders may be expected to yield about 35% response (Dillman, 1978; Heberlein & Baumgartner, 1978). Of course, one should realize that DBM is restricted to special populations who have access to a computer.

2.2. Effect on data quality

The technological possibilities of CASI have a positive influence on data quality. Item non-response is minimized by computer controlled routing and by checking whether an answer or a 'do-not-know' is entered before proceeding to the next question. A consistent finding in the literature is that item-nonresponse caused by respondent- or interviewer errors, is virtually eliminated, but that there is little reduction in rates of explicit 'do-not-know' and 'no-opinion' answers (Nicholls, Baker & Martin, 1997)

As respondents are generally positive about CASI, we expect that respondents will experience a higher degree of privacy and anonymity, which should lead to more self-disclosure and less social desirability bias. Support for this hypothesis is found in the literature. In a meta-analysis of 39 studies, Weisband and Kiesler (1996) report a strong and significant effect in favor of computer forms. This effect was stronger for comparisons between CASI and face-to-face interviews, but even when CASI was compared with self-administered paper-and-pencil questionnaires, self-disclosure was significantly higher in the computer condition. The effect reported was larger when more sensitive information was asked. Weisband and Kiesler (1996) also report the interesting finding that the effect has been diminishing over the years, although it did not disappear! They attribute this to a growing familiarity with computers and their possibilities among the general public.

The effect of computerization on the quality of the data in self-administered questionnaires has also been a concern in psychological testing. In general, no differences between computer assisted and paper-and-pencil tests were found in test reliability and validity (Harrel & Lombardo, 1984; Parks, Mead & Johnson, 1985). This is confirmed by a meta-analysis of 29 studies comparing conventional and computerized cognitive tests (Mead & Drasgow, 1993). However, there are some indications that time pressure interacts negatively with the perceptual and motor skills necessary for reading questions from a screen and typing in answers correctly. Therefore, respondents, especially when they are a special or 'difficult' group should never be put under time pressure (for a more detailed discussion see De Leeuw, Hox & Snijkers, 1995).

In sum: empirical comparisons between paper-and-pencil and computer assisted self-administered questionnaires point to less item-nonresponse and more self-disclosure in the computer assisted form. Furthermore, respondents like this method, which is reflected in its high response rates compared to paper questionnaires.

3. A DISK BY MAIL SURVEY OF PUPILS IN PRIMARY SCHOOLS[2]

In spring 1995 a Disk by Mail survey was implemented in 106 primary schools; they formed a sample of primary schools and were scattered all over the Netherlands. The respondents were 6428 pupils, aged 8-12; the topic of the questionnaire was bullying. The questionnaire of 99 questions focused on attitudes regarding bullying, handling of bullying by teachers and parents, and actual bullying, either as a victim or as active culprit.

Traditionally this type of research is done with group administration of paper self-administered questionnaires in the classroom. This method has two severe drawbacks: lack of motivation of pupils to complete a long paper test and the potential influence of the close proximity of classmates on the answers (Scott, 1997). As pupils are in general very reluctant to talk about bullying, even to their parents or teachers, we searched for a procedure that enhanced feelings of privacy and created a more informal, relaxed mood. To keep the children motivated it is important that the questionnaire *appears* simple and attractive. CASI can meet these demands. An additional point is that printing and mailing such a large number of questionnaires will be rather costly. Thanks to a large government sponsored project to improve computer literacy among the young, all primary schools in the Netherlands are equipped with personal computers of the same type, and teachers have a basic knowledge of computer technology. Therefore, the basic requirements for a successful DBM were met (cf Witt & Bernstein, 1992).

3.1. Logistics

A Disk by Mail version of the questionnaire was developed using the Ci3-program. Range checks were defined for all questions, and questions were randomized within blocks of related questions. A special code (9) was defined for 'do-not-know'; however, this possibility did not appear on the screen, but was stated in a special instruction. To accommodate this special population, the possibility was created for a temporary stop when a child was tired or when the teacher needed a pupil. The pupil could resume answering the questionnaire at a more convenient time. Also, to make the task as simple and attractive as possible, special attention was given to the screen lay-out. A paper version of both questionnaires was available as back-up. Six schools used this paper version; the main reason was that those schools were extremely large, and that it would take the teachers too much time to have their pupils take the individual computer questionnaire.

A small package, consisting of two or more disks (depending on the number of computers), three short printed instructions and an accompanying letter, was sent to the teachers of the participating schools. The disk contained automated batch-files for starting the questionnaires, pausing and resuming, saving the data, and making back-ups. Two of the printed instructions were for the teacher: one gave instructions on how to start up the children's questionnaire, one gave instructions to start up a special teacher's questionnaire. The third instruction, a yellow page with eight points in large letters, was developed for the pupils. This instruction was simple and to the point and was always kept besides the computer. Main points in the instruction were the use of <enter> and <back space>, and an explanation of the 'beeb' when a child gave an out

[2] For more details see Van Hattum & De Leeuw (1997)

of range answer or used <enter> without giving an answer. The instruction also stated that they were allowed to type in '9' if they REALLY could not give an answer to a specific question.

The teacher implemented the questionnaire and allocated pupils to answer the questionnaire individually on the computer. A telephone help desk was operating, and if necessary people were stand-by to go to a school with problems. Also several university laptops were available as back-up or as an additional computer for large schools. In one case, an assistant went to the school to give general support; this school had specifically asked for assistance because they were very worried if they were capable enough to do the 'computer things'.

3.2. Data quality

We investigated the acceptance of the method, the data quality, and the costs involved.

Acceptance: At the end of the data collection period the participating teachers received a personalized report based on the results of their class and were asked to complete a short evaluation questionnaire. The results were encouraging. Teachers were positive, even elderly teachers and teachers with limited computer experience. Furthermore, even the youngest children liked the procedure. The teachers also reported few problems during the data collection. The problems that were encountered were mainly general reading or language problems, not technical ones concerning the computer or keyboard.

Data quality: We could also compare the results of the CASI-questionnaire (245 classes) with those of the paper-and-pen questionnaires (PAPI) that were used in a limited group of very large schools (18 classes). The classes were comparable regarding background characteristics of the teacher (e.g., teaching experience, education, class level).

A far higher percentage of *missing values* (p=0.00) occurred in the PAPI-condition. In the CASI-group the mean percentage of missing values was 5.7 while in the PAPI-condition the mean of the percentage missing was 14.1. A very interesting result is that the corresponding standard deviations also differed strongly between the groups. In the CASI-condition the standard deviation was 3.4, in the PAPI-condition the standard deviation was 25.0. These results suggest that not only the average amount of missing data is less in computer assisted data collection, but also that the individual variability, indicated by the standard deviation, is less. This can be attributed to the fact that with a paper questionnaire children who are not very concentrated or who are careless can easily skip a question or even a whole page by mistake. CASI forces children to be more precise.

The main pupil's questionnaire also contained a short test measuring the tendency to give *socially desirable answers*, a high score on this 9 item-test indicates that a child has the tendency to give honest, socially undesirable answers. There was a significant difference (p= 0.00) between the two conditions. Children in the CASI-condition gave slightly more undesirable answers (mean= 30.6) than children in the PAPI-condition (mean= 29.9). The standard deviations did not differ between conditions.

Regarding *openness* and *self-disclosure* we looked at the answers on both the bullying test and the victimization test. Children in the CASI-condition reported that they were actively involved in more bullying than children in the PAPI-condition (p= 0.00). The mean score for the CASI-condition was 30.5, while the mean score in the PAPI-condition was 27.7. In the CASI-condition also more victimization was reported (p= 0.00). The mean score on the victimization test was 26.4

for the CASI-condition and 23.1 for the PAPI-condition. Again standard deviations did not differ between conditions.

Besides data quality, *costs* are an important factor too. Cost comparisons are always difficult. To give a fair comparison we calculated the costs we made, and compared this with the costs we would have made if we had done the same survey by paper-and-pen. The costs of sampling, of developing the questionnaire, and of keeping account of the returned questionnaires are not taken into account; these would have been approximately the same in both cases. In the CASI-case we included costs for acquiring the CI3-program, for computer disks, programming, staffing the help-desk and mailing. For the paper equivalent we include printing and mailing costs using the cheapest mailing procedures. We also included the costs for data entry and coding. For the DBM-procedure the total costs were $1.01 for each completed questionnaire, in the paper mail survey this would have been about $3.22.

In sum, we showed that:

1) A Disk-by-Mail survey can be successfully implemented in Dutch primary schools.
2) Children from the age of 8 years on can successfully complete a computer assisted self-interview, and enjoy it.
3) Data quality in the computer-assisted group was better than in the paper and pencil group.
4) DBM results in less costs for each completed questionnaire compared to a PAPI mail survey.

4. A MIXED-MODE CAPI AND CASI SURVEY OF VISUALLY IMPAIRED AND BLIND ADOLESCENTS AND YOUNG ADULTS[3]

The second challenge was a study of blind and visually impaired adolescents and young adults (aged 14-24). In total, 354 respondents scattered over the Netherlands had to be interviewed about their personal network, experienced social support, feelings of loneliness and self-esteem, well-being, and handicap-acceptation. This resulted in a complex questionnaire of more than 260 questions.

Especially the questions on the ego-centered network are very complex for interviewers to administer. First, every important network member in specific domains (e.g., family, friends, neighbors) has to be enumerated. This is followed by questions on practical and emotional support for each listed network member. To ease the task of the interviewer and to minimize interviewer error, a computer assisted procedure seemed appropriate. In CAPI (computer assisted personal interviewing) the interview program takes over and handles the questionnaire logic and question flow; interviewer errors are averted and the interviewer has more time to concentrate on the respondent and establish rapport (cf. De Leeuw, Hox & Snijkers, 1995).

The questions on self-esteem, well-being, and loneliness are of a sensitive and private nature. Therefore, a paper self-administered questionnaire was used in earlier Dutch studies among

[3] For more details on the background of the study and first results see Kef, Hox, Habekothé, 1997.

'sighted' adolescents and young adults. Because of the highly sensitive nature of these questions and for reasons of comparability, CASI was the best choice for this part of the questionnaire.

For this study a mixed-mode CAPI-CASI survey was the best choice, provided that specific adaptations of the procedures were made to accommodate the special needs of the blind and visually impaired respondents.

4.1. Logistics

A computer version of the questionnaire was developed using Ci3. Lists of persons were used in a roster-function with the network questions, and range checks were defined for most of the questions. Also additional interviewer reminders were programmed in; for instance, when to hand over the computer to the respondent for the CASI part. Some extra adaptations had to be programmed for the CASI-application.

We opted for a 'manual' Audio-CASI. At the time of our survey Audio-CASI equipment was still in the developmental stage (Johnston & Walton, 1995; O'Reilly et al, 1994), and no standard solution were available. We devised the following procedure:

The interviewer handed over the computer to the respondent, making clear by shifting audibly the chair that she could not see the screen or keyboard. The interviewer had the text of the questions in writing and read them out aloud to the respondent, who typed in the answers. To synchronize the text of the question on the screen with the one the interviewer was reading, a series of 'beebs' was programmed to sound after a response was typed in by the respondent. The questions were all rating-scale type, and the respondent had to type in just one numerical key. For the Audio-CASI a special hardboard template was developed to cover the keyboard. In the template the part for the numbers from 1 to 0 was cut out, since it was only necessary to use these keys. At the appropriate places above the keys the hardboard template had both braille and magnified numbers, enabling the respondents to use the keyboard themselves while answering.

To support the respondent's memory, we also developed paper flash-cards with the response-categories used. There were three versions: one with braille, one with a very large magnification and one with little magnification.

The questionnaire and the procedure were pre-tested extensively, using qualitative pretests and a small scale pilot study. Interviewers attended a three day course. Topics were standard interviewer training, handling the laptop, the contents of the questionnaire, an introduction in CAPI and CASI, and the structure of the computerized questionnaire. Very important issues in the training were the special adaptations in the interview and specific skills concerning our target population: blind and visually impaired adolescents. The training included a visit to a special school for the visually impaired.

The questionnaire was implemented on the laptops of the interviewers, together with an automated system for making backups and a virus-scanner. Before the fieldwork started each laptop was thoroughly tested, including the interview program and the back-up facilities. A disk-version of the questionnaire was available as stand-by. The stand-by version was implemented to run adequately on a diversity of MS-DOS computers; if the interviewer laptop should break down, the respondents own personal computer could be used. During the fieldwork period both laptops and software proved to be very robust. A paper field guide was prepared for the interviewers. It contained the text of the questions for the Audio-CASI part, a summary of basic

interviewer rules, and a short manual summarizing the main computer commands and help with problems. Also, a field manager could be consulted by phone, even at odd hours in the evening and during the weekend.

The fieldwork took five months (March-July 1996). During that period sixteen interviewers traveled all over the Netherlands, each approximately interviewing twenty respondents. An interview, including the CASI-part, took on average 90 minutes.

4.2. Data quality

For obvious reasons we did not have results on a paper questionnaire with which to compare our data. However, we did have several possibilities to check the acceptance of the methods used and the internal validity of the data.

To investigate respondents *acceptance* and to systematically list any problems that may have occurred during the data collection, we had structured interviewer debriefing sessions. As the knowledge of interviewers and the information they possess on the past interviews is often rather diffuse and unstructured, we used concept mapping. This is a qualitative, highly structured method specially developed to extract information and quickly proceed from fuzzy knowledge to an acceptable conceptual framework (Trochim, 1989). Also, available were the results of short evaluations of both respondents and interviewers, completed immediately after the finished interview.

The experiences of the blind and visually impaired adolescents were very positive. In the Netherlands, almost all blind and visually impaired young persons are very familiar with computers. In general, a computer means a lot to these respondents and is not frightening for them. Many respondents asked a large number of questions about the kind of laptop used and the reasons why we used a computer in this study. Our mixed-mode approach created interest and motivated the respondents. CASI gave the respondents more privacy and offered more variation in
the interview-situation, while CAPI proved efficient with the complex network questions. The interviewers stressed that it was important to clearly verbally state that they were not looking at the screen during the CASI-part. The hardboard device worked well and the respondents had no difficulties with the typing-in of the answers. Accidentally, some respondents pushed some not-important keys through the hardboard device. Since the questionnaire was programmed to accept only numerical input at this point, this created no problems.

The CAPI-part and its adaptation to the special population did not give any problem, the special cards with response categories in braille and large letter type worked extremely well. Again the interviewers mentioned that it was extremely important to verbalize every action. When interviewing visually impaired, only a limited channel capacity of communication is available (audio and touch). Interviewers had to heavily rely on verbal and paralinguistic communication (e.g., humming instead of nodding as a positive reinforcement).

To investigate the *internal validity* of the data, we checked missing values, psychometric reliability and interviewer variance. First of all, no *missing values* occurred.

To examine the psychometric *reliability* the responses to the multi-item scales were analyzed. For each multi-item scale Cronbach's coefficient alpha was computed for the whole group of respondents and for subgroups (i.e., blind vs visually impaired). We expected that it would be somewhat harder for the blind to use the CASI-part resulting in somewhat less consistent answers. This was not confirmed by the data. In the whole group and in the subgroup the multi-item scales had sufficient reliability. No significant differences in reliability of scales were found between sub-groups.

Finally, we investigated whether there were any interviewer effects for the question on network size. Again, we analyzed the data for the whole group and for the blind and visually impaired subgroups separately. Although we expected that the blind needed more assistance, resulting in a larger interviewer effect, this was not confirmed by the data. In fact, no interviewer effects on network size were found for the whole group, nor for the subgroups.

In sum:

1) A mixed CAPI-CASI or CAPI-only approach can be successfully used with visually impaired adolescents and young adults.
2) Given the high level of computer sophistication of Dutch young visually impaired and the fact that almost all own a PC with braille adaptations, a CASI-only survey could be successfully implemented.
3) Acceptance is high. Both interviewers and respondents were positive in their reactions.
4) The special adaptations using braille and Audio-CASI procedures worked well.
5) The combination of computer-assisted data collection and well-trained interviewers results in good data quality.

5. Current Best Methods Plus: Recommendations for computer assisted interviewing of special groups

In a successful survey of special groups, adaptations have to be incorporated in the Current Best Methods available for a quality survey: one needs CBM+. Just standard good practice with some adaptations is not enough. With a special group a slight error in the questionnaire or procedure is more difficult to compensate; its influence will be magnified and data quality will suffer more than usual. To optimize data quality, the best practices in survey research, computer technology, and adaptations to the group should be combined into one total survey design aiming at Total Quality Management. Main points in CBM+ are: 1) optimize the design by pre-analysis of goal of study, group to be surveyed, and logistics; 2) optimize questionnaire and proceedings by using the CAI-potential fully 3) check the TOTAL design by pretests of questionnaire, implementation, and procedures; 4) build in repairs for the rare cases that errors will occur. A CBM+ system is 'fool'-proof, and when the fool beats the system, there is a repair mechanism. We want to stress that CBM+ can be implemented using existing, flexible software, such as Sawtooth Ci3. We will give some examples.

5.1. Optimizing the design.

The most important step here is a systematic analysis of the group. Points for consideration are:

- development of cognitive skills of the respondent (e.g., different stages in children, elderly)
- available channel capacities in interview (audio, visual & paralinguistic)
- social customs (social customs may differ)
- hazards to eye-hand coordination (e.g., hospital patients)
- computer literacy
- easy access to computers, either their own or a company or school computer
- ease of safely providing the members with a computer on a temporary basis (e.g. have a computer delivered with some instruction for a key contact at a hospital)
- availability of key contacts as help to introduce the survey (e.g., a teacher, a trained matron in a hospital ward, a social worker)

Some examples: In Audio-CASI, the audio- and paralinguistic channels are most important to convey information. In some cases respondents have to rely on the audio channel only, making CASI resemble CATI more closely. The extensive research on CATI and data quality shows that only a limited number of response categories can be used. In our survey of the visually impaired we used Audio-CASI, combined with braille cards for the response categories to compensate for the limited channel capacity. When Audio-CASI is used one should use all channel capacities and have the text on screen in large letters too. Using both channels reduces the risk of information loss. In studies with very young children and illiterates it is wise not to rely on the visual channel and use questions with a limited number of response categories.

The survey of school children is a good illustration of the use of a key person on the spot to assist in the data collection. Another example is an evaluation study with hospital patients. As the questions were rather sensitive, the research firm decided to use CASI. A representative of the research firm visited the hospital with a laptop and gave some basic instructions to the matron or executive nurse. The matron could bring the laptop to a patient at an optimal time for hospital and patient and start the interview-program. Only very simple keystrokes were necessary to answer questions and screen contrast was heightened to enable using the laptop in bed.

5.2. Using CAI-potential fully

The strength of computer assisted interviewing is that intelligence can be built into the program. A very complex questionnaire, with checks of answers, complicated branchings, and randomization of response categories can be used safely. However, it is important that the questionnaire appears logical and simple. The magic word is **appear** simple and logical. What is seen on the screen should be simple, what happens in the program may be complex! The designer, programmer, and tester of the questionnaire may get headaches in solving problems, the respondent may not! These principles should be combined with CBM in questionnaire construction.

In constructing a CAI survey for special groups one should bear in mind that:

- The questionnaire should be experienced as simple and short and structured to compensate for fewer cognitive skills and smaller channel capacity.
- Point of reference is always the respondent. What is easy and logical for the respondent is not necessarily logical or easy for the questionnaire designer.
- Group questions in a logical order, use blocks of questions, use the same question format as far as possible, etc.
- Perceptual and motor skills necessary for responding to a computer assisted questionnaire are slightly more complicated and take somewhat more time than those necessary for paper-and-pen tests.
- Question texts are harder to read on a monitor than on paper, which implies that ergonomical text presentation and careful screen design is important.
- Easy key-stroke combinations should be available for answering. Respondent burden should be minimalized.
- Avoid mistakes, if possible use templates to cover keys that are not necessary or even 'dangerous.'
- Avoid any suggestion of time pressure, especially with inexperienced users. If eye-hand coordination is expected to be sub-optimal, allow for extra time.
- Respondents should be able to concentrate fully on the questions, they should not be distracted by extra tasks.
- When interviewers or 'helpers' on the spot are used, do not leave the solving of problems to the interviewer. Interviewer burden should be minimalized by well constructed and tested questionnaires. Interviewers need their attention for the special respondent NOT for the computer.
- Everything a system can do it should do. For instance, starting the questionnaire, making back-ups, keeping administrative records, stopping and resuming at the right point.

5.3. Pretest and check

Often there is not enough time and/or money to do extensive pretests and run a full pilot study. This should not be an excuse for omitting pre-testing altogether. Carefully planned, small scale pretests can be easily implemented at low costs. Qualitative, or cognitive, interviews with a small number of real respondents can detect many errors in the basic questionnaire. Dry-runs, after the programming, can be performed in-house. Observation of a respondent, in combination with in depth interviewing after the performance is a good method for testing the implementation.

In short:

- Pretest the questionnaire: does the respondent understand the meaning of the question, the meaning of terms used, the response categories. This can be done early with the paper version.
- Pretest routings (no respondents needed).
- Pretest the computer implementation (e.g., starting-up, making back-ups). After technical tests in-house, let a naive respondent try it out.
- End with a usability test on the final product. Check user-friendliness of system, but also screen lay-out, use of special keys, etc.

5.4. Build in repairs

Prevention is better than curing. But sometimes...

- Internal checks on 'out-of-range' answers and consistency checks are almost automatically employed in CAI. When employing these one should keep in mind, that a check alone is not all; the following message on screen should be clear to the respondent too!
- Have a short list on paper with instructions and meta-information. When something goes wrong, help-functions or a help-key often only confuse the flustered respondent. Use larger than standard letter-type without serif (e.g. Helvetica 20).
- Have a help-desk manned or use informed key-persons in the vicinity as 'help.'
- Make sure 'first-aid' diskettes are available with a complete back-up of the questionnaire, either with the key-persons or at the help-desk to be mailed out immediately.

5.5. Conclusion

DBM+ just asks a little bit extra. Most importantly is a systematic approach. Analyze the research problem and adjust your study accordingly. The above lists aid in the analysis and implementation of adjustments. It is not necessary to have software developed, quality standard software can be used to accommodate your special survey. The new developments in multimedia systems, using sound and video, will increase the power of the tools available for surveying special groups.

6. REFERENCES

Beckenbach, A. (1995). Computer assisted questioning: the new survey methods in the perception of the respondent. *BMS, 48,* 82-100.

De Leeuw, E.D. & Collins, M. (1997). Data collection method and survey quality: An overview. In: L. Lyberg et al. (eds). *Survey measurement and process quality*. New York: Wiley.

De Leeuw, E.D., Hox, J.J., & Snijkers, G. (1995). The effect of computer-assisted interviewing on data quality. A review. *Journal of the Market Research Society, 37,* 4, 325-344.

De Leeuw, E.D. (1992). *Data quality in mail, telephone and face to face surveys* (chap. 3). Amsterdam: TT-publikaties.

Dillman, D.A. (1978). *Mail and telephone surveys; The total design method*. New York: Wiley.

Harrel, T H & Lombardo, T A 1984. Validation of an automated 16PF administration procedure. *Journal of Personality Assessment*, 48, 216-227.

Heberlein, T A & Baumgartner, R (1978). Factors affecting response rates to mailed questionnaires; A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.

Higgins, C.A., Dimnik, T.P., & Greenwood, H.P. (1987). The DISQ survey method. *Journal of the Market Research Society*, 29, 437-445.

Johnston, J. & Walton, C. (1995). Reducing response effects for sensitive questions: A computer assisted self interview with audio. *Social Science Computer Review*, 13, 304-319.

Kef, S., Hox, J.J. & Habekothé, R. (1997). *(On)zichtbare steun*. In Dutch [(In)visible support; a study of visually impaired young adults and their personal network]. Amsterdam: Thesis publishers.

Mead, A D & Drasgow, F (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, 114, 449-458.

Nicholls, W.L.II, Baker, R.P., & Martin, J. (1997). The effect of new data collection technologies on survey data quality. In: L. Lyberg, et al. (Eds). *Survey Measurement and Process Quality*. New York: Wiley.

O'Reilly, J.M., Hubbard, M.L., Lessler, J.T., Biemer, P.P., & Turner, C.F. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics, 10*, 197-214.

Parks, B T, Mead, D E & Johnson, B L (1985). Validation of a computer administered marital adjustment test. *Journal of Marital and Family Therapy*, 11, 207-210.

Saltzman, A. (1993). Improving response rates in Disk-by-Mail surveys. *Marketing Research, 5,* 32-39.

Sawtooth, 1994. *Ci3 User Manual* (edited by Harla Hutchinson & Margo Metegrano). Evanston: Sawtooth Software.

Schwarz, N., Strack, F., Hippler, H-J, & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*, 193-212.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In: L. Lyberg, et al. (Eds). *Survey Measurement and Process Quality*. New York: Wiley.

Sudman, S. & Bradburn, N.M. (1974). *Response effects in surveys: A review and synthesis. Chicago: Aldine.*

Tourangeau, R., & Smith, T.W. (1996). Asking sensitive questions; the impact of data collection, question format, and question context. *Public Opinion Quarterly*, 60, 275-304.

Trochim, W.M.K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, 12, 1-16.

Van Hattum, M. & De Leeuw, E.D. (1996). *A Disk by Mail survey of teachers and pupils in dutch primary schools; logistics and data quality*. University of Amsterdam Department of Education, Methods & Statistics Series # 57.

Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta analysis and implications. *CHI '96* (http://www.al.arizona.edu/~weisband/chi/chi96.html).

Witt, K.J., & Bernstein, S. (1992). Best practices in Disk-by-Mail surveys. *Sawtooth Software Conference Proceedings*, Sawtooth Software: Evanston, Illinois.

Zandan, P., & Frost, L. (1989). Customer satisfaction research using disks-by-mail. *Sawtooth Software Conference Proceedings*, Sawtooth Software: Evanston, Illinois.

BEST PRACTICES IN INTERVIEWING VIA THE INTERNET

Karlan J. Witt
IntelliQuest, Inc.

INTRODUCTION

With the adoption of the Internet as a means of communication, researchers and marketing professionals in all types of companies are scrambling to identify what leverage can be obtained from this new medium. It can certainly offer faster, cheaper ways of collecting data, and in many cases, will even provide a more targeted list of respondents. The Internet brings with it a host of unique limitations, however, that impact any research effort in this area. The purpose of this paper is to explore each aspect related to collecting data over the Internet, describe the challenges which exist, and provide some suggestions for overcoming them. The scope of this paper is to examine the quantitative surveys which might be conducted online. The Internet lends itself to a whole new world for qualitative interviewing as well, but these online opportunities are not addressed here.

This paper is organized into nine sections, beginning with a brief review of the history and background of interviewing via the Internet. The next two sections describe different types of Internet data collection methods, and provide an extended discussion of criteria to use in evaluating the appropriateness of using these methodologies. The paper then discusses many factors which affect response rate, provide some typical response rate using this survey modality, and includes some reactions from respondents to taking Internet surveys. The next section summarizes the limitations associated with interviewing over the Internet, and illustrates the timing and costs associated with this new method as compared to more traditional ones. The paper then summarizes the best practices for interviewing via the Internet, and ends with a look at the future of Internet surveying.

HISTORY OF INTERNET INTERVIEWING

The history of Internet interviewing actually begins with electronic interviewing. Electronic interviewing encompasses interviewing over any type of network, within a company or over some other dedicated network. It also includes disk-by-mail (DBM) interviewing. In 1985, IntelliQuest began conducting electronic interviews over dedicated networks and collecting data for employee surveys over company networks. The limitations associated with these surveys were numerous, including the very defined nature of the sample which had access to the survey. However, where that audience was appropriate, it provided a faster, more economical, novel way to conduct surveys.

As DBM interviewing evolved and the use of personal computers proliferated, electronic interviewing had broader applications. With the development of commercially-available software such as Sawtooth Software's Ci2 and Ci3 programs, researchers found ways of leveraging this new technology in many ways. If the audience they wanted to survey were not computer users, they had the ability to recruit respondents to a central location to take surveys on PCs

provided for their use. As technology has continued to evolve to allow the use of video and graphics in the interviewing applications, the use has become broader and broader.

With the commercialization of the Internet, the vision is to migrate all that has been developed from other forms of electronic interviewing to the instantaneous, virtually cost-free environment now available. The methods are not interchangeable in all aspects, however, and careful consideration must be made in selecting the method which provides the best combination of research results, cost and timing.

TYPES OF INTERNET DATA COLLECTION

Throughout this paper, the terms "Internet data collection" or "interviewing via the Internet" are used. Before discussing details of using this mode of interviewing, the terms need to be defined. There is not one single way of collecting data over the Internet. While many share common strengths and weaknesses for different types of surveying applications, it is important to note their differences as well.

1. **Survey posted on a web site.** This is a survey which must be accessed through a browser[1] on a computer or Web TV-type device. This survey is one which users who happen across the web site can self-select themselves to complete. Another means for obtaining respondents for this type of survey is to advertise (over the Internet, or through more traditional media like print advertising) to get users to go to the site to complete the survey. This is a true convenience sample of Internet users.

2. **Survey on a web site accessible only by targeted respondents.** This type of survey differs from the first not in the technology used to create it, but rather in who the target respondents are. The respondents for this type of survey have received a specific phone call which screens them for certain qualifications, or in the event of a known list, perhaps an email which provides the location of the survey for the respondent to complete. The surveys are typically protected from general public viewing, and the password a respondent is given can only be used once, preventing a chain-mail broadening of your sample without your knowledge. Target respondents can be sent an email message with a link which will take them directly to the survey web site, but currently not all email software applications support this method. For others the email message can contain the URL[2] with specific instructions regarding how to navigate to the site.

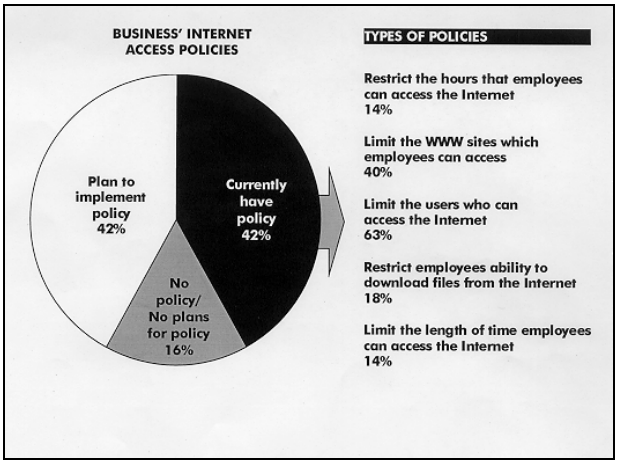
3. **Survey sent out as an attachment to an email message, or as part of the email itself.** The primary difference between this and the second type is where the interview is completed. In both cases the respondent must be screened by phone, or in some other way qualified and their email address obtained. The survey can then be sent to them to complete and return rather than their coming to a central site to complete the survey. As of the second quarter of 1997, almost half of all Internet users pay for time spent online, as opposed to having unlimited access to the Internet. The option of sending a survey to the

[1] A browser is a software application which allows users to view information. The most common uses of browsers are for viewing information available on the world wide web, and to view information on an organization's intranet. The browser provides the mechanism for navigating to the information the user desires, and depicting the text or graphics available.

[2] URL is the Universal Record Locator and is the name of the web site to which you are directing the respondent. For instance, "intelliquest.com" is the URL for IntelliQuest's web site.

respondent is the only option which doesn't require them to pay for the time required to complete the survey. As the adoption of the Internet continues, however, we are seeing that other issues may impact the ability to send questionnaires to some types of respondents. In surveying business Internet users, we have already seen organizations which have security mechanisms including blocking attachments to messages which can pose a security threat.[3]

Figure 1



It is possible to send out this type of survey to individuals who use the Internet, but do not necessarily have a browser to surf the World Wide Web (WWW). However, IntelliQuest has found that the number of Internet users who do not have access to the WWW is very small, and shrinking over time. In Q2 of 1996, 24% of Internet users did not access the WWW. This had fallen to 15% by Q4 1996 and 12% by Q2 1997. The majority of this paper addresses the use of graphical surveys, emphasizing those which can be viewed through the WWW.

WHEN TO COLLECT DATA OVER THE INTERNET

The wide-spread adoption of the Internet offers researchers a unique environment in which to conduct surveys. The decision about whether or not to collect data over the Internet is a complicated one, owing to many aspects of the research design. This section of the paper addresses these design issues, and their impact on the selection of the Internet as a data collection modality.

1. **Research sample.** There are five primary considerations regarding the appropriateness of Internet surveying for the target population of respondents.

   - **Respondents have or can be given access.** Potentially the biggest limitation of interviewing over the Internet is that it is only appropriate when the population of interest all have access to the Internet. In certain instances, the benefits of conducting surveys via the Internet may be compelling enough to warrant providing Internet access to target respondents to avoid sample bias by including only those who already have access. This is obviously not appropriate for all non-user audiences, and care must be taken to examine the other issues about target respondents to ensure this is an appropriate survey modality. An

[3] IntelliQuest's IntelliTrack IQ™ Intranet Study Q1 1997.

example of when this is an audience that many companies might want to survey relates to the evaluation of a company's web site, that might be used for marketing, service and support, and even sales. IntelliQuest in conjunction with *USA Today* runs a biweekly program evaluating companies' web sites. An example of those results appears below.

Figure 2

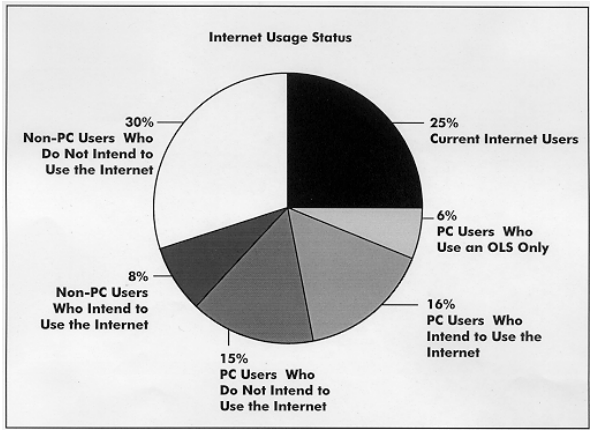
In general, any evaluations having to do with the look and feel of your web site can very effectively be measured using an online survey.

But what if you aren't sure whether or not an Internet survey is right for your target audience? Here are three issues to consider:

◊ **Absolute Internet usage percent penetration.** There has been much hype surrounding the wide-spread adoption of the Internet. The Internet, it seems, is even driving people to adopt a PC for the first time, just in order to "surf the web." Respondents to a recent IntelliQuest survey were queried as to their current or intended Internet usage. Of the people who intended to begin using the Internet, 33% were not current PC users.[4] Despite the real phenomenon as well as the hype, as of the second quarter of 1997, only about 25% of the U.S. population had access to the Internet (including all locations),[5] which would preclude its use for a random study of U.S. residents. The incidence of Internet usage is growing, however. From Q2 of 1996 to Q2 of 1997, we saw it increase from 16% to 25% of the U.S. population.[6]

[4] IntelliQuest's Worldwide Internet/Online Tracking Service Q4 1996.
[5] IntelliQuest's Worldwide Internet/Online Tracking Service Q4 1996.
[6] IntelliQuest's Worldwide Internet/Online Tracking Service Q2 and Q4 1996.

Figure 3



One issue that is particularly relevant when screening for Internet users is that disclosing the subject matter of the survey during an introductory statement will likely cause those potential respondents who are not users, or are not even aware of the Internet, to refuse to complete the interview. If any questions relate to market incidence of users, or comparing habits or perceptions of users to non-users, this must be avoided. Appendix A contains an example question series which accomplishes that by capturing the incidence information prior to zeroing in on the subject matter for the survey:[7]

◊ **Representativeness of Internet usage penetration.** In addition to identifying the absolute Internet usage penetration, it is also useful to examine the representativeness of that audience. As Internet adoption increases, the representativeness of that audience increases as well.[8] It is still, however, not a clean proxy for the overall U.S. population. As shown below, it is also not a clean proxy for PC users.

Figure 4

	US POPULATION	INTERNET USERS Q2 1996	INTERNET USERS Q4 1996	INTERNET USERS Q2 1997	PC USERS Q2 1997	INTERNET INTENDERS Q4 1996
Age	16-34 = 42% Median = 40	16-34 = 49% Median = 35	16-34 = 55% Median = 36	16-34 = 53% Median = 36	16-34 = 49% Median = 35	16-34 = 47% Median = 37
Education	20% College Grads	46% College Grads	45% College Grads	37% College Grads	30% College Grads	26% College Grads
Gender	48% Male	64% Male	55% Male	53% Male	52% Male	42% Male
Income	57% < $50K	40% < $50K	45% < $50K	45% < $50K	49% < $50K	63% < $50K

◊ **International issues.** When attempting to conduct research across many geographies, the issue of whether or not to survey via the Internet is even more complicated.

[7] IntelliQuest's Worldwide Internet/Online Tracking Service Q2 1997.
[8] IntelliQuest's Worldwide Internet/Online Tracking Service Q2 and Q4 1996.

- **Offers consistency across geographies.** The survey can be offered in one or many languages, and the administration of the survey via the Internet eliminates the variations that can occur when conducting telephone or mail surveys in many countries.

- **Penetration is even more restricted in other geographies.** Depending on which countries are to be included in the survey, the issue of absolute penetration and the representativeness of it is compounded further. In Q2 1996 the incidence of Internet access in the U.S. was 16%, while in Germany, France and the U.K., incidence rates were 6%, 3%, and 5% (respectively). The complexion of users in other countries varies greatly as well. In the U.S., Internet adoption was driven largely by home users, and remains the largest segment, although business is growing rapidly. In Europe, where the regulation of the telecom industry complicates the cost and availability of Internet access, users are predominantly business users.[9]

Figure 5

	US	UK	GERMANY	FRANCE
Overall Incidence of Internet Usage	31.6M	2.2M	4.0M	1.3M
% of Overall Population	16%	5%	6%	3%
% of Users Accessing from Work	44%	60%	57%	38%
% of Users Accessing from Home	58%	32%	44%	17%
% of Users Accessing from School	23%	31%	22%	30%

- **Email addresses of target respondents is known or can be obtained.** Addresses may be known when using a listed sample or a panel who has been profiled for Internet access. The use of these sample sources has its own limitations, however.[10] If the sample source is random and email addresses must be obtained, there is added cost involved in conducting prescreen interviews.

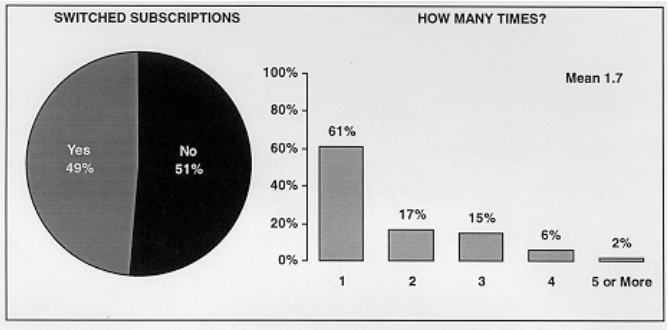
  Listed samples which contain email addresses for potential respondents may become stale as Internet users move from company to company to switch Internet or online service providers. While the churn rate among home Internet users does not currently approach those of long distance providers, it is still an issue in constructing a valid research sample. From IntelliQuest's Worldwide Internet/Online Tracking Service (WWITS™), it is clear that the types of individuals who switch providers for the latest deal are systematically different from those who have never switched.[11]

[9] IntelliQuest's Worldwide Internet/Online Tracking Service Q2 1996.

[10] Rich, Clyde L. (1977). "Is Random Digit Dialing Really Necessary?" *Journal of Marketing Research*, Volume XIV. 300-305.
Blankenship, A.B. (1977). "Listed versus Unlisted Numbers in Telephone-Survey Samples." *Journal of Advertising Research*, Volume 17, Number l. 39-42.

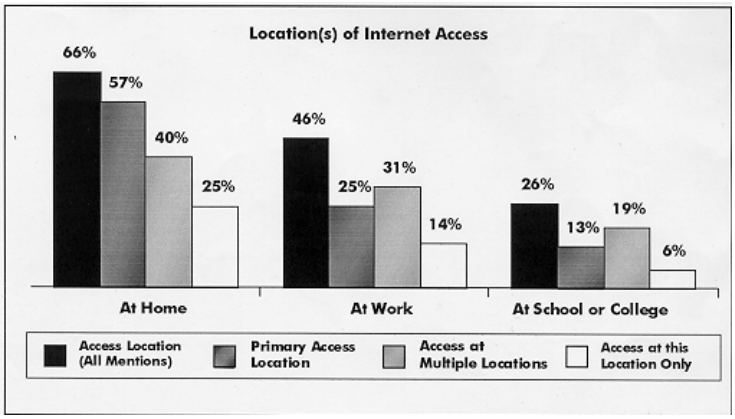
[11] IntelliQuest's Worldwide Internet/Online Tracking Service Q4 1996.

Figure 6



- **Location of access to Internet.** As mentioned above, Internet users may access the Internet from home, work, school, the neighborhood library, a friend or relative's house, etc. The location from which they access the Internet may impact the ability (or desirability) to include them in a survey sample. Companies often have policies regarding the web sites which employees may visit, as well as the amount of time they are allowed to use it.

Figure 7



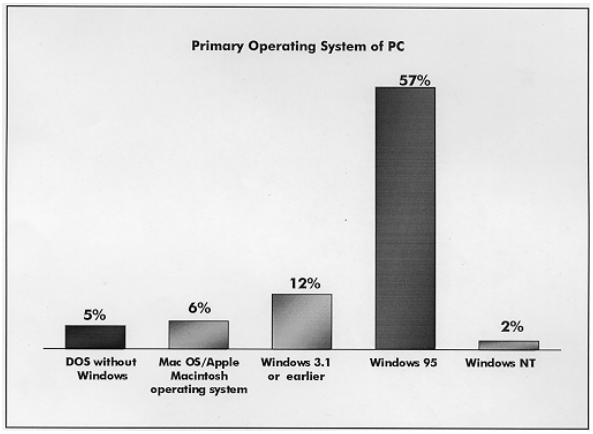
Although the home segment is currently the largest, the business and school segments are growing more quickly than the home market.

- **Appropriateness of an Internet survey for the target audience.** Once you have identified a population that uses the Internet, it is still relevant to question their comfort level in navigating to the desired web site and interacting with a survey through their browser. This might be determined through direct questioning, or by examining the types of

activities the target respondents perform while online. For example, respondents who use the Internet only for email may be uncomfortable with the skills required to participate in an Internet survey.

- **Need for surveying users of multiple platforms.** Often times researchers' information needs are not dependent on whether a person most often uses a PC running Windows or OS/2, a Macintosh, or even a UNIX-based workstation. Surveying over the Internet allows the creation of one survey which can be viewed across all these different platforms.[12] As of the second quarter of 1997, IntelliQuest found that a representative group of Internet users used the following operating systems:

Figure 8



2. **Questionnaire design.** The second set of variables which impact the selection of the Internet as a means of data collection are related to the design of the questionnaire itself.

   - **Reasons for using an electronic survey.** Electronic surveys offer capabilities that surpass any other data collection modality. These include the ability to incorporate stimuli such as graphical images and video or audio segments. These can be used for specific types of studies such as ad testing, new product design, and publication readership. Electronic surveys also offer enhanced capabilities to include complex programming such as conjoint analysis, randomized discrete choice exercises, and complex display or skip logic. Below is an extended list of issues in the questionnaire design phase that might drive the use of some sort of electronic survey.

     ◊ Programmed automatic skip patterns give respondents only relevant questions.

     ◊ Display logic can be used to construct lists with only the relevant responses based on answers to past questions.

     ◊ The survey has questions with long lists which are easiest viewed directly by the respondent.

     ◊ The survey contains many technical terms or acronyms, which are easily read, but difficult to discuss in a telephone interview.

[12] IntelliQuest (1996). "Interactive Questionnaire Software for Today's Global Marketplace."

- ◊ Responses can be constrained within appropriate bounds (e.g., numbers in response to a numeric question, range limits, constant sums, etc.).
- ◊ The survey can incorporate adaptive or intelligent modules, such as Adaptive Conjoint Analysis (ACA), which are most easily self-administered where they can customize questions based on previously-given information.
- ◊ The program can randomly select a subset of questions to show a respondent from a much larger set; a task that would be arduous at best on paper.
- ◊ Open-end questions capture accurate, lengthy verbatim answers without interviewer bias[13].
- ◊ Respondents perceive the survey to take less time to complete than it actually does[14].
- ◊ Randomization reduces order bias within lists and across questions.
- ◊ Less respondent fatigue than for a phone survey.
- ◊ Prevent respondents from looking ahead to concept or follow-up questions.
- ◊ Respondents cannot look ahead, as they can in a paper survey that is too long or too complex.
- ◊ Allows for a greater range of measurement (allows the researcher to use scales not possible to administer via the telephone, such as sliding scales, long response lists, etc.).
- ◊ Provides a more natural environment for surveys with many technical terms and achronyms.
- ◊ Provides rapid turnaround of data without waiting for manual data entry.
- ◊ Perceived by respondents to be innovative and novel.
- ◊ Use survey software designed to allow incorporation of graphical images, video or audio clips.

- **Software limitations on commercially available electronic survey packages.** While the software available for disk-by-mail surveys has been available the longest, and offers the richest set of features, the same cannot be said of Internet surveying packages. The vast majority of Internet surveying conducted today is done with a paper or "forms-based" paradigm. A forms-based survey looks to the respondent like a paper survey would, only they access it over the Internet. The benefits of electronic surveying, such as advanced logic, are not widely available. There are recently a small number of software packages on the market which do support most electronic survey features over the Internet. These include In2itive, Decisive Survey, Quantime, Socratic Software, Ronin's Results For Research package, and IntelliQuest's NetQuest™ survey software.

---

[13] Witt, Karlan J. and Steve Bernstein (1992). "Best Practices in Disk-by-Mail Surveys." *Sawtooth Software Conference Proceedings*. 1-26.

[14] Zandan, Peter and Lucy Frost (1989). "Customer Satisfaction Research Using Disks-By-Mail." *Sawtooth Software Conference Proceedings*. 5-17.
Higgins, C.A., T.P. Dimnik, and H.P. Greenwood (1987). "The DISKQ Survey Method." *Journal of the Market Research Society*, Volume 29, Number 4.

FACTORS AFFECTING RESPONSE RATE

As with any survey, there are many factors which impact response rate. For surveys conducted over the Internet, non-responders have the potential to be systematically different from responders in at least one respect: their access to the Internet. Although potential respondents may be screened for Internet usage, this may introduce a source of non-response bias.

While that one issue has been addressed earlier in the paper, there are other sources of non-response bias that need to be considered in order to create a representative study. These include:

- **Saliency of the survey topic to respondent.** The more interesting and relevant the topic of the survey is to the target audience, the higher the resulting response rate. If the topic is somehow more relevant to some potential respondents in the research sample than others, the non-response rate may differ by type of respondent, introducing bias into the study.
- **Length of survey.** There are two components to survey length which elicit behavioral responses from potential respondents. The first is the expected length of time to complete the survey, reported during the prescreen phone interview, or in the email soliciting the respondent's participation. This eliminates certain respondents who are unwilling to commit that amount of time to the interview. The second component is *perceived* time elapsed while taking the survey. While some respondents may begin an interview, they may terminate if they perceive the survey is too long.

  An interview is "too long" if it takes longer than expected to complete. It may also be "too long" if it bores the respondents, or if respondents have a difficult time answering the questions.[15]
- **Respect for respondents' time; high professional ethics.** While there is evidence that respondents will respond to longer surveys using an electronic modality[16], it is the responsibility of the researcher to always respect respondents' time.
- **Composition of the research sample.** Certain populations, such as purchase influencers and senior executives, are frequently asked to participate in surveys and place a very high value on their time. Both of these groups typically demonstrate lower-than-average response rates in research studies.
- **Convenience of taking the survey.** An electronic survey typically provides the convenience of completing the survey at a time of the respondents' choosing. This convenience provides an advantage of electronic surveys over all other survey modalities.

[15] Bahner, Lesley (1987). "Long Self-Administered Questionnaires." *Sawtooth Software Conference Proceedings*. 11-21.
[16] Witt, Karlan J. and Steve Bernstein (1992). "Best Practices in Disk-by-Mail Surveys*." Sawtooth Software Conference Proceedings*. 1-26.

- **Bandwidth of the Web server hosting the survey.** Sending an invitation to hundreds or thousands of Internet users to come to a given web site can quickly bring many servers to their knees. If the company sponsoring the research has let respondents know who they are, it can reflect poorly on them. It can also directly affect response rate negatively, as people will not wait online indefinitely while the screen says "contacted host: waiting for response." One suggestion is to send out the invitation to visit the site in groups, or replicates of sample, spread out over a short amount of time. Another option is to rent space on a server capable of handling the large volume of inbound traffic.

- **Sponsorship of the survey disclosed.** One of the key factors impacting response rate is whether or not the sponsor of the research is disclosed. While it is certainly not appropriate in most studies, disclosure is recommended when possible. This will have the benefit of increasing the response rate. Further, the sponsorship is most effective when the survey sponsor is respected by the target audience, such as in product follow-up surveys.

  Disclosing the sponsor may also benefit the sponsoring company. In one IntelliQuest customer satisfaction study, 35% of respondents stated that their attitudes towards the sponsor improved as a result of receiving the survey from the sponsor.[17]

- **Guarantee of anonymity or confidentiality.** Because the industry is still so new, respondents' impressions of the true confidential nature of surveys conducted over the Internet is yet to be determined. Respondents are currently less comfortable providing their email addresses than their street addresses[18]. It is likely that respondents' experiences with reputable and trustworthy research companies will increase their comfort level over time. Disguised sales pitches and having names sold to "spam"[19] mailing lists will greatly hinder future research endeavors with this audience.

- **Incentive.** Incentives are one of the most interesting and most debated response rate enhancers in survey research. Most sources report that incentives of any kind increase response rate.

  Electronic surveys offer the ability to offer incentives that would otherwise not be viable. Examples of these include free screen saver applications, computer games, and free software. These are inexpensive when purchased in bulk, can be delivered electronically at almost no cost, and provide an immediate, tangible "thank you" to the respondent. These can be programmed to allow respondents to obtain them immediately after completing the last screen of the survey. Based on IntelliQuest's experience, these have a higher perceived value as compared to $1 included in a mailed survey, and also have the benefit of being customizable to the target audience.

  Another potential incentive for this group is to pay for some of their Internet-access time. Based on proprietary work conducted by IntelliQuest, almost half of Internet users pay based on the amount of time spent online, and completing a survey would literally cost them money from their Internet provider.

---

[17] Zandan, Peter and Lucy Frost (1989). "Customer Satisfaction Research Using Disks-By-Mail." *Sawtooth Software Conference Proceedings*. 5-17.

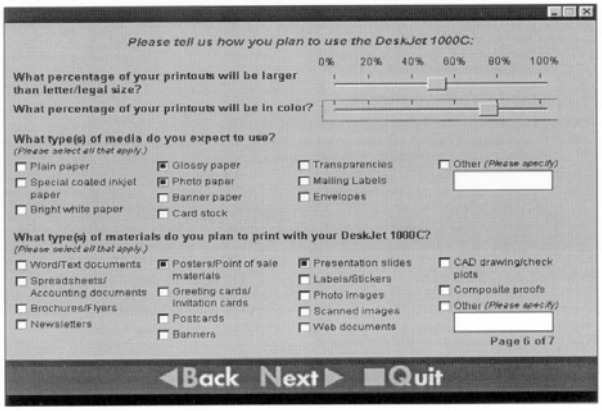
[18] IntelliQuest's Worldwide Internet/Online Tracking Service Q2 1997.

[19] Spamming is a technique for distributing unsolicited email to Internet users. It is the junk mail of the Internet.

- **Professional presentation of materials (survey programming, cover note, glossary of terms, etc.).** The quality of the materials the respondent receives at each stage in the process has an impact on their likelihood of responding. This includes the initial call or letter from the research company, the look and feel of the survey application, and the perceived quality of the questions themselves. Materials need to be complete and concise, absent of typos, and extremely professional. A toll-free number should be available for the respondent to call for help at any stage in the process.

  The look and feel of the survey application itself can greatly enhance or limit response rates. IntelliQuest has experienced response rates as much as double with the same sample group when employing the use of graphical surveys such as the one below:

Figure 9



  The Web surfers whose attention is targeted by these surveys are bombarded daily with Web sites that have employed the best and the brightest to create enticing graphics for their Web pages. These surfers can be likened to cable TV surfers, where they flip from "channel" to "channel" quickly, and a survey has to be designed to keep their attention.

- **Timing of the survey (time of year).** In analyzing surveys of all modalities over the last ten years, IntelliQuest has found that in the U.S., December is the single worst time to conduct interviews. Not only is the total response rate lower, but the effort required to maintain a representative sample is greater.

  Throughout the world, there are holiday times to be avoided. Some are religious or secular holidays, while others are just traditional vacation time, such as August in France. These specific times are ones your research provider will anticipate, and account for in the project schedule.

- **Prescreened by phone versus unsolicited survey.** In many studies, it is necessary to contact respondents in advance of the electronic survey to:

  ◊ Identify the individual who should receive the survey

  ◊ Pre-qualify individuals for the study

- ◊ Identify to which market segment or quota group a respondent belongs
- ◊ Screen for access to the Internet
- ◊ Verify email address

Even in instances where it is not necessary to conduct pre-screening phone calls for the reasons stated above, IntelliQuest has found that it increases response rate to pre-notify respondents, either by mail or by phone, prior to receipt of the electronic survey. Pre-notification legitimizes the survey and communicates its importance to the survey sponsor.

Additionally, pre-qualifying respondents by telephone ensures that all respondents receiving the survey are eligible to participate. If non-qualified respondents receive survey disks and do not respond, they are likely to be counted in the non-response. It is not non-response bias if an *unqualified* respondent does not respond.[20]

- **Time between pre-screen or pre-notification and receipt of survey.** It is important for respondents to receive the survey soon after the pre-notification. For a telephone pre-notification, IntelliQuest has found it most effective for respondents to receive the survey within three to seven days. With written pre-notification, IntelliQuest has found it most effective for the survey to be received within five to ten days.
- **Follow-up reminder.** As with pre-notification, a reminder call, postcard, or fax increases response rate. This follow-up may be used to thank respondents if they have already responded, and gain share of mind among those who have not yet responded.

TYPICAL RESPONSE RATES ON INTERNET STUDIES

Response rates on studies conducted over the Internet vary greatly. At this time, much of the research conducted is done so with listed samples. The response rate, then, varies with the quality and frequency of use of the list. Considering all the factors described above, researchers should expect response rates in the range of 35% to 80%.

When IntelliQuest first began collecting data via the Internet, we received emails from respondents commenting on the pleasure of taking the survey using this method, and indicating that their positive experience made them willing to participate in future Internet surveys. While this novelty will eventually wear off, it is illustrative of the fact that respondents do react favorably to the medium.

[20] Pilon, Thomas and Norris C. Craig (1988). "Disks-By-Mail: A New Survey Modality." Sawtooth Software Conference Proceedings. 387-396.

INTERNET SURVEY LIMITATIONS

As with any medium for conducting surveys, the Internet has its limitations. Among them are:

- **Abuse of the medium.** Electronic surveys are subject to the same misuses other data collection methodologies have experienced, as well as some misuses unique to the medium. In particular, some misuses include:
  - ◊ Over-burdening the respondent with a questionnaire that is too long
  - ◊ Excessive branching so that too few respondents get particular questions and data is meaningless
  - ◊ "Spamming" target respondents
  - ◊ Attempting to sell respondent something under the guise of a survey
- **Last minute changes.** Changes in questionnaire content and flow after a questionnaire has been programmed cost time and money, and introduce possibilities for error. This is becoming increasingly true in survey software designed to collect data via the Internet. Surveys designed for the Internet lend themselves to this form of abuse because changes can be made right up to the last minute.
- **Respect for the respondent.** Respondents value their time, and the researcher must provide the respondents with surveys that are professional in presentation, and, as with all surveys, ask important, relevant questions so that respondents do not feel that completing the survey is a waste of their time.

COST AND TIMING COMPARISONS

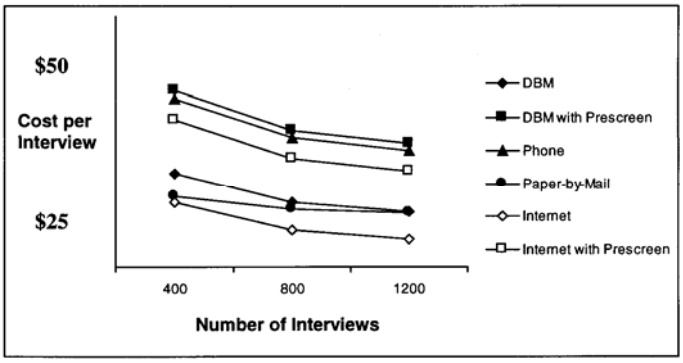
The low cost and quick turnaround that can be attained using the Internet are often drivers for choosing this modality. While surveys conducted this way can be turned around within one to seven days, we wanted to quantify the difference in cost to conduct a survey using this modality versus other, more traditional, alternatives.

All forms of electronic surveys are efficient for collecting complex data and for administering lengthy surveys. For comparison, the table below shows a per-interview cost comparison for a lengthy survey which could be administered by phone, paper-by-mail, disk-by-mail, or via the Internet. Data collection estimates are for a survey which would take 20 minutes by phone. Cost estimates are based on the following assumptions:

- For phone interviews: 1 completed interview per interviewer hour, programming the CATI
- For disk-by-mail interviews: a 40% response rate, $1 incentive, programming the disk-based survey, on the pre-screen option assume four completed screening interviews per interviewer hour
- For paper surveys: a 25% response rate, $1 incentive, 6-page (3-page duplex) survey, data entry of coded responses, but not verbatim responses on open ended questions

- For Internet surveys: a 40% response rate, screen saver incentive, the programming of the web-based survey, on pre-screen option assume four completed screening interviews per interviewer hour

Figure 10



In addition to the illustrated cost savings, the Internet is real-time, eliminating the transit time and mailroom processing time for a DBM study. Conducting a study over the Internet can easily be done within one week, and has been done in as little as one day. Whenever possible, we recommend leaving the survey open to respondents for one full week (including the weekend), because people get online at different times of day and different days of the week. The number of interviews to be completed does not extend the time required in field as it might a telephone survey, but rather simply increases the number of respondents to be solicited to participate.

BEST PRACTICES FOR INTERNET SURVEYS

To aid in applying the information, we have compiled the highlights for designing and managing a study using Internet-based data collection into the following key areas.

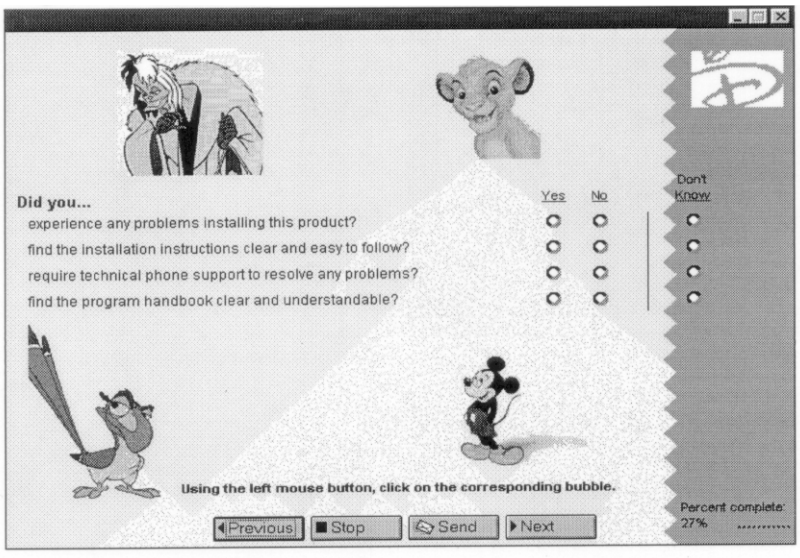
1. **Research sample.** In order to ensure that no biases are being introduced due to data collection methodology, the target audience should be considered:

   - All potential respondents must have access to the Internet (or have it somehow be provided to them)—surveys collected using this modality are only representative of the "wired" population
   - Evaluate whether the Internet and the browser software would be intimidating to any segments of the target population—this could result in non-response bias
   - Determine whether a pre-screen interview is necessary to identify the correct respondent, classify the respondent's market segment or quota group membership, confirm access to the Internet, and obtain the email address
   - Pre-notify (through phone, fax, email, letter, or postcard) of approaching survey to increase response rate

2. **Questionnaire design.** The decision to use an electronic survey may be driven by the objectives of the research. Once this decision has been made, the following steps will guide the execution.

   - Develop the questionnaire on paper, as usual, to provide an easy form of communication between the client and researcher, and to create questionnaire text which can be imported for use in the electronic survey software.
   - Finalize question types, question order, respondent instructions, display logic, and skip patterns before programming on disk.
   - Where possible, pre-test the survey on paper prior to programming, and then again once it has been programmed.
   - After all questionnaire changes have been made, access the survey with different brands of browsers in order to confirm the questionnaire code is entirely compatible with the leading brands (at a minimum Microsoft Internet Explorer, Netscape Navigator/Communicator, and America Online (AOL)'s browser). Individual users often set preferences within their browsers which will change the way that pieces of the survey appear. Surveys must be programmed for the lowest common denominator, and tested thoroughly with each browser.
   - When programming the survey using any graphical images, place them at the top or the bottom or the page, as shown below. Resist the urge to wrap text around images, as it will cause problems with some users' systems currently.

Figure 11



From a survey administration standpoint, the various browser standards are troublesome. However, it is a very real issue. Below are the percent of Internet users who use each of these brands of browsers:

Figure 12

Primary Web Browser Used	Q2 1997
Netscape Navigator	43%
Microsoft Internet Explorer	15%
AOL Browser	14%

3. **Survey programming.** To lessen the likelihood of respondents terminating during the course of the interview, and to enable them to provide accurate, actionable answers, the following guidelines are suggested when interviewing via the Internet:

   - The layout of the questions should be consistent, professional, and non-distracting from the content of the questions
   - Use appealing, parsimonious screen designs
   - Use survey software which is very easy for the respondent to use
   - Send the invitation to take the survey with a very professional email note
   - Use graphics, audio or video where appropriate without slowing it down
   - Ensure from pre-tests that all respondent instructions are clear

4. **Sending email messages/soliciting respondents.** To maximize the response rate and minimize non-response bias, we recommend the following steps:

   - Provide clear instructions to the respondent for each step in the process: how to navigate to your web site (if applicable), how to launch the survey and navigate within it, how to submit the completed survey, how to receive the offered incentive, and how to obtain help if needed
   - Communicate the benefits of participating in the survey
   - Describe the incentive(s) clearly
   - Include copyright notices as needed

5. **Fielding.** During the execution phase of the fielding, here are a few key recommendations:

   - Provide an 800 number for respondents to call toll free with questions about the web-based survey.
   - Provide an alternate online help method at your web site.

- Communicate the deadline for completing the survey—IntelliQuest recommends a maximum of one to two weeks for web-based surveys (preferably including a weekend).
- Set up enough bandwidth on the web server hosting the survey to enable the anticipated number of simultaneous users to respond to the survey. If bandwidth is limited, stagger the times at which the invitations to respondents are sent to minimize the number of busy signals respondents receive when trying to access your site. Monitor bandwidth and busy signals religiously. No access or slow response times can increase survey non-response.

6. **International.** There are many issues which are unique to international studies, in addition to those previously mentioned in this document:
   - Questionnaires should be translated to the language of the target country by a native speaker, and then reverse translated by a different party, normally in the same city as the client to confirm that it has been correctly translated. There must be final agreement by the native translator, the local translator, and the client to avoid confusion over instances where technology transcends local language or where local customs supersede linguistic tradition.
   - Questionnaires MUST be reviewed by someone familiar with the customs and peculiarities of the country, as well as the product category.
   - Legal requirements should be verified regarding obtaining lists of email addresses, collecting certain types of information (for example, demographics), and transmitting data to companies outside the country. This concept is best defined by the terms contracting in and out of research and is enforced to varying degrees both within regions (European Community) and even within target groups. For example medics in France can be contracted for research even if they contract out of research and list inclusion, however all lawyers in France are automatically contracted out of research and list inclusion.
   - When possible, provide respondents with a local, market numbers to call if they encounter problems with the survey. Sometimes this should be local regional rather than just local country. In particular this is true in Japan and Germany.
   - Incentives should be appropriate and legal for each country. In Japan handkerchiefs and tokens for books or telephone calling works very well. In Germany it is donations to environmental charities that are effective. Online versus traditional incentives should both be evaluated for their potential effectiveness within the target audience.
   - Translating is an evolving process. For on-going or tracking studies, translations should be continually reviewed—at least twice per year.
   - Oral translating is different from textual translating. A translation that is done for an electronic survey which will be read by the respondent will be slightly different from a translation prepared for a telephone interview where it will be spoken.

THE FUTURE OF INTERNET SURVEYING

The commercialization of the Internet has already left its mark on the marketing and market research communities. To understand fully its impact, many more studies must be conducted. IntelliQuest's belief is that over the next ten years, it can become the primary means of data collection among this audience. Having collected literally millions of electronic interviews over the past twelve years, IntelliQuest sees many opportunities for not just migrating the existing types of research to the "net", but also for creating new paradigms.

Despite this seemingly unflagging optimism, IntelliQuest believes there are many inappropriate uses of the Internet for surveying, today and in the future. There is also a great danger of overuse, with costs for surveying in this medium offering a low barrier to entry.

While many research companies will apply high ethical standards for their work in this area, the majority of the surveys currently being conducted via the Internet are done by individuals within companies who are typically not trained market researchers. Over use and general abuse will likely lead to a backlash of potential respondents, similar to that currently seen in the telephone arena.

APPENDIX A: EXAMPLE SERIES OF QUESTIONS DESIGNED TO CAPTURE INCIDENCE OF INTERNET USERS

I'm going to read a list of topics and we would like to know your level of familiarity with each. For each topic, please respond with "Never Heard of It," "Have Only Heard of It," or "Know What It Is." How familiar are you with <INSERT TOPIC>? Would you say you have never heard of it, have only heard of it, or know what it is? (topics are randomized)

Topics

(a) The Internet
(b) The television show Baywatch
(c) Digital television
(d) Ice beer
(e) The new drug Romazyne (PRONOUNCE: ROW-MA-ZEEN)

Response List

(a) Never heard of it
(b) Have only heard of it
(c) Know what it is
(d) Don't know

Do you or anyone else in your household currently own . . .
(Response list is randomized and is read. Multiple responses allowed.)

(a) An answering machine
(b) An automobile
(c) A bicycle
(d) A personal computer for use at home
(e) A device which attaches to a television and is used to access the Internet
(f) None

Do you or any members of your household subscribe to the following for use at home? (Response list is randomized and is read. Multiple responses allowed.)

(a) A daily newspaper
(b) A magazine
(c) Cable TV or satellite
(d) A computer online service or other Internet access service
(e) None

These questions serve to build rapport and obtain usage information which will qualify respondents.

COMMENT ON WITT

Edith De Leeuw
Department of Education
University of Amsterdam

1. INTRODUCTION

When I designed my first Disk by Mail survey, Witt & Bernstein's "Best Practices in Disk by Mail Surveys" was on my desk as a guide and checklist. I am certain that when I design my first Internet survey, Witt's paper on "Best Practices in Interviewing via the Internet" will prove to be a great checklist too.

Although one of the earliest experiments with an electronic survey through a network was done as early as 1983 at Carnegie Mellon University (Kiesler & Sproull, 1986), it took almost 15 years before Internet surveys became fashionable (cf. Ramos, Sedivi & Sweet, 1996). Among the early pioneers were the U.S. Navy, who in 1985 used Ci2 for a survey system on Bulletin Boards and diskette (Somer & Murphy 1989) and, of course, IntelliQuest who also in 1985 started electronic interviews over dedicated networks (Witt, 1997). The accumulated know-how is summarized in the 1997 Witt paper. I stated before that this paper will be a classic for everyone designing an Internet survey, and I do not have many critical comments. There are however two points on which I would like to exchange ideas. These are 'what makes an Internet survey special' and 'how to improve response'.

2. WHAT MAKES INTERNET SURVEYS SPECIAL?

Internet is a new medium that in its social codes is somewhere between written paper messages and spoken telephone messages. It is not as formal as paper mail and at the same time not as fleeting as a telephone conversation. When browsing the net, people have a short attention span, and usually spend only a limited time on each separate item. Therefore, Internet surveys should appear to be short and attractive as Witt stated. Another consequence of the short attention span is that one should keep an Internet survey as simple and clear as possible. Psychological research shows that time pressure or lack of concentration interferes with the perceptual and motor skills needed for reading the screen and typing in answers. Therefore a pretest, focusing on human-computer interaction, would greatly benefit data quality. Also, the full potential of electronic questionnaire design should be used to aid the respondent. This means no 'flat' website surveys with just a page that has to be scrolled, but an adequate design as in Computer Assisted Personal or Telephone Interviewing with automatic routings and internal error checking.

A growing concern with security on the 'net' is a second issue that should be met. To get adequate response rates and good quality data it is important that potential respondents can be assured of the confidentiality of their answers. For designers of surveys this implies maintenance of security, and respondent answers should at least be encrypted automatically; the de-facto standard is now Secure Socket Layer (SSL) at least for the US government. It also means that the respondent should be reassured about security, for instance with a short statement that the answers will be encrypted and that e-mail addresses will not be made available to others.

Finally, one should consider the respondent's costs. In traditional paper-and-pen mail surveys, and also in Disk by Mail the investigator pays for postage. In an Internet survey the *respondent* pays the telephone company and the Internet provider for the time spent online.

Witt suggests that compensating for costs could be an effective incentive, I will go even further and state that respondents should always be compensated for their online-costs. In addition one may add incentives (free screensaver, computer game) as Witt suggests.

3. How to improve response

In the past thirty years an impressive body of knowledge has been compiled on improving response rates for paper mail surveys (cf Heberlein & Baumgartner, 1978). In view of the similarities between traditional mail surveys and Internet surveys, much can be learned from this earlier work. Witt has translated many of the principles from mail surveys to electronic surveys. But we can go one step further. A well-researched and successful framework for mail surveys is the Total Design Method (TDM) of Dillman (1978, see also Clayton & Werking, 1996). According to the TDM, response rates can be maximized when the rewards for the respondents are maximized, the costs of responding are minimized, and that a feeling of trust is established between respondent and investigator. Costs and rewards can be both tangible and intangible. Intangible rewards are, for instance, respect for respondents' time, and making respondents feel that their time, effort and comments are valued, for instance by a thank-you message or a short summary of the results. Tangible rewards are incentives, which are most effective when they are promised in advance and sent immediately after the respondent has completed the questionnaire. The researcher can minimize the intangible costs of time and effort by making the questionnaire as simple, short, and attractive as possible (see also 2 above). Also the tangible out-of-pocket costs of the respondent should be reimbursed; in a paper survey, for instance, postage paid envelopes and a toll-free number for information. Finally, trust can be established by noting the affiliation of the survey organization, by disclosing sponsorship, and by guarantees of confidentiality. Clayton & Werking (1996) do not explicitly address the out-of pocket costs for the respondent to an Internet survey (e.g., costs for connect time). But they do translate many of the TDM recommendations to Internet data collection. An example is to use hypertext links to provide information for those who want it, thereby keeping the general appearance of the questionnaire short, simple, and attractive.

Paper mail surveys may seem boring and old-fashioned, but it took much effort to make mail surveys respectable and successful. We can learn from the past, use these ideas creatively, and transpose the old findings to fit the new technology!

4. The future of Internet surveying

Internet offers a great potential and new tools to present questions in ways impossible with paper mail surveys. Among them are links to background information about the survey, audio and video displays. Also, Internet surveys are very cost efficient.

Up till now only narrowly defined groups can be surveyed through the net. But provided that Internet becomes as widespread as the telephone net is at present, it will be a great tool for social science and marketing research. Until then, Internet surveys can be used for special populations

or form part of a multi-mode-multi-technology survey. One can survey parts of the population through Internet, parts through DBM, or CATI, or even paper-and-pen mail or FAX-surveys.

The multi-mode-multi-technology approach demands a well-defined sampling frame and in many cases a screening phase. It also demands research into potential mode and technology effects on data quality, to avoid an 'apples and oranges' effect. For software developers it poses new challenges. The ultimate multi-mode-multi-technology package will produce comparable questionnaires for distribution over the Internet, by Disk by Mail, CATI, CAPI and on paper forms.

There is also a second condition that should be met, one of security and ethics. Especially when sensitive information is asked, the respondent should be reassured about security. Lengthy reassurances only will make respondents shy and extremely aware of the potential risks involved. But the respondent should have the feeling that his answers are safe, and for instance encryption combined with an icon of a key should do the trick. Ethics is even more important. What happens with the bad guys who abuse trust? Both the postal system and the telephone system in most countries have laws protecting the confidentiality of the messages sent. At present the laws covering the Internet have yet to be made (cf. Clayton & Werking, 1996). In the meantime, the survey industry has to protect its name and develop a code of ethics, and at the same time develop means to convey to the respondents the legitimacy of their surveys.

REFERENCES

Clayton, R.L. & Werking, G.S. (1996). Business surveys of the future: The World Wide Web as a data collection method. Paper presented at the InterCASIC 1996 conference, San Antonio, Texas.

Dillman, D.A. (1978). Mail and telephone surveys; The total design method. New York: Wiley.

Heberlein, T.A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires; A quantitative analysis of the published literature. American Sociological Review, 43, 447-445.

Kiesler, S., & Sproull, L.S. (1986). Response effects in the electronic survey. Public Opinion Quarterly, 50, 402-413.

Ramos, M., Sedivi, B.M., & Sweet, E.M. (1996). Computerized self-administered questionnaires (CSAQs). Paper presented at the InterCASIC 1996 conference, San Antonio, Texas.

Somer, E.P. & Murphy, D.J. (1989). Computer interviewing applications in the Navy. Sawtooth 1989 conference proceedings. Sun Valley: Sawtooth

Witt, K.J. (1997). Best Practices in Interviewing Via the Internet, this volume, pages 15-34.

Witt, K.J., & Bernstein, S. (1992). Best practices in Disk-by-Mail surveys. Sawtooth Software Conference Proceedings, Sawtooth Software: Evanston, Illinois.

AN ALTERNATIVE APPROACH TO BRAND PRICE TRADE-OFF

Ray Poynter
Deux

INTRODUCTION

This paper reviews the traditional approach to Brand Price Trade-Off (BPTO) and develops an alternative procedure, which overcomes many of the problems that have been associated with the traditional BPTO methodology. For sake of clarity, I shall refer to this alternative methodology as Purchase Equilibrium Pricing (PEP). Equally, for the sake of clarity, I shall refer to the 'standard' BPTO as Trading-Up.

THE TRADITIONAL TRADING-UP APPROACH

Brand Price Trade-Off has been around for a considerable length of time. In 1972 Market Facts Inc published a paper on the technique, another article was published by Frank Jones in the Journal of Marketing (1975). A review by Chris Blamires (1981) places the introduction of this technique into Europe, from the USA, in the mid-1970s.

The exact methodology for the technique is flexible in its manifestation but the underlying algorithm remains constant. A very clear exposition of the standard Trading-Up approach is provided by Chris Balmires (1987). The technique is also covered in a range of handbooks such as Birn, Hague, and Vangelder (1990). Therefore, I shall confine myself, in this report, to presenting the key features.

Basic Trading-Up Methodology

A group of products (typically 4 to 8) is selected by the researcher. For each product a range of prices is defined and sorted from lowest to highest. An example is shown below:

Table 1

Product A	Product B	Product C	Product D
$2.50	$2.00	$2.50	$2.75
$2.75	$2.25	$2.75	$3.00
$3.00	$2.50	$3.00	$3.25
$3.25	$2.75	$3.25	$3.50

The respondent is then presented with each product, at its lowest price. The respondent selects the product they feel they are most likely to purchase at the prices shown. The lowest price for that product is removed and the next highest price is revealed. This process is then repeated, either a set number of times, or until all the prices for one product are chosen, or until all the prices for all of the products have been displayed and chosen.

The interview can be conducted in a range of different ways. The products can be shown as piles of cards with each card comprising a product name (or picture) and a price. The cards would then be set out as four piles with the lowest prices at the top. The interview can be conducted using a shelf situation with the prices adjusted as choices are made. Alternatively the interview can readily be computerised. The computerised versions have the advantages of being able to have flexible start prices and the use of percentage increments.

The analysis of the data can be as simple as determining the rank order preference, for each respondent, for the array of prices or products. Alternatively, the data can be arranged as a sequence of equations allowing the values to be estimated via linear regression. The regression approach has the advantage of permitting the utility of intermediate prices to be estimated.

Note: this methodology requires that prices start at their lowest level and, having made a decision, respondents must trade-up to a more expensive option. Hence, the use in this paper of the term Trading-Up.

It should be noted that it is not possible to simply reverse the technique, i.e. to start with prices at the most expensive and then reduce them in a sort of Dutch auction. The reversal of the process does not work because there is no simple way of choosing which product should have its price reduced.

THE SHORTCOMINGS OF TRADING-UP

The basic trading-up approach has a number of shortcomings. These include: the form of the interview, the need to produce fixed prices, and the need to determine the upper and lower limits of the prices in advance. However, the most significant shortcoming is that the technique can produce disappointing results (Richard Johnson & Kathleen Olberts). Despite these shortcomings “For many people, BPTO still remains the pinnacle of research pricing techniques.” (Pete Comley).

Some Reasons for Failure

In their paper, Johnson and Olberts point to several observations they had made about the trading-up technique. In particular they mention the way the interview appears to the respondent as a game. Some respondents appeared to treat it as an intelligence test, meticulously choosing the cheapest. Others appeared to treat it as a challenge to their brand loyalty, remaining loyal even when their brand was increased in price to an unreasonable level. Pete Comley also refers to the respondents’ tendency to ‘play the game’ and also to the problems of treating price as a conscious variable in a market where it does not function that way.

In addition to the reasons identified by Comley and Johnson and Olberts, a number of other problems can be readily identified. The trading-up approach starts at the minimum prices to be tested. In most studies this is a very unrealistic set of prices, for each brand individually and as a collective range. The expectations created in the mind of the respondent will be partially determined by these start prices. During the interview prices only go up. This monotonic increase contributes to the transparency of ‘the game’. Many respondents find it an insult to their intelligence to be asked questions in this mechanistic way. The structure of the interview, the prices and the changes, are such that the interview becomes pattern inducing.

Another area of concern in all pricing research is the degree to which respondents are able to answer direct questions about price and their likelihood to purchase specific products in a way that allows meaningful and useful results to be obtained. A paper by Rory Morgan (1987), discusses 41 key issues which can determine whether the research is likely to be effective. These issues include points such as 'Repertoire versus single brand', 'Is the purchase a gift?', and 'Low involvement'.

Our experience has been that direct questioning works best when: the product is easy to envision, prices are accepted as being a feature in the purchase decision, and respondents have some awareness of the relative prices of products. Conversely, direct price-questioning works less well where the product is hard to envision, where prices are not thought (by consumers) to be an issue, or where consumers are not familiar with prices. We have also found that price research is more challenging when it is attempting to assess, from a laboratory situation, the likely long-term effect of a price change on a very frequent purchase. For example the effect of a 1% increase in the respondent's daily train, bus, taxi, or newspaper choice tends to be harder to assess when compared with FMCG research. This last problem is not unique to pricing research. The same problems face product and advertising tests, where the daily ongoing impact is being assessed from the short-term reaction to stimulus.

In a text-based interview, such as traditional CAPI, a product may be hard to envision because it is complicated or because its appeal relies on its appearance. For example, if the product relies for much of its business on the appeal of its pack (for example luxury biscuits attracting casual or gift purchases) then it is unlikely this will be captured by text in the interview. If the product is complicated, for example if the product is the full specification for a car purchase, it is unlikely that the respondent will be able to assess both the prices and the descriptors. It should be noted that graphical systems can remove some of these envisioning problems, for example in the case of a luxury biscuit.

Respondents may be reluctant to rationalize price as a factor in a number of situations. Typical problem areas are ones where:

Price differences are very small.

The products are value loaded.

The prices are low in absolute terms.

The respondent believes they do not know the price of other competitive products.

In these situations we have found direct questioning to often be of limited value. A quality graphical image, or constructed shelf situation, can sometimes create a more 'real' situation and better represent the shopping experience.

DEVELOPING AN ALTERNATIVE APPROACH

To help highlight the approach being developed, a simple 5 product example is used. It should be noted that in real studies things are seldom as straightforward as this example:

Realistic Start Prices

Our first step, in attempting to improve the BPTO algorithm, was to move the starting point. We felt the best place to start the process was with product shown at realistic prices.

There is an old saying, in the UK, about a villager who when asked by a tourist "What is the best way to get to xxxx." replied "Oh, if I was going there, I wouldn't start from here!". Likewise if we want to assess the current price values for products, we shouldn't start from the lowest levels under consideration.

The initial prices can be set in a variety of ways. They may represent the client's view of the current market price. They may reflect the respondents' views about the prices they expect in their locality, and in their regular store. Alternatively, the prices can be formed in a more exotic way. For example, in a health insurance study the start prices may be determined, arithmetically, by using the answer to questions about age, profession, habits, and health.

Table 2

	It 0
Product A	2.50
Product B	3.00
Product C	2.75
Product D	3.15
Product E	2.50

The table above shows an example set of start prices. In terms of our methodology this is referred to as Iteration zero.

Selecting the Price Adjustments

In the PEP algorithm, all of the products change in price after each respondent selection. The selected product is increased in price, just as in the trading-up methodology. The products that were not selected all have their prices reduced.

The choice about how much the price of a product moves up or down in price is critical and has to be made very carefully. Since more products will move down than up, the system would soon become unbalanced if the price increases and decreases were equal in size. The response to this is to ensure that a suitable, larger figure is selected for the increase, compared with the decrease.

Table 3

Visible	**It 0**	**It 1**
Product A	2.50	2.45
Product B	3.00	3.15
Product C	2.75	2.70
Product D	3.15	3.09
Product E	2.50	2.45
Selection		
Product A	0	
Product B	1	
Product C	0	
Product D	0	
Product E	0	
Increment		
Product A	-3.0%	
Product B	5.0%	
Product C	-3.0%	
Product D	-3.0%	
Product E	-3.0%	

The table shows a possible scenario. The prices for Iteration 1 are determined by the increments set out in Iteration 0. Product B was selected, its price is accordingly increased by 5%. The price of the other four products is decreased by 3%.

The Purchase Equilibrium Price

The object of the interview is to determine the Purchase Equilibrium Price for each of the brands. The Purchase Equilibrium Price is the price, for each brand, where the respondent would be equally likely to select each of the brands. Given enough iterations the system finds this point for all of the brands.

In practice, we have found that we can halt the process after about 10 to 15 iterations and have a reasonable estimate for the brands under consideration. Two things affect the number of iterations needed. The first is the number of brands being tested. If there are more brands, then the number of iterations needs to be higher. The number of iterations also governs the maximum and minimum prices. If there are 10 iterations then the maximum price any product can achieve is 150% of the start price (assuming an increment of 5%). Likewise, with 10 iterations the lowest price any product can reach is 70% of the start price (assuming a decrement of 3%). The price estimate for brands that would not be purchased, tend to be over-estimates. If there had been more iterations the product could have been rejected at lower prices. However, these estimates should normally be adequate in conveying the message that the product would not normally be purchased by this respondent at any price which is likely to be offered.

Refining the System

A further refinement to the process is to modify the increments and decrements based on earlier selections. If a product has been rejected at every stage then it is reasonable for the price decrement to stay at the same level. Equally, if a product has been selected at every stage then it is reasonable for the price increments to stay at the same level. However, if a product has been selected on some occasions and rejected on others, its price is approaching its equilibrium point. In these cases we normally reduce the size of its increments and/or decrements.

By comparison with the traditional trading-up approach, the PEP algorithm creates fewer patterns in the responses elicited from the respondent. For example, when all but one of the products becomes cheaper the respondent can only select one of these – removing one of the simple patterns we have observed in the trading-up approach.

The pattern problem can be further addressed by removing, from each choice set, one of the products that would have been on offer. This option can be selected on a random basis and noticeably reduces the respondent's ability to form patterns in their responses.

The table below shows a typical set of iterations, incorporating variable increments/ decrements and blank options.

Table 4

Visible	**It 0**	**It 1**	**It 2**	**It 3**	**It 4**
Product A	2.50		2.43	2.35	2.28
Product B	3.00	3.15		3.30	3.23
Product C	2.75	2.67	2.59	2.71	
Product D	3.15	3.06	2.96	2.87	3.01
Product E	2.50	2.43	2.35		2.28
Selection					
Product A	0		0	0	0
Product B	1	1		0	1
Product C	0	0	1	0	
Product D	0	0	0	1	0
Product E	0	0	0		0
Increment					
Product A	-3.0%	0.0%	-3.0%	-3.0%	
Product B	5.0%	5.0%	0.0%	-2.5%	
Product C	-3.0%	-3.0%	4.5%	-2.0%	
Product D	-3.0%	-3.0%	-3.0%	4.5%	
Product E	-3.0%	-3.0%	-3.0%	0.0%	

At each iteration, the prices are determined as being the prices in Iteration zero multiplied by the sum of the increments plus 100%. If a product was not displayed then its accumulated increment/decrement is not affected by that iteration.

Estimating the Purchase Equilibrium Prices

The starting point for the price estimates is the price set the products reach at the end of the iterations. However, as the chart of a typical response patterns shows, Chart 1, the iterative nature of the process means that, the last iteration is quite likely to produce an estimate above or below the true iterative endpoint. A good method of improving on the estimate is to use regression to estimate the nth price, where n is the number of iterations. An example of this process is shown in table 5.

Chart 1

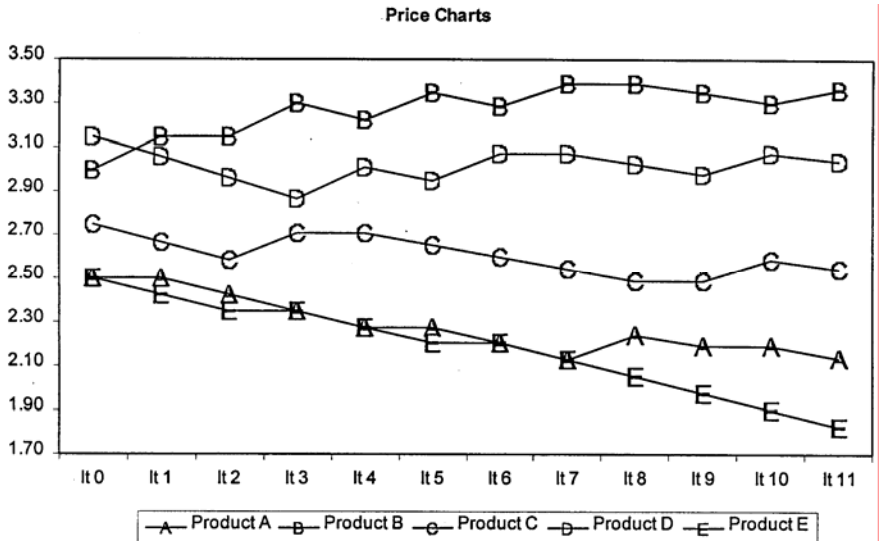


Table 5

	Iteration 11	Trend Estimate
Product A	2.14	2.15
Product B	3.36	3.35
Product C	2.54	2.51
Product D	3.04	3.05
Product E	1.83	1.84

The trend estimate was formed using linear regression and the results of the last seven iterations. The seven iterations were fed in and regression used to estimate what the last iteration should have been, thereby removing any 'wobble' in the iterative process.

Implementing a Study

One key disadvantage of the PEP technique is that it removes the possibility of the study being conducted using paper and pencil. We have found that it is necessary to implement the PEP studies utilizing Computer Assisted Interviewing. All of the steps outlined in this paper can be

coded using Ci3. However, it is worth noting that the basic algorithm could be programmed into any CAPI system that incorporates variables and arithmetic.

What Do the Estimates Mean and How Are They Used?

PEP produces price estimates for each respondent and for each brand. The price estimates that are produced are Purchase Equilibrium Prices. That is, prices at which each of the products would be equally preferred.

It should be noted that there is not a unique equilibrium set for each respondent. For example, if prices P1, P2, P3, and P4 were an equilibrium set for a respondent for products B1, B2, B3,
and B4, then so would be the set P1*101%, P2*101%, P3*101%, and P4*101% for the same 4 products.

The system has been used with a number of objectives. These objectives have included: producing a price positioning for a new product, estimating brand or option values, estimating the values of conjoint features (see note below). However, the most frequent use of PEP has been to provide the input to a Brand/Price simulation model.

Typically the PEP estimates are used in a simulation model to estimate brand choices for any given set of price levels. The utility for a product, for a respondent, is determined by dividing the price estimate by the model price. For example if a respondent values product A at $3 and Product B at $4, then if both products are assumed to be on sale for $3 we see the utility, to this respondent, of A is 100% but the utility for B is 133%. Assuming a first choice model we would therefore assume the respondent would purchase product B.

A simple model can readily be created by extending the calculation of product utility to a range of products and for all of the respondents. By counting all of the first choices, for all of the brands, a simple and flexible first choice model can be constructed. Models of this type can be implemented using spreadsheets, such as Excel. Models constructed in this way are both powerful and open (in the sense that users can access all of the model's workings). A useful introduction to this type of modeling is provided in a paper by Ray Poynter (1996).

Using PEP with Conjoint Analysis

Much has been written on the problems of using price in a conjoint study, for example Dirk Huisman (1992). PEP can be used, in conjunction with conjoint analysis, to obviate this problem. One option is to conduct the conjoint analysis without price. At the end of a conjoint interview the respondent can then conduct a PEP. The 'products' used in the PEP should be options constructed from specific attributes and levels within the conjoint.

At analysis time it is possible to calculate what each of the options in the PEP is worth in terms of utility points and also in terms of money. By combining this information it is possible to create a price variable which is scaled to reflect the utility values in the conjoint study.

Experience to Date

To date this technique has been used on over 30 studies. The product areas have included: consumer durables, financial services, transport, alcoholic beverages, telecommunications, agrochemicals, FMCGs, and pharmaceuticals. The reactions of clients, particularly in comparison

with the Trading-up technique, have been very favorable. In those cases where external data exist the results have produced a good fit with other data.

CONCLUSIONS

The BPTO methodology is still popular and widely used, despite suffering from a range of defects that have caused many respected practitioners to shun it. The techniques outlined in this paper tackle many of the fundamental weaknesses in the BPTO methodology. However, this is achieved at the expense of making the technique dependent on CAPI.

Good pricing research demands two things. Firstly, the correct technique has to be selected for each individual situation. Secondly, the interview has to be constructed in such a way that it is possible, and likely, that the respondent will answer the questions in a way that is helpful to the research.

The Purchase Equilibrium Pricing technique provides an extra option to the researcher, when the researcher has to select an appropriate tool for a specific problem.

REFERENCES

Chris Blamires (July 1981), "Pricing Research Techniques: a Review and a New Approach," Journal of the Market Research Society

Chris Blamires (April 1987), "'Trade-Off' Pricing Research: a Discussion of Historical and Innovatory Applications," Journal of the Market Research Society

Robin Birn, Paul Hague, and Phyllis Vangelder (1990), "A Handbook of Market Research Techniques," Kogan Page

Pete Comley (January 1997), "Pricing Research", Admap

Dirk Huisman (July 1992), "Price-sensitivity Measurement of Multi-attribute products," Sawtooth Software Conference Proceedings

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

Richard Johnson and Kathleen Olberts (1996), "Using Conjoint Analysis in Pricing Studies: Is One Pricing Variable Enough?", Sawtooth Software Technical Paper

Frank Jones (1975), "A Survey Technique to Measure Demand under Various Pricing Strategies," Journal of Marketing

Rory Morgan (April 1987), "Ad Hoc Pricing Research—Some Key Issues," Journal of the Market Research Society

Ray Poynter (September 1996), "Open the Box or Take the Money?", Survey and Statistical Computing 1996 (Association for Survey Computing)

CREATING END-USER VALUE WITH MULTI-MEDIA INTERVIEWING SYSTEMS

Dirk Huisman
SKIM Group

DEVELOPMENT OF MULTI-MEDIA INTERVIEWING

In a traditional computer interview, mainly text-based and with a few images on hand-outs, the researcher seems in control of the interview: he can randomize, he can create a balanced research design, and when analyzing the interview he will be able to trace and link the response and the feature/value that triggered the response. But once you start to lard your interview with sound and digitized visuals (pictures, video, animation, three-dimensional imaging, 360-degree views), identification of the feature/value that triggered the response becomes a quite complex task. Likewise, randomization (of all these new stimuli) stops being the relatively simple task that it used to be.

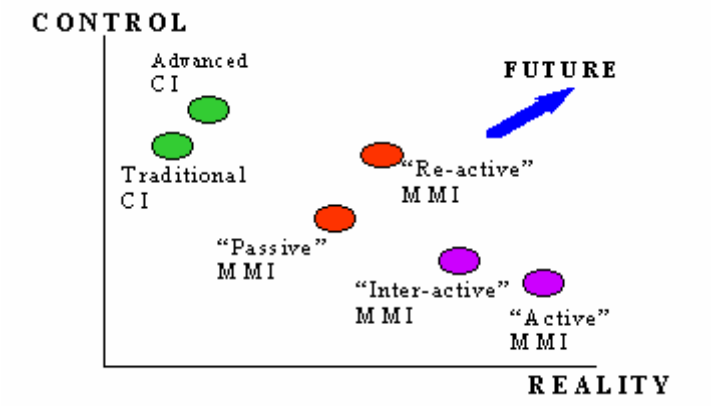
Based on the degree of interaction between the interviewee and the number of stimuli he/she is exposed to, we can distinguish the following levels of multi-media interviewing:

- "Passive." The researcher has full control over the combination of stimuli he exposes the respondent to, and records the reaction to the combination of features. The next combination of stimuli he presents to the respondent is not dependent on the response to the previous combination of stimuli.
- "Reactive." The combination of stimuli presented to the respondent may be based on the response to a previous set of stimuli. The researcher still has full control of the combination of stimuli he presents. But because not all combinations of stimuli will be realistic, possible or imaginable, this will be a complex task, and the researcher will have to make a trade-off : will he go for complexity and the relatively high costs of an interview with all combinations (the unbiased research design)? or will he save costs and accept an identifiable bias?
- "Active." The respondent can walk through a mall, look into a car, take a product from a shelf to observe all sides, and is invited to react on the basis of these experiences. The researcher is still in full control: he created a database with many combinations of stimuli and a large decision tree with many decision paths that the respondent can follow. Winding his way, the respondent decides without explicitly specifying his response. At a certain point he explicitly specifies what he wants and the next path (e.g. an animation) is activated. All paths are predefined.
- "Interactive." The respondent is able to interact with the stimuli. Based on his reactions the stimuli are adapted. The researcher is still in control: he defined the decision rules (or algorithms) that decide on the next combination of stimuli to be shown. But not all the paths have been predefined.

The higher the degree of action or interaction, the more difficult it is to keep in control. Therefore today most applications of multi-media interviewing are “passive” and “reactive”.

The information technologies underlying the multi-media interviewing systems do not only provide control of the stimuli to which the respondent can be exposed, but also provide better control options in recording the respondent’s reactions. At this stage of development this is primarily full audio control and control of the “mouse” or the movements the respondents makes. However, the technology enables the researcher to capture many more reactions from the interviewee (measuring tension or stress, eye-ball tracking, attention measurement).

Traditional computer-assisted interviewing systems like Ci2 offered a high level of control but, being text-based, the interviewee sometimes had to strain his imaginative faculties. With the multi-media systems the market researcher can create lifelike interviews, but on the road to this level of reality interview control can grow extremely complex. Accordingly, the development of the multi-media interviewing systems can best be described along two dimensions: the control dimension and the “reality” dimension, as in the following chart:



MULTI-MEDIA INTERVIEWING WITH ACA

Unfortunately the information regarding the impact of multi-media interviewing in practice is mainly casuistic and hardly ever based on surveys or comparative studies. The information is captured in case studies presented at the ARF and ESOMAR as well as in the case descriptions provided by the developers of the multi-media systems. In line with this type of information, users of multi-media interviewing systems are thrilled. The enthusiasm expressed in the cases is centered around two benefits: *Doing what we could not do before* and *It is more appealing and more realistic*.

“Doing what we could not do before” regards, in the first place, using information and stimuli that could not be used before, like, for instance, measuring the impact of tire profile and tire sound (different sounds you hear in a car when different tires are installed) on tire preference, and rotating commercials in an ad test. The claim that it is “more appealing and more realistic” is derived from the spontaneous reactions of interviewees, interviewers and end-users. But all this is verbal praise, and so far hard data are apparently not available.

SKIM Analytical recently conducted a multi-media survey in the USA. In this survey we tried to measure if visualization of a part of the (ACA) attributes would influence the utility of these attributes. Of a total of 360 respondents, 145 conducted an Adaptive Conjoint Analysis using the Sensus Trade-Off system (a Windows application built around ACA), and 215 conducted the traditional ACA. We arrived at the following conclusions:

- because visualization adds information, it influences the sensitivity to the attributes and the utility of the attributes (and in this study the visuals shown on the screen were only "reminders" because, preceding ACA in both samples, all respondents had been shown the visuals);
- visualization does not by definition increase the utility of the visualized attributes;
- the effective interview time in both samples was identical; so, in this case visualization did not lead to shorter interviews;
- there was a slight difference in product preference between the two samples, but because we did not ask to choose from a number of (hold-out) product concepts we can not conclude which method leads to better predictions.

Example 1: The impact of visualizing the attributes in ACA

To test the impact of the visualization in the trade-off process about 40% of the sample (145) conducted an Adaptive Conjoint Analysis using the Sensus Trade-Off system (a Windows application built around ACA) and about 60% (215) conducted the traditional ACA. In total 15 attributes and 45 attribute levels were traded off. Preceding the ACA module in the interview all interviewees were shown (in Ci3 for Windows) the 15 attributes and 45 attribute levels and had reacted to a number of questions regarding the 15 product features. Consequently, the differences found do not reflect the effect of visualization in general but only the impact of the visualization of the features in the ranking and the trade-off process.

The 15 attributes can be classified as follows:

- did not have any visual connotation, e.g. "speed of the meter". These were visualized by symbols and the core difference (number of seconds) was specified in large letters;
- for 2 attributes the visualization was purely dimensional in nature ("size of the meter" and "size of blood drop");
- for 4 attributes the visualization reflected a specific application or benefit as well as a dimensional element (for instance "type of battery" reflects the duration of the battery and the standard avail-ability as well as the size of the battery);
- for 4 attributes the visualization reflected a specific application or benefit of the meter without a dimensional element.

Analyzing the impact of the visualization we compared the average utility values of the 45 attribute levels. In total the average utility value of 7 attribute levels (from 6 attributes) differed 10 or more points.

It is primarily the dimensional element that causes the difference in sensitivity.

Not one of the attributes which did not have any visual connotation differed. Of the four attributes for which the visualization reflected only the application without a dimensional element, only one attribute differed, but this was the attribute from which two levels differed more than 10 points. Clearly

the visualization of this attribute showed the application and, probably, the perceived benefit better than the phrasing of the application did. Of the four attributes for which the visualization reflected a dimensional element as well as the application, three attribute levels differed more than 10 points. And finally, for both attributes of which the visualization was purely dimensional the preferred level differed more than 10 points.

Of the seven attribute levels which differed more than 10 points, four were more positive in the Sensus Trade-Off version and three in the textual ACA conjoint. So, visualization does not always make a feature more attractive. For instance, the importance of the difference in "size of the meter" when visualized was less than when specified in millimeters or inches. On the other hand, when visualized the "size of blood drop" was more important than when specified in micro-liters.

The relative importance of the attributes without visual elements (but shown as symbols) was not affected. This is important because otherwise it might have been purely the visualization, independent of the content, which influenced the importance of a feature.

In addition to the analysis of the utility values we also compared the length of the interview. There was no significant difference at all between the textual ACA and the Sensus TradeOff version. Consequently, the hypothesis that the interview runs faster when the attributes are visualized had to be rejected.

Finally, for the whole interview (Ci3 for Windows + Sensus TradeOff / ACA) we compared the stated interview length and the real duration of the interview. The perceived interview length was 45 minutes, and was significantly less than the recorded interview length, which was 58 minutes. We have not measured a difference of 22% before. Based on our experiences in Europe and on other studies regarding the interview length[1], we know that the difference between the perceived interview length and the actual interview length is a function of the complexity of the response task and the interest and involvement in the subject of the study. The target group in our US example is involved in the subject, which may partly explain the difference, but previous studies among the same target group in Europe never generated a difference in stated versus real interview length of more than 10%. Consequently, at least half the difference may be attributed to the use of multi-media.

MULTI-MEDIA INTERVIEWING WITH CBC: HEADING FOR THE VIRTUAL STORE

The fact that multi-media interviewing today is primarily "passive" and "reactive", while the multiple stimuli are exposed as one fixed set, reduces the complexity and the problems in practice. The problems faced are primarily hardware problems (no standard configurations and the configurations do not fit with the specification required) and problems related to a lack of experience.

If multi-media interviewing will stay as it is today, passive or reactive, we may do things we could not do before and it may be more enjoyable for the interviewee, but the interviewee will not really be taken up in the interview and position himself in the virtual real world. Neither shall we meet the requirements of the end user, because we are less flexible in the survey design. What is needed is that, during the interview, from the beginning to the end, multi-media are used, so the interviewee will familiarize himself with the environment. At the same time, the interview must include methods that enable us to collect data to answer the strategic questions.

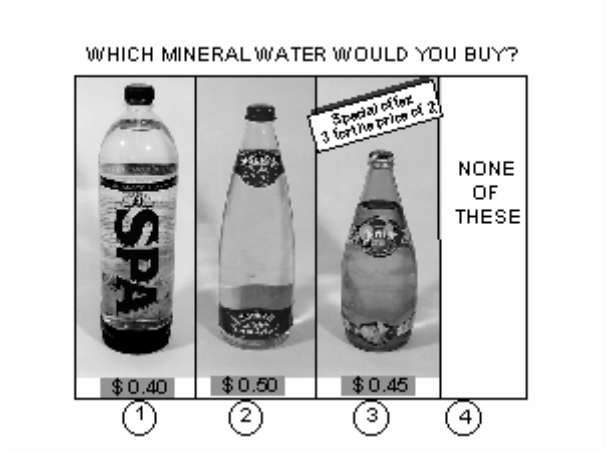
Today it is possible to create a virtual environment, like the "Virtual Store" of the Harvard Business School's Marketing Simulation Lab[2]. Building the model (creating a store in which the

[1] Yearbook of the Dutch Society of Market Research, 1980.
[2] R.R. Burke. Virtual Shopping: Breakthrough in Marketing Research, Harvard Business Review - April 1996

three dimensions are defined) is a complex task and requires special programs and computers. Once the model is built, it is relatively easy to place the products (data files capturing a 3D visualization of the product) on the shelves. Because products, merchandising material, packagings and advertisements are more and more available in electronic form and stored in data bases, it will become relatively easy to furnish the store. The virtual store is operational and used to test new strategies.

According to R.R. Burke sales measured in the virtual store correlate with actual sales. I predict that the virtual store will be marketing research practice in the near future: new strategies can be tested in a virtually real environment and all kinds of interactions can be measured, provoked and simulated. The use of the virtual shop at this moment is to be typified as a controlled experiment. After a number of virtual purchasing events the new stimulus, for instance the new product or promotion to be tested, is placed in the shop and effects are measured. Reality is simulated in the purchasing situation and the market simulation is directly derived from the behavior in the virtual store.

However, the virtual store is not common market research practice yet, so we will have to simulate with the tools and techniques available. Visualization of the choice tasks in Choice-Based Conjoint is the first step to the virtual store.



The advantage of this approach over the virtual shop is that, underlying the visualization of the choice task, there is a model based on attributes which enables us to trace for the different classes of customers the impact of the different product characteristics on the choices made. The results of a survey regarding price and promotion sensitivity of mineral water have been compared with an analysis based on Nielsen Data, and the effects measured were identical. This leads to the conclusion that the model can be used to forecast effects ex-ante. However, to create this simple multi-media choice model a number of hurdles had to be jumped (see example).

[4]

Example 2: Measuring price-sensitivity using Choice-Based Conjoint and multi-media

To measure price-sensitivity a choice model was designed including price, brand, packaging (size and material) and type of mineral water. The interviewee had to complete 3 x 6 choice tasks. The bottles to choose from were positioned on a shelf on the screen. At the bottom of the shelf we showed the price tab.

In this simple model there were two major problems. Packaging is partly brand specific, so either we had to make a number of exclusions or we had to visualize non-existing packaging brand combinations. We did both. The other problem we encountered is that each bottle has its specific width and height, so we had to show all bottles in proportion. Consequently, each "picture" required a specific space on the screen, but we also wanted to randomize the bottles on the shelf without the wide bottles overlapping the thin bottles.

To show the bottles as realistically as possible, we had to define an acceptable resolution, which is memory intensive. Because each bottle is captured in a file we needed quite some memory and hard disk space. Compared with BPTO surveys, the price-sensitivity measured for the total sample was less. This was due to the fact that there was a high brand-loyalty and the loyalists ignored prices.

The limitation of this approach is that we assume no impact from other product categories (e.g. ice tea) on the product choice and price-sensitivity. So, to make it more realistic, we now try to place the shelf with the choice task between other shelves with competing products. For two reasons we are not able to include these competing products in the wider choice task. First because there will be too many products and attributes to deal with in the choice model, and secondly, if each competing product would be captured in a file we would run out of memory. The solution to this is to make a file of the "environment" and identify which spot in the environment the interviewee selects.

Using small price tabs, as in reality, we found a limited price-sensitivity. But in reality there are, in addition to the price tab, price promotions, rebate posters and promotion stickers. So, with the help of multi-media we simply placed rebate stickers on the shelf and randomized over the brands. But using latent class analysis it became clear that only a specific group reacted to the rebate stickers: the stickers were not convincing enough to pass the receptor barrier of most of the respondents. But the total price-sensitivity measured increased.

What we will try to do next is to add in-store 3D price promotions instead of 2D stickers. And that's where we start to meet the virtual store.

Compared with what we did in the past we now communicate visually from the beginning until the end of the interview; other products and image criteria are also visualized. The whole flow of the interview is more natural and almost self-completion, whereas in the past, in this example, either BPTO questions with real bottles or verbalized choice tasks with real bottles on the desk were asked. It has become much easier to add other stimuli without specifically drawing the attention of the interviewee to these stimuli.

We experienced that to learn the opportunities and limitations of the multi-media systems, creating the multi-media stimuli yourself is a beneficial experience. But there are specialized 2D and 3D design agencies and increasingly often the data files are already available. Consequently, it may be better to concentrate on the design of the survey and the simulation model, and to subcontract the visualization.

At this stage of development multi-media interviewing is still more expensive than traditional computer-assisted interviewing. In our mineral water case, the price was 25% above a comparable survey without the visualizations. All interviews were studio interviews or mall intercept interviews using 17" screens and multi-media desk tops with 16MB internal memory.

Based on the requests we receive and based on our experience with multi-media interviewing, most opportunities of MMI regard fast moving consumer goods. The reason for this is that, for these products, habits and subtle differences play an essential role in the choice process and in the marketing of these products. With multi-media interviewing we are better able to communicate these subtle differences and to trace these habits while the interviewees react more naturally.

A second industry which is very interested in multi-media interviewing is the service industry. The needs for and opportunities of multi-media interviewing in these industries are obvious, but it is very hard to define the stimuli because they are intangible.

Both the FMCG and the service industries are interested in multi-media interviewing primarily because they want to "observe" and measure the impact on behavior or preference without asking for rationalized answers.

In contrast with these two categories of users interested in multi-media interviewing, a third category is interested because information can be provided which could not be provided before or only at high costs: for instance, in the case of the design of the interior of planes, new cars and new buildings. Market research can meet the needs of this last category by conducting MM interviews.

NEED FOR RESEARCH

As indicated, multi-media interviewing is only in its infancy, the opportunities for market research to speed up and get in line with its changing environment are there. But that requires that the market research industry is able to change the application of multi-media interviewing from passive and reactive to active and interactive. The hypothesis and the observation is that, by using multi-media, the interviewee reacts more naturally and we can trace which subtle differences trigger the response or add value for the end-user. The information available to prove this is casual or based on information collected in addition to market surveys. To strengthen its position and to have a greater impact on strategic decisions, there is a clear need for research that shows the potential, the real benefits and impact of multi-media interviewing for the end-users of market research.

COMMENT ON HUISMAN

Karlan Witt
IntelliQuest, Inc.

Dirk has shared in this paper some valuable research in an area that clearly needs closer examination. He addresses an issue that among selected audiences is becoming an ever-increasing issue: a receptor barrier. PC users overall, and Internet users in particular, are confronted continuously with the latest graphics, style, and content available electronically. A survey that offers much less can fail to hold respondent's interest, introducing a potential bias into survey results. However, I will disagree with Dirk that multi-media interviewing is a goal toward which we should all blindly march.

I believe there is a place for all the forms of interviewing Dirk describes (Passive, Reactive, Active, and Interactive). Further, I believe that over-using multi-media in a survey can detract as much, if not more, than it adds. Multi-media should be used appropriately, and with purpose, not just for the sake of using it.

In this paper Dirk desI'm attachingribes what seems to be a very appropriate use of multi-media as a shopping model. Computerized shopping models offer great insight into the purchase process. Shoppers gain a better understanding of the various attributes being studied, and can simulate their search through a variety of sources of information used during the shopping process. There is little published, however, about how to fully leverage the non-compensatory data collected in this type of shopping model.

Overall I am thrilled to see work being shared in this area, and would call for additional efforts to make this a mainstream survey research option.

A COMPARISON OF FULL- AND PARTIAL-PROFILE BEST/WORST CONJOINT ANALYSIS

Keith Chrzan and Ritha Fellerman
IntelliQuest, Inc.

BACKGROUND

Traditional full-profile conjoint analysis requires respondents to evaluate stimuli that specify a level for each attribute in a study. In a study with many attributes this could overwhelm respondents, diminishing their ability to take all attributes adequately into account. Biased or otherwise less reliable parameters may result. One solution is to use "partial profile" stimuli. A given partial profile conjoint question specifies levels for only a subset of the total number attributes in a study. This is the strategy in the "pairs" section of an ACA interview (Sawtooth Software 1993).

It is also the strategy used in partial profile choice-based conjoint analysis. Empirical comparisons built into four recent commercial studies showed partial profile conjoint stimuli to outperform full profile conjoint stimuli (Chrzan and Elrod 1995, Chrzan, Bunch and Lockhart 1996). In each case, full and partial profile models had parameters that were equivalent up to a multiplicative scaling constant. The partial profile parameters, however, contained less unexplained error variance than did the full profile models' parameters, making for greater efficiency. The efficiency gains more than offset inefficiency due to non-orthogonal design matrices that partial profile experiments sometimes employ.

Best/Worst Conjoint Analysis

The present study extends this comparison of full and partial profile model parameters to an innovative conjoint technique called "maximum difference" or "best/worst" conjoint analysis (Swait, Louviere and Anderson 1995). In best/worst conjoint analysis utilities are derived *via* multinomial logit modeling of respondents' choices of best and worst levels from each of several questions like this one:

Exhibit 1. Sample Best/Worst Conjoint Question

Please pick the one statement that **most** makes you want to buy stereo A and the one statement that **least** makes you want to buy Stereo A.

Stereo A	**Most**	**Least**
Sony	[ ]	[ ]
100 watts of power	[ ]	[ ]
6 disk CD changer	[ ]	[ ]
Single, auto-reverse cassette deck	[ ]	[ ]
No graphic equalizer	[ ]	[ ]
3 way speakers	[ ]	[ ]
Rack design	[ ]	[ ]
Dolby B noise reduction	[ ]	[ ]
Surround sound	[ ]	[ ]
High-speed synch dubbing	[ ]	[ ]
Price: $699	[ ]	[ ]

Note that the respondent's task is choice of one most and one least liked attribute level from a single product profile. This differs from both ACA and partial profile choice-based conjoint analysis wherein the respondent task is choice (or preference) of one product profile from a set of two or more such profiles.

Best/worst conjoint analysis has some limitations relative to choice-based conjoint analysis: designed for "plain vanilla" modeling, one cannot use best/worst conjoint for modeling interactions, brand-specific effects or disproportionate cannibalization. Still, best/worst conjoint analysis has some valuable advantages. First, best/worst conjoint produces rough utility estimates at the individual respondent level, allowing one to model respondent heterogeneity through segmentation analysis. Another advantage is that best/worst model parameters contain less unexplained error than do the parameters of a choice-based conjoint model (i.e. they are more reliable). The most unique strength of best/worst conjoint analysis, however, is that it eliminates the arbitrariness of the scale origins of the individual attributes: in all other conjoint estimation procedures, levels cannot be compared across attributes because the different attributes have different (arbitrary) origins (Johnson, Shocker and Wittink 1991). Best/worst conjoint analysis includes the attributes' origins among the parameters estimated by the model so that all attributes' levels can be put on a common scale and thus compared directly.

Swait, Louviere and Anderson (1995) describe experimental designs for best/worst conjoint analysis stimuli. They note that partial profile designs may offer a simpler task for respondents and they describe experimental design strategies for partial profile best/worst conjoint experiments. A partial profile best/worst question for the stereo study might look like this:

Exhibit 2. Sample Partial Profile Best/Worst Conjoint Question

Please pick the one statement that **most** makes you want to buy stereo A and the one statement that **least** makes you want to buy Stereo A.

Stereo A	**Most**	**Least**
Sony	[ ]	[ ]
100 watts of power	[ ]	[ ]
Rack design	[ ]	[ ]
Dolby B noise reduction	[ ]	[ ]
Surround sound	[ ]	[ ]
Price: $699	[ ]	[ ]

In this case the best/worst question contains only six of the 11 attributes. Other partial profile questions would include different subsets of the 11 attributes. The composition of attribute subsets varies according to an appropriate experimental design (e.g., Lazari and Anderson 1994).

STATISTICAL EVALUATION OF FULL- VERSUS PARTIAL-PROFILE BEST/WORST CONJOINT ANALYSIS

We draw on tests of two hypotheses to evaluate full-profile versus partial-profile best/worst conjoint analysis. The first has to do with whether the two models have parameter vectors that differ beyond simple rescaling by a multiplicative constant:

$$H_A: \beta_1 = \beta_2$$

The second concerns whether or not the two models contain different amounts of unexplained error variance. Because the "logit scale parameter," μ, is related to the inverse of parameters' unexplained error variance as follows (Ben-Akiva and Lerman 1985),

$$\sigma^2 = 6\pi/\mu^2,$$

H_B is at once a hypothesis about equality of unexplained error variances and its opposite, parameter reliability:

$$H_B: \mu_1 = \mu_2$$

Tests of these hypotheses, and their rationale, are described in Swait and Louviere (1993). Below we employ them for comparing full and partial profile best/worst conjoint analysis model parameters.

EMPIRICAL STUDY

Experimental Design

The best/worst conjoint experiment focused on eleven attributes ($9 \times 7^3 \times 5^2 \times 4 \times 3^2 \times 2^2$) of an office equipment product. Each respondent received both full and partial profile questions, in separate blocks. Blocks were rotated so that about half of the respondents received the full-profile questions first and half received the partial-profile questions first. Six full-profile questions contained levels for all 11 attributes while 11 partial profile questions contained levels for just six attributes each. Thus the two blocks of questions are "fair" in that each contained an equal number, 66, of attribute stimuli. Attribute presence in the partial profile questions varied according to the 12 run 2^{11} design (with a null set) from Hahn and Shapiro (1966). Levels for attributes in both types of questions were assigned using the randomizing capability of the *Ci3* disk-based survey software (Sawtooth Software 1995).

Administration

Respondents, members of the IntelliQuest Technology Panel, were pre-qualified as influencers of the product category under study. A total of 903 respondents completed and returned the disk-by-mail survey. About 52% of the respondents (468) received the partial profile task first and 48% (435) received the full profile task first.

Results

"Don't know" responses were more common for the partial profile questions than for the full-profile questions (8.4% of partial-profile questions versus 6.7% of full-profile questions). Obviously a respondent's range of options is more limited in the partial profile question, so she may more often be confronted with choosing a best or a worst from a list lacking elements from either extreme. Interestingly, respondents used the "don't know" response almost twice as often for choice of the least preferred level as for choice of the most preferred level, both for full- and partial-profile questions.

Analysis *via* multinomial logit yields the sets of coefficients for the full and partial profile models shown in Table 1.

Table 1. Coefficients for Raw B/W Models

	Full Profile		Partial Profile	
Parameter	**β**	**se**	**β**	**se**
Brand	-.06	.05	.04	.03
Processor	**2.57**	**.03**	**1.88**	**.02**
Feature 1	**.55**	**.03**	**.50**	**.02**
Feature 2	**-.61**	**.03**	**-.56**	**.02**
Feature 3	**-1.25**	**.05**	**-1.25**	**.03**
Feature 4	**.89**	**.03**	**.74**	**.02**
Feature 5	**-.22**	**.03**	**-.16**	**.02**
Feature 6	-.05	.04	**-.14**	**.02**
Feature 7	**-1.04**	**.05**	**-.52**	**.04**
Feature 8	**.44**	**.05**	**.51**	**.03**
Brand Level 1	**-.26**	**.08**	-.10	.06
Brand Level 2	**.42**	**.09**	**.51**	**.06**
Brand Level 3	**.26**	**.09**	.09	.06
Processor Level 1	**-.24**	**.04**	**-.08**	**.04**
Processor Level 2	-.03	.04	.02	.04
Processor Level 3	.03	.04	-.05	.04
Processor Level 4	.01	.04	-.05	.04
Feature 1 Level 1	-.11	.09	-.10	.06
Feature 1 Level 2	**.34**	**.09**	**.33**	**.06**
Feature 1 Level 3	**.30**	**.09**	**.35**	**.06**
Feature 1 Level 4	.05	.09	.06	.06
Feature 1 Level 5	**.27**	**.09**	**.41**	**.06**
Feature 1 Level 6	**.53**	**.08**	**.32**	**.06**
Feature 1 Level 7	**-.64**	**.09**	**-.66**	**.06**
Feature 1 Level 8	**-.51**	**.10**	**-.43**	**.06**
Feature 2 Level 1	**.39**	**.09**	**.22**	**.06**
Feature 2 Level 2	**.60**	**.09**	**.53**	**.06**
Feature 2 Level 3	**.17**	**.08**	.03	.06
Feature 2 Level 4	**-.24**	**.08**	**-.15**	**.06**
Feature 2 Level 5	-.14	.08	.07	.06
Feature 2 Level 6	**-.38**	**.07**	**-.41**	**.05**
Feature 3 Level 1	**.69**	**.05**	**.61**	**.04**
Feature 4 Level 1	**.55**	**.07**	**.54**	**.05**
Feature 4 Level 2	**1.04**	**.06**	**.89**	**.04**
Feature 4 Level 3	**.40**	**.07**	**.39**	**.05**
Feature 4 Level 4	**.98**	**.06**	**.66**	**.04**
Feature 5 Level 1	.11	.09	**.14**	**.06**
Feature 5 Level 2	**.73**	**.09**	**.74**	**.06**
Feature 5 Level 3	.05	.09	-.08	.06
Feature 5 Level 4	**.48**	**.09**	**.48**	**.06**
Feature 5 Level 5	**-.18**	**.09**	**-.38**	**.06**
Feature 5 Level 6	.09	.09	**.15**	**.06**
Feature 6 Level 1	**.26**	**.09**	**.16**	**.06**
Feature 6 Level 2	**.19**	**.09**	**.13**	**.06**
Feature 6 Level 3	**.21**	**.09**	**.13**	**.06**
Feature 6 Level 4	.10	.09	**.15**	**.06**
Feature 6 Level 5	**.20**	**.09**	**.29**	**.06**
Feature 6 Level 6	.15	.09	**.12**	**.06**
Feature 7 Level 1	**.23**	**.05**	**.40**	**.04**
Feature 8 Level 1	**-1.14**	**.07**	**-1.24**	**.05**
Feature 8 Level 2	**.48**	**.07**	**.63**	**.05**
Feature 9 Level 1	**.78**	**.07**	**.83**	**.05**
Feature 9 Level 2	**.26**	**.06**	**.09**	**.04**

Bold: coefficient significantly different from 0 at $p < .05$.

Though not uniformly so, most coefficients are larger in the full profile model than in the partial profile model. This suggests that the partial profile model contains more unexplained error variance than does the full profile model (Ben Akiva and Lerman 1985). In other words, the full profile model's parameters appear to be more reliable than those of the partial profile model.

Statistical tests for the equality of model parameters and for the equality of unexplained error variance (Swait and Louviere 1993) utilize log likelihood statistics for

1. the full profile models (LL_1);
2. the partial profile model (LL_2);
3. the "pooled" model that results from concatenating the two design matrices (LL_p),
4. the "scale parameter adjusted" pooled model that equalizes unexplained error in the concatenated design matrix and maximizes the log likelihood of the pooled model (LL_μ).

If the test of H_A identifies a difference in model parameters, it means that the full and partial profile models measure different cognitive processes. A non-significant result for the test of H_A, makes possible the test of H_B. If the test of H_B identifies a difference in unexplained error variance, it means that respondents answered one of the sets of best/worst questions more reliably than they answered the other set.

Resulting log likelihood statistics are

$LL_1 = -34{,}270.09$

$LL_2 = -64{,}905.68$

$LL_p = -99{,}439.26$

$LL_\mu = -99{,}290.19$

The test statistic for H_A is λ_A $= -2[LL_\mu - (LL_1 + LL_2)]$

$= -2[-99{,}439.26 - (-99{,}175.77)]$

$= 228.84$

The critical value for this χ^2 statistic, at $\alpha = .05$ and 54 degrees of freedom, is 72.15, so we reject H_A and conclude that the full and partial profile models measure different cognitive processes.

Table 2–Coefficients for Scale-Adjusted B/W Models

	Full Profile		μ-Adjusted Partial Profile	
Parameter	**β**	**se**	**β**	**se**
Brand	-.06	.05	.05	.04
Processor	**2.57**	**.03**	**2.36**	**.03**
Feature 1	.55	.03	.62	.03
Feature 2	**-.61**	**.03**	**-.71**	**.03**
Feature 3	**-1.25**	**.05**	**-1.57**	**.04**
Feature 4	.89	.03	-.20	.03
Feature 6	**-.05**	**.04**	**-.17**	**.03**
Feature 7	**-1.04**	**.05**	**-.66**	**.05**
Feature 8	**.44**	**.05**	**.64**	**.04**
Brand Level 1	-.26	.08	-.12	.07
Brand Level 2	**.42**	**.09**	**.65**	**.07**
Brand Level 3	**.26**	**.09**	**.12**	**.07**
Processor Level 1	-.24	.04	-.10	.05
Processor Level 2	-.03	.04	.03	.04
Processor Level 3*	.03	.04	-.06	.05
Processor Level 4	.01	.04	-.06	.05
Feature 1 Level 1*	-.11	.09	-.12	.08
Feature 1 Level 2	.34	.09	.41	.07
Feature 1 Level 3*	.30	.09	.44	.07
Feature 1 Level 4	.05	.09	.08	.08
Feature 1 Level 5	**.27**	**.09**	**.51**	**.07**
Feature 1 Level 6*	.53	.08	.40	.07
Feature 1 Level 7	-.64	.09	-.83	.08
Feature 1 Level 8*	-.51	.10	-.54	.08
Feature 2 Level 1	.39	.09	.28	.08
Feature 2 Level 2	.60	.09	.67	.08
Feature 2 Level 3	.17	.08	.04	.07
Feature 2 Level 4	-.24	.08	-.19	.07
Feature 2 Level 5	-.14	.08	.08	.07
Feature 2 Level 6	-.38	.07	-.51	.06
Feature 3 Level 1	.69	.05	.77	.04
Feature 4 Level 1	.55	.07	.68	.06
Feature 4 Level 2*	1.04	.06	1.11	.05
Feature 4 Level 3*	.40	.07	.48	.06
Feature 4 Level 4	.98	.06	.83	.06
Feature 5 Level 1	.11	.09	.17	.08
Feature 5 Level 2	.73	.09	.93	.08
Feature 5 Level 3	.05	.09	-.10	.08
Feature 5 Level 4	.48	.09	.60	.08
Feature 5 Level 5	**-.18**	**.09**	**-.47**	**.07**
Feature 5 Level 6	.09	.09	.18	.08
Feature 6 Level 1	.26	.09	.21	.08
Feature 6 Level 2	.19	.09	.16	.08
Feature 6 Level 3*	.21	.09	.17	.08
Feature 6 Level 4	.10	.09	.18	.08
Feature 6 Level 5	.20	.09	.36	.08
Feature 6 Level 6	.15	.09	.15	.08
Feature 7 Level 1	**.23**	**.05**	**.50**	**.05**
Feature 8 Level 1	**-1.14**	**.07**	**-1.56**	**.06**
Feature 8 Level 2	**.48**	**.07**	**.79**	**.06**
Feature 9 Level 1*	**.78**	**.07**	**1.04**	**.06**
Feature 9 Level 2	.26	.06	.11	.06

Bold: coefficients differ significantly in full and partial profile models
*: coefficient affected by IIA violations

Table 2 contains these significantly differing parameter vectors, after adjusting as much as possible for any difference in unexplained variance. Bold font identifies coefficients differing significantly between the two models. The differences evince no obviously interpretable pattern. We know that the models differ, but any rationale underlying the pattern of differences escapes us.

The logit scale parameter that best equalizes the two models was .797 for the partial profile model (relative to 1.0 for the full profile model). This means our best guess is that the partial profile model contains 57% more unexplained error variance than the full profile model ($1.57 = 1/.797^2$). Recall that the rejection of H_A, however, prevents rigorous statistical testing of H_B.

Chrzan and Skrapits (1997) analyzed this data set for violations of the assumptions of the multinomial logit model. They determined that the partial profile model contains "IIA violations." IIA is an assumption underlying the use of multinomial logit such that its violation invalidates multinomial logit estimation. In the present case, the IIA violations affect a different subset of the attribute levels than differ between the full and partial profile models (indicated with asterisks in Table 2). Thus IIA violations do not seem to explain the two models' differing parameter vectors.

Discussion

The evidence of this study contradicts the assertion of Swait, Louviere and Anderson (1995) that partial profile designs may be used with best/worst conjoint analysis: full and partial profile best/worst models are not equivalent measures of preference structure in this study. As noted above, however, partial profile versions of choice-based conjoint models have been shown in four separate commercial studies to measure the same preference structure as full profile choice-based conjoint models. Apparently there is something specific to best/worst measurement that makes it not work with partial profiles, so we recommend only using full profile stimuli with best/worst measurement.

REFERENCES

Ben-Akiva, Moshe. and Steven. R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge: MIT Press.

Chrzan, Keith, David S. Bunch and Daniel C. Lockhart (1996) "Testing a Multinomial Extension to Partial Profile Choice Experiments: Empirical Comparisons to Full Profile Experiments," paper presented at the INFORMS Marketing Science Conference, Gainesville, FL.

Chrzan, Keith and Terry Elrod (1995) Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes," paper presented at the American Marketing Association's Advanced Research Techniques Forum, Monterey, CA.

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

Chrzan, Keith and Mike Skrapits (1997) "Testing for IIA Violations in Partial Profile Conjoint Models," paper presented at the 1997 INFORMS Marketing Science Conference, Berkeley.

Hahn, G. J. and S. S. Shapiro (1966) "A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial Experiments," Technical Report 66-C 165. Schenectady: General Electric Research and Development Center.

Johnson, Richard M., Allen D. Shocker and Dick R. Wittink (1991) "The Effect of Design and Estimation Program on Conjoint Utility Limits: A Comment," *Marketing Research*, 3, 45-49.

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.

Sawtooth Software (1995) *Ci3 System*, Evanston: Sawtooth Software.

Sawtooth Software (1993) *ACA System*, Evanston: Sawtooth Software.

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

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

COMMENT ON CHRZAN AND FELLERMAN

Marco Hoogerbrugge
SKIM Analytical

The conclusion of this paper is clear: there is a difference between full profile and partial profile Best/Worst analysis. However, it remains a little mysterious **why** this is the case. It seems quite important to me to discover the reason, if it were only to get better acquainted with the mechanisms of this new technique. That is why I have two suggestions:

1. Because it is to a certain extent possible to calculate an individual's utilities, it might be worthwhile to check if the difference between full- and partial-profile is caused by a particular group of respondents.
2. When you check the table with scale-adjusted utility levels carefully, it strikes that most differences occur in attributes with few levels. Now my hypothesis is that in full-profile the results of few-level-attributes might get biased because a respondent evaluates each level much more often than levels of many-level-attributes. In partial-profile this problem is not so big because in every task many attributes are not shown. If this hypothesis would be correct, the conclusion of the paper should be the other way around: use partial-profiles best/worst analysis and don't use full-profile!

EFFICIENT EXPERIMENTAL DESIGNS USING COMPUTERIZED SEARCHES

Warren F. Kuhfeld
SAS Institute, Inc.

ABSTRACT

In the past few years, marketing researchers have been increasingly using sophisticated computerized search algorithms to find experimental designs. This paper reviews some fundamentals of experimental design, orthogonality, and balance, and introduces the idea of design efficiency. It then compares some widely available design software including Sawtooth Software's CVA and SAS Institute's OPTEX programs.

INTRODUCTION

Conjoint analysis is used to study product purchase decisions when the products have several attributes or factors. Consumers "consider jointly" all of the attributes of a set of products, make trade offs, and then report their preferences for the products. The design of experiments is a fundamental part of conjoint analysis. Experimental designs are used to construct the hypothetical products.

For much of the history of experimental design and statistics, researchers used orthogonal designs that they looked up in tables. When an ANOVA model is fit with an orthogonal design, the parameter estimates are uncorrelated, which means each estimate is independent of the other terms in the model. More importantly, orthogonality *usually* implies that the coefficients will have minimum variance and hence maximum precision. For these reasons, orthogonal designs are usually quite good. ANOVA was widely used before the wide-spread availability of modern computers. With orthogonal designs, relatively simple formulas were available for hand or calculator ANOVA computations. Even as late as the 1970's, this was an important reason to use orthogonal designs. However,
in the last ten to twenty years, general linear models software that does not require orthogonality has become widely available, so orthogonality is not as important as it used to be.

Like ANOVA, the early history of conjoint analysis is based on orthogonal designs. However, for many practical problems, particularly in marketing research, orthogonal designs are simply not available. Examples:

- when there are many attributes
- when the number of attribute levels is different for most of the factors
- a nonstandard number of cards is desired
- when some combinations are unrealistic, such as of the best product features at the lowest price.

In these and other situations, *nonorthogonal* designs must be used. During the past several years, marketing researchers are increasingly using efficient nonorthogonal designs (Kuhfeld, Tobias, and Garratt, 1994). These designs are efficient in the sense that the precision of the parameter estimates is maximized. Efficient designs can be found with the aid of a computer for nonstandard situations in which there are no orthogonal designs. A computerized search, with software such as the Sawtooth Software CVA (Conjoint Value Analysis) designer or the SAS Institute (1995) OPTEX procedure can be used to find good, efficient, and realistic conjoint designs.

Before exploring experimental design in detail, it is instructive to compare forms of conjoint analysis. CVA can be used to perform standard full-profile conjoint analysis where subjects rank or rate one product at a time. CVA can also be used for pair-wise presentation of products where subjects are asked to compare two products. CVA is typically used for paper and pencil administered studies. ACA (Adaptive Conjoint Analysis) is another widely used method for conjoint analysis. ACA interactively administers a conjoint study, adapting its questions to the individual respondent. ACA was designed for problems that generally could not be handled by full-profile methods, such as larger problems. CBC (Choice Based Conjoint) is used for fitting a multinomial logit model to discrete choice data. CBC, like ACA, collects data interactively, directly administering the study on the computer. However, CBC also has a paper-and-pencil module. ACA adapts its questions to the respondent; CBC and CVA do not.

CVA creates an efficient conjoint experiment using a computerized search. ACA does not attempt to create an optimal design. Instead, it is guided by another criterion, asking maximally informative questions. For choice models, it is impossible to create an efficient design without first knowing the “true” parameters. Hence, the construction of choice designs must be guided by other principles. CBC strives to make sure that for each pair of attributes, each level is paired with each other level (at least nearly) equally often.

ORTHOGONAL EXPERIMENTAL DESIGNS

An experimental design is a plan for running an experiment. The *factors* of an experimental design are variables that have two or more fixed values, or *levels*. Experiments are performed to study the effects of the factor levels on the dependent variable. In a conjoint study, the factors are the attributes of the hypothetical products or services, and the response is preference or choice. For example, Price could be a factor with levels $1.49, $1.99, and $2.49. A design is *orthogonal* if all effects can be estimated independently of all of the other effects (excluding the intercept). A design is *balanced* when each level occurs equally often within each factor, which means the intercept is orthogonal to each effect. Imbalance is a generalized form of nonorthogonality, which increases the variances of the parameter estimates.

A *full-factorial design* consists of all possible combinations of the levels of the factors. For example, with five factors, two at two levels and three at three levels (denoted $2^2\ 3^3$)*, there are 108 possible combinations. In a full-factorial design, all main effects, all two-way interactions, and all higher-order interactions are estimable and uncorrelated. A full-factorial design is balanced and orthogonal. The problem with a full-factorial design is that, for most practical situations, it is too cost-prohibitive and tedious to have subjects rate all possible combinations. For this reason, researchers often use *fractional-factorial designs*, which have fewer cards than full-factorial designs. The price of having fewer cards is that some effects become confounded. Two effects are *confounded* or *aliased* when they are not distinguishable from each other.

A special type of fractional-factorial design is the *orthogonal array*, in which all estimable effects are uncorrelated. Orthogonal arrays are categorized by their resolution. The resolution identifies which effects, possibly including interactions, are estimable. For resolution III designs, all main effects are estimable free of each other, but some of them are confounded with two-factor interactions. For resolution V designs, all main effects and two-factor interactions are estimable free of each other. Higher resolutions require larger designs. Orthogonal arrays come in specific numbers of cards (such as 16, 18, 20, 24, 27, 28, ...) for specific numbers of factors with specific numbers of levels. Resolution III orthogonal arrays are frequently used in marketing research. The term "orthogonal array," as it is used in practice, is imprecise. It refers to designs that are both orthogonal and balanced, and hence optimal. It also refers to designs that are orthogonal but not balanced, and hence potentially nonoptimal.

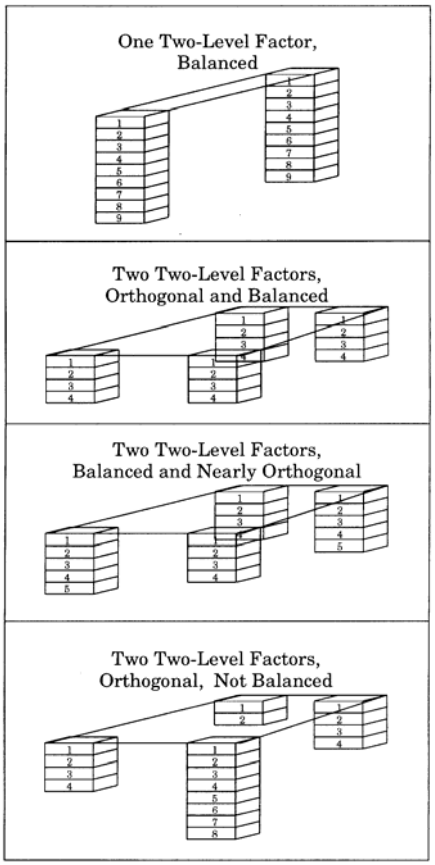
* $2^2 3^3$ means: 2-level factors (there are 2 of them) 3-level factors (there are 3 of them)

ORTHOGONALITY AND BALANCE

A good metaphor for discussing experimental designs is a raft. A raft is a flat boat that you hope will support your weight and keep you from getting wet. An experimental design forms the basis of a conjoint study, and you hope it will provide you with good information to support your marketing decisions. If your raft is not properly constructed, you will fall in the water and get eaten by alligators. If your experimental design is nonoptimal, you will have less information to use to make important decisions, and if your decisions are wrong, you will be eaten alive by your competitors.

A design with a single two-level factor is like a board, a one dimensional raft, supported by Styrofoam floats. For maximum stability with N_D floats (cards), put $N_D/2$ floats under each end of the board. A design constructed according to this principle is balanced. If you put floats in the middle or more floats on one end, the board will be less stable. See Figure 1.

Figure 1. Balance and Orthogonality, Illustrated with Rafts



Two two-level factors are like an ordinary square raft, supported by Styrofoam floats. For maximum stability, with N_D floats (cards), put $N_D / 4$ floats under each corner of the raft. A design constructed according to this principle is orthogonal and balanced. If you put floats in the middle or more floats on some corners, the raft will be less stable. Sometimes it is not possible to equally support all corners. Consider $N_D = 18$ with two two-level factors. Then the best you can do is four cards with the *(a, a)* combination, four cards with the *(b, b)* combination, five cards with the *(a, b)* combination, and five cards with the *(b, a)* combination. A design constructed according to this principle is balanced and nearly orthogonal. Orthogonal designs can be very unbalanced. This leads to much less information being collected about some combinations than others. See Figure 1.

With three-level factors or more than two factors, the raft analogy is harder to imagine. However, the principles are the same. In orthogonal and balanced designs, the corners of the design space are well supported and equally supported. Nearly orthogonal and balanced designs where the corners are nearly equally supported are often the best that you can do in practice.

CODING

Before a design is used, it must be coded. One standard coding is the *binary* or *dummy variable* or (1, 0) coding. Another standard coding is *effects* or *deviations from means* or (1, 0, –1) coding. For evaluating design efficiency, we prefer an *orthogonal coding*. Standard nonorthogonal codings such as effects or binary coding are generally correlated, even for orthogonal designs. We use orthogonal codings so that we can get efficiency statistics scaled to range from 0 to 100. Efficiencies computed using nonorthogonal codings will have a smaller range (except for the special case of two-level factors).

- First a column of ones is coded for the intercept.
- A two-level factor *(a, b)* is replaced by one column.

Binary coding:	*a*	is replaced with	1
	b		0
Effects coding:	*a*	is replaced with	1
	b		-1
Orthogonal coding:	*a*	is replaced with	1
	b		-1

- A three-level factor *(a, b, c)* is replaced by two columns.

Binary coding:	*a*	is replaced with	1	0
	b		0	1
	c		0	0
Effects coding:	*a*	is replaced with	1	0
	b		0	1
	c		-1	-1
Orthogonal coding:	*a*	is replaced with	1.224745	-0.707107
	b		0	1.414214
	c		-1.224745	-0.707107

- A four-level *(a, b, c, d)* factor is replaced by three columns.

Binary coding:	*a*	is replaced with	1	0	0
	b		0	1	0
	c		0	0	1
	d		0	0	0
Effects coding:	*a*	is replaced with	1	0	0
	b		0	1	0
	c		0	0	1
	d		-1	-1	-1
Orthogonal coding:	*a*	is replaced with	1.414215	-0.816497	-0.57735
	b		0	1.632993	-0.57735
	c		0	0	1.73205
	d		-1.414214	-0.816497	-0.57735

- The orthogonal coding for an *n*-level factor is found by creating an $n \times n$ matrix C, with an intercept column and $n - 1$ columns containing the effects coding, then creating $\sqrt{n}\ C(C'C)^{-1/2}$ and discarding the first column.

DESIGN EFFICIENCY

Efficiencies are measures of design goodness. Common measures of the efficiency of an ($N_D \times p$) orthogonally coded design matrix **X** are based on the information matrix **X´X**. The variance-covariance matrix of the vector of parameter estimates β in a least-squares analysis is proportional to $\mathbf{(X'X)}^{-1}$. The variance of $\hat{\beta}_i$ is proportional to the x_{ii} element of $(\mathbf{X'X})^{-1}$. An efficient design will have a "small" variance matrix, and the eigenvalues* of $(\mathbf{X'X})^{-1}$ provide measures of its "size." Two common efficiency measures are based on the idea of "average variance" or "average eigenvalue". *A-efficiency* is a function of the arithmetic mean of the variances, which is given by trace $((\mathbf{X'X})^{-1})/p$. (The trace is the sum of the diagonal elements of $(\mathbf{X'X})^{-1}$, which is the sum of the variances and is also the sum of the eigenvalues of $(\mathbf{X'X})^{-1}$.) *D-efficiency* is a function of the geometric mean of the eigenvalues, which is given by $|(\mathbf{X'X})^{-1}|^{1/p}$. (The determinant, $|(\mathbf{X'X})^{-1}|$, is the product of the eigenvalues of $(\mathbf{X'X})^{-1}$, and the *pth* root of the determinant is the geometric mean.) A third common efficiency measure, *G-efficiency,* is based on σ_M, the maximum standard error for prediction over the candidate set. All three of these criteria are convex functions of the eigenvalues of $(\mathbf{X'X})^{-1}$ and hence are usually highly correlated.

A-efficiency is based on the average of the variances of the parameter estimates. A-efficiency is perhaps the most natural criterion to use in evaluating design goodness. As orthogonality decreases, both the off-diagonal and diagonal elements of $(\mathbf{X'X})^{-1}$ increase. Looking at the average variance while ignoring the off-diagonal covariances, is reasonable because the variances increase as the covariances increase. D-efficiency is perhaps less intuitive than A-efficiency, but both provide a measure of the average size of the variance matrix. D-efficiency is used more often in practice for two reasons. Relative D-efficiency† is invariant under different codings; relative A-efficiency is not. Also D-efficiency is easier to update, so programs based on D-efficiency run faster.

---

* Eigenvalues are proportional to squared lengths. To understand eigenvalues, visualize a slightly deflated American football. Imagine holding it so the longest dimension is horizontal. Since it is partly deflated, imagine it positioned so the next longest dimension is vertical, and the shortest dimension corresponds to depth, the direction perpendicular to horizontal and vertical. The squared horizontal length is the first eigenvalue, the squared vertical length is the second eigenvalue, and the squared depth length is the third eigenvalue. These three numbers provide information about the size of the space occupied by the football. The eigenvalues of a variance matrix give information about the sizes of the variances.

† Relative efficiency is the ratio of two efficiency statistics.

For all three criteria, if a balanced and orthogonal design exists, then it has optimum efficiency; conversely, the more efficient a design is, the more it tends toward balance and orthogonality. Assuming an orthogonally coded **X:**

- A design is balanced and orthogonal when $(\mathbf{X}'\mathbf{X})^{-1}$ is diagonal.
- A design is orthogonal when the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$, excluding the row and column for the intercept, is diagonal; there may be off-diagonal nonzeros for the intercept.
- A design is balanced when all off-diagonal elements in the intercept row and column are zero.
- As efficiency increases, the absolute values of the diagonal elements get smaller and the diagonals approach $1/N_D$.

These measures of efficiency are scaled to range from 0 to 100:

$$\text{A-efficiency} = 100 \times \frac{1}{N_D \text{ trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right)/p}$$

$$\text{D-efficiency} = 100 \times \frac{1}{N_D \left|(\mathbf{X}'\mathbf{X})^{-1}\right|^{1/p}}$$

$$\text{G-efficiency} = 100 \times \frac{\sqrt{p/N_D}}{\sigma_M}$$

These efficiencies measure the goodness of the design relative to hypothetical orthogonal designs that may be far from possible, so they are not useful as absolute measures of design efficiency. Instead, they should be used relatively, to compare one design to another for the same situation. Efficiencies that are not near 100 may be perfectly satisfactory.

Figure 2. Candidate Set and Optimal Design

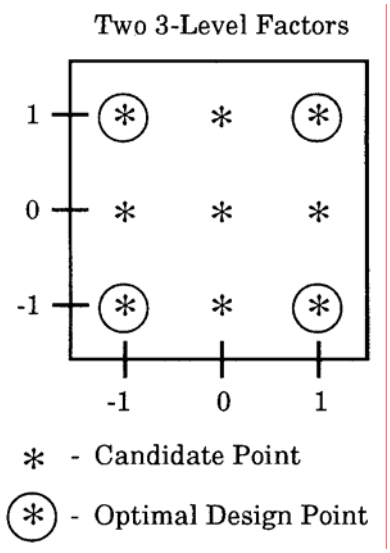


Figure 2 shows an optimal design in four cards for a simple example with two factors, using interval measure scales for both. There are three candidate levels for each factor. The full-factorial design is shown by the nine asterisks, with circles around the optimal four design points. As this example shows, efficiency tends to emphasize the corners of the design space. Interestingly, nine different sets of four points form orthogonal designs—every set of four that forms a rectangle or square. Only one of these orthogonal designs is optimal, the one in which the points are spread out as far as possible.

COMPUTER-GENERATED DESIGN ALGORITHMS

When a suitable orthogonal design does not exist, computer-generated nonorthogonal designs can be used instead. Various algorithms exist for selecting a good set of *design points* from a set of *candidate points*. The candidate points consist of all of the factor-level combinations that may potentially be included in the design, for example the nine points in Figure 2. For small problems, such as $2^2 3^3$, a good candidate set is the full-factorial design, since it contains only 108 cards. For larger problems, fractional-factorial designs make good candidate sets. When the full-factorial is more than say 1024 cards, it is always a good idea to try a fractional-factorial candidate set. Even with software that can handle several thousand candidates, it is good to also try small good candidate sets, because it is easier for the computer to find good designs when the search is limited to a small region.

N_D, the number of cards, is chosen by the researcher. Unlike orthogonal arrays, N_D can be any number as long as $N_D \geq p$, where p is the number of parameters.* The algorithm searches the candidate points for a set of N_D design points that is optimal in terms of a given efficiency criterion. There usually is not enough time to list all N_D-run designs and choose *the* most efficient or optimal design. For example, with $2^2\ 3^3$ in 18 cards, there are $108! / (18!(108 - 18)!) = 1.39 \times 10^{20}$ possible designs. Instead, nonexhaustive search algorithms are used to generate a small number of designs, and the most efficient one is chosen. Usually, an initial design is randomly selected from the candidates, then it is iteratively refined. The algorithms select points from the candidate set for possible inclusion or deletion, then update the efficiency criterion. The points that most increase efficiency are added to the design. These algorithms invariably find efficient designs, but they may fail to find *the* optimal design, even for the given criterion. For this reason, we prefer to use terms like *information-efficient* and *D-efficiency* over the more common *optimal* and *D-optimal*.

There are many algorithms for generating information-efficient designs. We will begin by describing some of the simpler approaches and then proceed to the more complicated (and more reliable) algorithms. Dykstra's (1971) sequential search method starts with an empty design and adds candidate points so that the chosen efficiency criterion is maximized at each step. This algorithm is fast, but it is not very reliable in finding a globally optimal design. Also, it always finds the same design (due to a lack of randomness). These next algorithms are typically run repeatedly for a given candidate set and different random initial designs, then the most efficient design is chosen. The Mitchell and Miller (1970) simple exchange algorithm is a slower but more reliable method. It improves an initial design by adding a candidate point and then deleting one of the design points, stopping when the chosen criterion ceases to improve. The DETMAX algorithm of Mitchell (1974) generalizes the simple exchange method. Instead of following each addition of a point by a deletion, the algorithm makes excursions in which the size of the design may vary. These three algorithms add and delete points one at a time.

The next two algorithms add and delete points simultaneously, and for this reason, are usually more reliable for finding the truly optimal design; but because each step involves a search over all possible pairs of candidate and design points, they generally run much more slowly (by an order of magnitude). The Federov (1972) algorithm simultaneously checks each candidate point and design point pair, then makes the swap that most increases efficiency. Cook and Nachtsheim (1980) define a modified Federov algorithm that checks each candidate point and design point pair and makes every swap that increases efficiency. The resulting procedure is generally as efficient as the simple Federov algorithm in finding the optimal design, but it is up to twice as fast.

---

* The number of parameters is the sum across all attritubes of the number of levels of each attribute, minus the number of attributes, plus one for the intercept.

CVA DESIGNER

The CVA design software automatically: creates a candidate set, excludes prohibited pairs, creates an initial design, uses the modified Federov algorithm to improve the efficiency, then it discards the candidate set and performs additional iterations to improve balance and overall efficiency. It repeats this process a user controlled number of times (five by default) then outputs the best design. Here is more detail on the algorithm:

- CVA generates the candidate set with a guided randomization process. For each attribute, CVA randomly picks a pair of levels from all permitted pairs, that have been presented least often. Pairs of levels are not repeated until all other permitted pairs have been shown. This creates a candidate set with good balance. For example, when a 20-profile design is requested, CVA by default creates a candidate set with 120 profiles.
- Next, CVA creates an initial design. It starts with the full candidate set and excludes one card at a time, the card that contributes the least to the design. CVA considers excluding each point and checks its effect on efficiency. It performs the exclusion that leads to the maximum efficiency. At first, efficiency may actually increase as the points that provide the least information are removed. Then, typically, efficiency will start to decrease. The initial design has the same number of profiles as the final design—for example 20.
- Next, CVA uses the modified Federov algorithm, swapping (previously excluded) candidates back into the design until efficiency stops improving.
- For the final step, the candidate set is discarded. CVA looks for imbalance and identifies the levels that appear most often. CVA considers changing those levels, making sure that prohibited pairs are not introduced. If changing a level to improve balance also increases efficiency, it is done. In effect, CVA is using a virtual candidate set in this step. Possible candidates include every card in the full factorial (minus prohibited pairs), but only a relatively few candidates are considered, those that improve balance.
- The entire process is repeated and the best design is chosen.

We will investigate the CVA designer for use in full-profile conjoint experiments. Other capabilities of CVA such as its ability to generate designs for pair-wise presentation are beyond the scope of this paper.

THE OPTEX PROCEDURE

The OPTEX procedure requires the user to create a candidate set. A good candidate set for a small problem is a full-factorial design. Resolution III, IV, V, and perhaps larger designs are good candidate sets for larger problems (Kuhfeld, 1996). Unrealistic or undesirable combinations can then be excluded from the candidates. PROC OPTEX starts with a random initial design and then iteratively improves it. PROC OPTEX has sequential, exchange, DETMAX, and Federov algorithms, but I usually use Modified Federov. The entire process is repeated (10 times by default), and the best design is chosen.

AN EMPIRICAL EVALUATION OF CVA AND PROC OPTEX

This section compares CVA and PROC OPTEX with problems. The first three examples are plausible conjoint studies. The next two examples are harder problems than you are likely to find in a reasonable conjoint study.

The first test was a relatively simple problem, $2^2\ 3^3$ in 18 cards. The optimal design, described by Kuhfeld, Tobias, and Garratt (1994), is nonorthogonal. CVA requires the user to enter the names of the factors, the levels, and the number of cards. CVA generates the candidate set and performs the searches. I created a full-factorial design for the PROC OPTEX candidate set. Both CVA and PROC OPTEX found the optimal design, with D-efficiency = 99.861 and perfect balance, in a matter of seconds.

The next test was harder, $2^2\ 3^3\ 5^2$ in 30 cards, but still small enough to be realistic for a conjoint experiment. CVA found a good design with D-efficiency = 96.2149 in about three minutes. The balance was perfect. All of my attempts with CVA to find a better design, by both generating more designs and changing the size of the candidate set failed. Using PROC OPTEX, I was able to find an unbalanced design with D-efficiency = 97.6690. With subsequent tries, I found a perfectly balanced design with D-efficiency = 98.0327. Since the full-factorial design at 2700 cards is large, I started with fractional-factorial candidate sets and worked my way up to the full factorial. The CVA design was slightly (98%) less efficient than the PROC OPTEX design, and both programs found perfectly balanced designs.

The next test was harder still, $2^2\ 3^2\ 5^2\ 7$ again in 30 cards. CVA found a good design with D-efficiency = 88.2956, which I was able to increase to 89.8463 in subsequent tries. The balance was excellent. With PROC OPTEX, I found designs with efficiency ranging up to 92.2954. The CVA design was slightly less efficient than the PROC OPTEX design but much better balanced.

Next, I tried a larger and much more difficult problem, 2 3 4 5 6 7 8 9 10 11, using the CVA recommended 168 cards. This is not a realistic design for a full-profile conjoint analysis (at least without blocking). Still, it seemed reasonable to test CVA's performance with a larger and more difficult problem. CVA found a design with efficiency

94.3472, whereas PROC OPTEX found a design with efficiency 96.3529. Again, the PROC OPTEX design was slightly more efficient, and the balance in the CVA design was
excellent and much better than the PROC OPTEX design.

Last, I tried a large problem, $2\ 3\ 4^2\ 5^2\ 6$ with 24 cards* and prohibited pairs. CVA allows the user to specify pairs of attribute levels that should never be presented together, for example largest size and smallest price. PROC OPTEX allows any combination to be excluded from the candidate set. The following pairs were prohibited: (x1 = 1, x2 = 1), (x2 = 1, x3 = 1), (x3 = 1, x4 = 1), (x4= 1, x5 = 1), (x5 = 1, x6 = 1), and (x6 = 1, x7 = 1). Using a full-factorial candidate set and generating ten designs, PROC OPTEX found a design with D-efficiency = 86.9362 in eight minutes. Generating 100 designs took one hour and resulted in a D-efficiency of 88.4463. Balance was good but not perfect. The first level tended to occur less often, particularly in the two- and three-level factors due to the prohibited pairs. I easily found a CVA design with D-efficiency = 81.9513. Letting CVA run overnight resulted in D-efficiency = 84.2510. The CVA design was better balanced than the OPTEX design.

CVA is easier to use than PROC OPTEX, particularly for the less-experienced user, because CVA automatically creates the candidate set. In contrast, for the more-experienced user, PROC OPTEX is more likely to find a more efficient design because the user can control the candidate set. PROC OPTEX typically runs faster than CVA, but with more user set-up time. For all but very small problems, PROC OPTEX is typically run several times with different candidate sets, then the best design from the best candidate set is chosen. For difficult problems, there are many ways to create reasonable candidate sets, and it is impossible to predict which way will work best. Learning how to create good candidate sets is not easy.

CVA usually finds good designs with excellent balance. For small problems like you would typically encounter in a full-profile conjoint study, CVA seems to do an excellent job. However, for larger and more difficult problems, it often fails to find more efficient designs that can be found with PROC OPTEX. The differences in efficiency between the two programs are small and may in part be offset by CVA's better balance. Balance is *very* important. You do not want any level (particularly for attributes like brand and price) appearing a lot more often than some other level. Some analysts generate many designs, output the most efficient few, and them pick the most balanced design from the most efficient few designs, even if the most balanced design is not the most efficient. In the next section, I will discuss ways in which CVA and PROC OPTEX might be improved.

* This design is almost saturated since there are 23 parameters, so this example is not realistic.

AN EVALUATION OF CVA AND PROC OPTEX ALGORITHMS

CVA starts by creating a guided random candidate set with good balance. It then creates an initial design by excluding cards from the candidate set. The approach typically used with PROC OPTEX is for the user to create a full-factorial design for small problems, and resolution III, IV, V, and perhaps larger candidate sets for larger problems. PROC OPTEX by default uses a random sample of the candidate set as the initial design. The next step for both methods is the modified Federov swaps, which is quite standard and works quite well. CVA has a final step that iterates further, simultaneously increasing efficiency and improving balance. This last CVA step is new, innovative, and I believe a *very* good idea.

The reason that PROC OPTEX can often find a more efficient design than CVA is due to the first two steps. I suggest that Sawtooth Software seriously look at using a full factorial for the candidate set with small problems and fractional factorials for larger problems. Candidate sets with well over one thousand cards should not pose any problems on today's PC, although it is frequently the case that a smaller candidate set is actually better. Perhaps using CVA's guided randomization to augment a core fractional-factorial candidate set would be a good idea. (I have never actually tried this.) I also suggest that Sawtooth Software consider using a random sample from the candidate set as the initial design.

For small problems, the CVA modified Federov swaps are reasonably fast. For larger problems I think they could be made faster. The final efficiency and balance optimization is no doubt the reason why CVA does such a good job of finding (at least nearly) balanced designs. However, it is slow for large problems and could be made faster.

PROC OPTEX would benefit from a graphical user interface, an option for automatic candidate set creation, and an option to optimize balance like CVA does.

CONCLUSIONS

Computer-generated experimental designs can provide both better and more general designs for conjoint studies. Classical designs, obtained from books or computerized tables, are good when they exist, but they are not the only option. When the design is nonstandard and when there are restrictions, a computer can generate a design, and it can be done *quickly*. For most conjoint studies, a good design can be generated in a few minutes. Furthermore, when the circumstances of the project change, a new design can again be quickly generated. The computerized search usually does a good job, it is easy to use, and it can create a design faster than manual methods, especially for the nonexpert. In nonstandard situations, simultaneous balance and orthogonality may be unobtainable. Often, the best that can be hoped for is optimal efficiency. Computerized algorithms help by searching for the most efficient designs from a potentially very large set of possible designs.

I am very pleased that more marketing researchers and more software packages are now using efficiency to guide their design search. PROC OPTEX does an excellent job in finding efficient designs even for very large problems, however less-experienced users may find it hard to use. CVA does an excellent job with small problems (which are the kinds of problems for which it was designed), a good job with larger problems, produces designs with excellent balance, and is particularly well suited for less experienced users. The final stage of the CVA designer algorithm is innovative, and does an excellent job of producing at least a nearly balanced designs.

REFERENCES

Cook, R. Dennis and Christopher J. Nachtsheim (1980), "A Comparison of Algorithms for Constructing Exact D-optimal Designs," *Technometrics*, 22 (August), 315-24.

Dykstra, Otto (1971), "The Augmentation of Experimental Data to Maximize $|(\mathbf{X}'\mathbf{X})^{-1}|$," *Technometrics*, 13 (August), 682-88.

Federov, Valery V. (1972), *Theory of Optimal Experiments*, translated and edited by W.J. Studden and E.M. Klimko, New York: Academic Press.

Kuhfeld, Warren F., Tobias, Randall D., & Garratt, Mark (1994) "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31 (November), 545-557.

Kuhfeld, Warren F. (1996) "Marketing Research Methods in the SAS System," an unpublished collection of papers and handouts.

Mitchell, Toby J. and F.L. Miller, Jr. (1970), "Use of Design Repair to Construct Designs for Special Linear Models," *Math. Div. Ann. Progr. Rept*. (ORNL-4661), 130-31, Oak Ridge, TN: Oak Ridge National Laboratory.

Mitchell, Toby J. (1974), "An Algorithm for the Construction of D-optimal Experimental Designs," *Technometrics*, 16 (May), 203–210.

SAS Institute Inc. (1995), *SAS/QC Software: Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

COMMENT[1] ON KUHFIELD

Joop J. Hox
Department of Education
University of Amsterdam

In his paper, Warren Kuhfeld gives a nice overview of why both balance and orthogonality are important in experimental designs, and provides us with a clear explanation of what programs for conjoint analysis do to achieve these goals, or at least to come close to that achievement. He mentions three reasons why having a balanced and orthogonal design is important: (1) they produce uncorrelated parameter estimates, (2) they are generally efficient, meaning that the estimates have a high precision, and (3) they are computationally simpler than designs that are not orthogonal and balanced. With the advent of powerful software
and cheap computing, the last reason is not very important anymore. In fact, if constructing a nonorthogonal design has its own advantages, such as having many fewer cards to present and consequently less expensive data collection, there is no longer any computational reason not to do so. However, the other two reasons why balanced and orthogonal are attractive are still valid, and it pays to continue to construct designs that are close to that ideal.

The keywords in this paper are *efficiency*, *balance*, and *orthogonality*. The criterion used by programs to construct good nonorthogonal designs is in fact a specific definition of efficiency: the relative precision of the design compared with a balanced and orthogonal design (which may in fact not be attainable given the constraints such as exclusions or number of cards allowed). This is reasonable, because the more efficient designs tend to be more balanced and more orthogonal as well. Having a high precision is by itself not necessarily a very important goal, because a somewhat lower precision can always be countered by taking a somewhat larger sample. If the precision of a design is 90%, one can obtain the precision of a fully efficient design by collecting data from about 10% more respondents. Thus, if data collection costs are low, this is not a problem.

However, taking a larger sample of respondents does not change the degree of imbalance or nonorthogonality. Imbalance means that some (combinations of) levels are presented more often than others. In general, the parameters of such levels will be estimated with a larger precision than the parameters of the other levels. More importantly, if the imbalance is severe, the respondents may notice this, and form their own ideas about

[1] This comment is based on a preliminary version of the paper. After the presentation at the conference, there was a lively discussion, and I have tried to include some of the remarks made there. Thus, not all ideas presented here originate with the author. However, I claim full responsibility for any errors.

what the investigator is really after. Generally, investigators do not want their subjects to form their own ideas about the research. In social psychology this is known as the *demand characteristics* of the experiment, and it may introduce all kinds of bias in the results. Thus, substantial imbalances in the design are undesirable.

Similarly, large nonorthogonalities in the design are also undesirable. If the design is not orthogonal, at least some of the parameter estimates are correlated. If the correlations between the estimates are large, it becomes difficult to interpret them independently from each other. It is difficult to give a rule of thumb when the degree of nonorthogonality becomes a problem, partly because nonorthogonality is not a global attribute of the design. For example, all parameter estimates could be slightly correlated, or most of the parameter estimates could be uncorrelated, while a few correlate rather high. The former would be no problem, and whether the latter is a problem could depend on *which* parameter estimates correlate. In the discussion a value of 0.30 was mentioned as an acceptable upper limit to the correlation between any two parameter estimates. This sound reasonable, since a correlation of 0.30 implies about 10% confounded error (sampling) variance. This would not hinder independent interpretation of the corresponding parameters. It would be useful if the programs that construct experimental designs include these correlations as diagnostics.

In the simulations CVA was either better or equally good as PROC OPTEX in producing balanced designs. Since PROC OPTEX was either better or equally good as CVA in producing efficient designs, it follows that PROC OPTEX's designs are more orthogonal (I thank Warren Kuhfeld for pointing this out to me). The only way to improve both balance and orthogonality is to find a more efficient design, something at which PROC OPTEX seems a little bit better. Thus slight superiority comes at the expense of *much* more time needed to set up the software and to specify a good candidate set. In the discussion after the presentation it was made clear that for *all* solutions found in all simulations both the degree of balance and orthogonality were in fact very good, even with the difficult last simulation problem. Again, since both imbalance and nonorthogonality could be present in only part of the design, it is difficult to give a simple global rule for when a design is acceptable. Given the range of problems in Kuhfeld's paper, I suggest that as a rule of thumb a design with a d-efficiency of at least 0.80 could be considered as *acceptable*, with an efficiency of at least 0.90 as *good*, and with an efficiency of 0.95 as *excellent*. However, even with acceptable designs it would still be worthwhile to investigate the diagnostics for specific parts of the design.

A note about the algorithm. Kuhfeld notes that CVA uses a clever but elaborate procedure to find a good initial design as input for the Federov optimization. Still, PROC OPTEX, which relies much more on simple brute force calculations to construct a initial design, appears to produce slightly more efficient designs. This is not totally surprising, since brute force methods work quite well in many applications. I agree with Kuhfeld's recommendation to use a random sample of a full factorial or fractional design as the initial design. Given current computer capacities, it should be simple to generate about 100 initial designs, and use the best 10 or so as input for the Federov optimization procedure.[2] Programming this is probably straightforward, and I anticipate that it would work very well.

[2] I use the imprecise terms 'about' and 'or so' to indicate that it should be easy to include an option to let users specify their own numbers here.

PRACTICAL WAYS TO MINIMIZE THE IIA-BIAS IN SIMULATION MODELS

Rainer Paffrath[1]
Simon, Kucher & Partners

1. INTRODUCTION

For our purposes, the preference model comprises a conjoint measurement model which yields individual partworth utilities for each attribute level which are relevant in the purchase decision. For the products ("stimuli") in the respective markets under consideration, one could for example use a simple summation to calculate the total utilities.

The second model phase within the brand selection decision lies in the decision model. The total utilities calculated in the first phase are now transformed into shares of preference.

With the decision models, the family of LUCE-Models (cf. Luce 1959, pp. 5 ff.) plays a special role because they have a relatively simple analytical form and are thus widely used in practice. The preference formation is deterministic in these models, but the actual selection process is probabilistic. In contrast to the First Choice Model, one cannot say with 100% certainty which product will be chosen. Even those products with minimal total utility receive a share of preference greater than zero. The models in the LUCE-Family include the BTL-Model and the Logit-Model (Product k out of set K, P indicates the probability, i indicates an individual):

$$P_i(k|K) = P_{ik} = \frac{u_{ik}}{\sum_{k=1}^{K} u_{ik}} \quad \text{(BTL-Model)}$$

$$P_{ik}(\alpha) = \frac{u^{\alpha}{}_{ik}}{\sum_{k=1}^{K} u^{\alpha}{}_{ik}} \quad \text{(generalized BTL-Model)}$$

$$P_{ik}(\alpha) = \frac{e^{u_{ik}\alpha}}{\sum_{k=1}^{K} e^{u_{ik}\alpha}} \quad \text{(Logit-Model)}$$

[1] This essay does not only represent the ideas of the author. Many discussions with Bernhard Böffgren, Dr. Jan Engelke, Claus Kolvenbach, Dr. Meinhard Kneller, and Dieter Lauszus supplied more than fruitful impulses. Without the valuable translation by Frank Luby and Anja Lenninger, this essay would have been held back from the English-speaking world.

The use of the LUCE-Models is connected with a problem: the models have the property known as “Independence from irrelevant alternatives” (IIA). According to this property, the relative selection probabilities are independent from alternatives. Formula:

$$P(k_\alpha|K) = P(K^-|K) * P(k_\alpha|K^-),$$
$$P(k_\beta|K) = P(K^-|K) * P(k_\beta|K^-),$$

which yields:

$$\frac{P(k_\alpha|K)}{P(k_\beta|K)} = \frac{P(k_\alpha|K^-)}{P(k_\beta|K^-)} \text{ für } K^- \subset K; k_\alpha, k_\beta \in K^-$$

Example. The effect is often illustrated with the following example. (cf. Ben-Akiva and Lerman 1994, pp. 51 f.). A commuter has the following choice probabilities for alternative forms of transportation:

$$P(car) = .5$$
$$P(blue\,bus) = .5$$

Now we introduce a bus line with a red bus. Except for the color, the bus service is identical to that of the blue bus. If one uses a model of the LUCE-family, the relationships among the choice probabilities of the previously available alternatives remain constant. Thus, the choice probabilities (red bus line included) would be as follows:

$$P(car) = .33$$
$$P(blue\,bus) = .33$$
$$P(red\,bus) = .33$$

This is unrealistic, because the color of the bus makes no actual difference to the commuter (the color of the bus is totally irrelevant). The selection probabilities should therefore be as follows:

$$P(car) = .5$$
$$P(blue\,bus) = .25$$
$$P(red\,bus) = .25$$

The classic “red bus/blue bus-example” shows that the IIA feature is not desirable in simulation models based on Conjoint Measurement. The selection axiom of the LUCE-Models is only applicable for alternatives which are ‘alike as possible.’

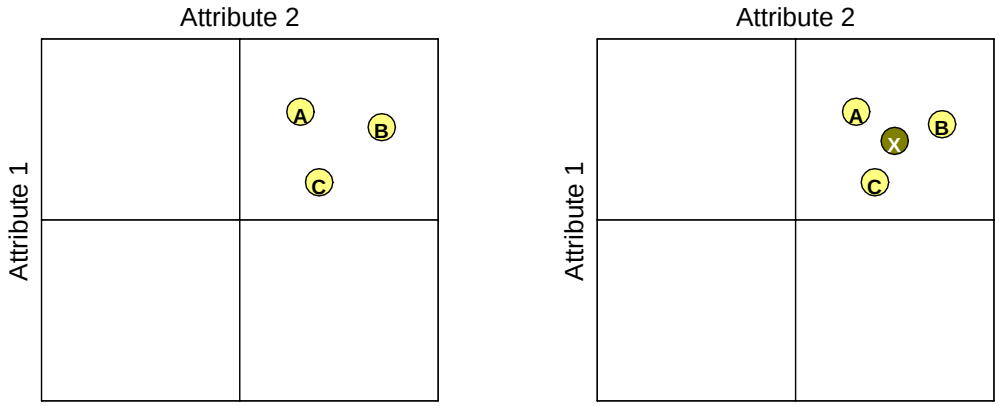
2. When do you need to watch out for IIA-Bias?

Researchers who are aware of the IIA problem would first consider whether the mentioned problems do actually occur in his case. IIA generally first plays a role when the substitution relationships modeled with the LUCE-Models are not sufficient, i.e. when products with varying degrees of dissimilarity are to be considered in the model. A representation in a mapping is helpful in this context for the analysis of similarities.

The multidimensional scaling, which is used to generate such a mapping, presents a n-dimensional space in two dimensions, and with as little information loss as possible.

Let us assume that a mapping contains three products which are all located the same distance from each other. If you introduce a product which is positioned exactly in the middle of these other products, it is rational to assume that the new product takes a correspondingly equal market share, i.e. choice probability away from the other three products which previously shared the market equally. But if in addition to the three equidistant products one introduces a product which is 'very far away' from the others, one assumes that the new product will receive little if any market share from the 'very far away' product, but would receive a relatively large amount of market share from the similar products.

Fig. 1 Mapping Situation 1



If one introduces in Situation 1 a product X to the established products A, B, C, the IIA property does not pose a problem. If A, B, C originally shared the market equally, i.e.

$$P(A) = P(B) = P(C) = .33,$$

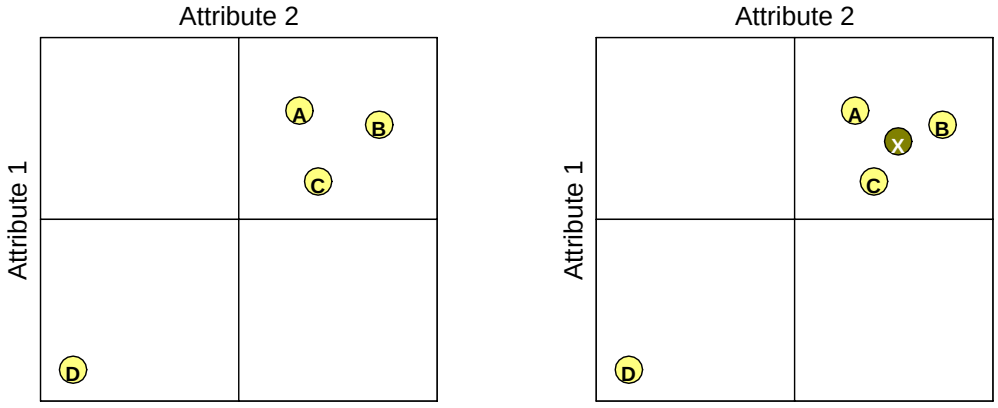
then the assumption of constant relative choice probabilities is realistic, i.e.

$$P(A) = P(B) = P(C) = P(X) = .25$$

Situation 2 is different. In this situation A, B, C, D would share the market as follows:

$$P(A) = P(B) = P(C) = .2$$
$$P(D) = .4$$

Fig. 2 Mapping Situation 2



The relative choice probabilities are therefore

$$\frac{P(A)}{P(D)} = \frac{.2}{.4} = .5 = \frac{P(B)}{P(D)} = \frac{P(C)}{P(D)}.$$

The introduction of product X therefore leads to the following choice probabilities:

$$P(X) = .25$$
$$\Rightarrow P(A) = P(B) = P(C) = .15$$

and $P(D) = .3$. But this is unrealistic, because X should receive proportionally more share from A, B, and C than from X.

To summarize, the IIA problem is particularly acute with heterogenous similarity relationships. This description is first of all inaccurate. One possibility would be to conduct formal tests (cf. Ben-Akiva and Lerman 1995, pp. 183 ff.). An example of an application is the extension of product lines under which slightly modified products which differ

minimally from other products in the product line but relatively significantly from competitive products are introduced into the market.

The paper is not intended to present a fundamentally new method for the reduction or elimination of IIA bias. Instead it will present modifications to available methods and then compare existing and modified methods on the basis of an evaluation scheme. On the one hand, correction methods will be discussed. These methods permit the continued use of the LUCE-decision models and corrects for the IIA bias. The paper will also discuss alternative and simple-to-use decision models, which will bypass the IIA feature or minimize its influence. This should provide researchers with a set of instruments which provides them in the appropriate situations with an optimal method (on the basis of the evaluation scheme) for the calculation of choice probabilities. The means to correct for IIA bias are either inadequate or non-existent in current standard software.

This paper is built along the train of thought described above. First, a catalog of demands/requirements for correction methods and for alternative decision models will be put together. Afterwards, corrective procedures and then alternative decision models will be described and discussed.

3. CATALOG OF DEMANDS/REQUIREMENTS

One basic assumption underlying the corrective procedures and alternative decision models described below is that when two products are relatively similar, one should expect a relatively large substitution relationship between them. On the other hand, when two products are very dissimilar, there will be little or no effect on the selection decision. In other words, **similarities among products are only a means by which one can describe substitution relationships**.

The following criteria should be met by measures designed to minimize IIA bias:

1. The corrective method or alternative decision model should be an **"appropriate" procedure** for taking similarities into account. This means that a new alternative should receive a relatively large amount of market share from relatively similar products and relatively little market share from the relatively dissimilar products. In the extreme case, the introduction of an identical product in the corrected model should lead to a reduction of one half in the share of preference of the identical product.

2. The corrective method should be usable **at the individual level**. The following example helps explain this point: similarity is without doubt a subjective phenomenon. The auto buyer sees for example the station wagons from different manufacturers as very similar, because he wants to buy an automobile large enough to accommodate his family. Another car buyer who instead thinks very economically will find those automobiles in his self-defined price range to be

similar and therefore—in contrast to the station wagon buyer—lumps automobile styles (sedan, hatchback, station wagon) into one basket.

3. Equivalent to this demand is the following: **the individual importance structure should be taken into account**. On the basis of the individual partworth utilities, one can determine the relative importance of the features under consideration. This information is valuable for the determination of similarities. In the first example above, the style of the automobile is the most important feature, because the auto buyer is planning to buy a station wagon and nothing else. The price plays a less important role. The more economical buyer, in contrast, is more price-oriented, i.e. the price is the most important feature for him.

4. It is of general importance that the features used to determine similarities comprise all those **features which are necessary to describe the relevant product**. This is, however, one of the prerequisites for a conjoint measurement study and will therefore not be discussed in detail in this paper. If one forgets to include features which are important for determining similarities, the results obtained can be easily misleading.

5. An additional requirement for an appropriate corrective method or alternative decision model is that **individual evoked sets** are considered. This requirement follows on the assumption that a buyer rarely has more than n products in his or her evoked set (n is dependent of the investigated market, $1 < n < 10$). In the automotive market, for example, even when only one class of cars is considered, e.g. luxury cars, it is not unusual to have 100 or more different elements, especially when one accounts not only for engine size and model but also for the type of car (sedan, station wagon etc.).

6. A rather technical demand on corrective methods: the worsening of a product (which could involve, for example, a price increase) generally has two effects: first, the share of preference declines, because the product is perceived as worse than before. Second, the product will either become more similar or more dissimilar to existing products. In the case that the product becomes more dissimilar and the corrective method is “appropriate” within the context described above, the product will be “improved” by the corrective method. In the extreme case, this “improvement” via the corrective method can offset the worsening of the product. This emergence of this effect should be prevented.

7. This next requirement also has to do with changes in the products. The requirement is that “slight” product changes only lead to “slight” changes in similarity and therefore only to minimal changes in the share of preference. Significant changes in the shares of preference should only be felt after certain threshold levels have been crossed. As will be shown later, this requirement contradicts the procedure used in ACA and CBC.

8. Until now nearly all statements have only been made regarding the direction of the correction. It is difficult to make generalized and universally applicable statements about the extent of the correction. But there are nonetheless two reference points available. First, the result of a first-choice simulation is advisable. The first-choice decision model produces results that are not subject to the IIA-bias. If you use
   the logit-model with a "high" exponent ("high" depends on how the utilities are scaled and standardized) you will produce similar results to those of the first-choice model. In other terms the first-choice model is a special case of the logit-model. Another reference is produced with a simulation considering only one representative of each product category. A prerequisite for this method is the existence of product categories or hierarchies. Refer to section 5.2 (alternative decision models).

9. The idea behind this paper is to allow for the application of the LUCE-model though these models are subject to the IIA-bias. Thus the amount of extra programming for corrections or alternative decision models should be kept at a minimum. Otherwise one could just as good use more "academic" models (i.e. PROBIT, Elimination by aspects).

4. SOME GENERAL REMARKS ON THE USE OF DIFFERENT DATA INPUT FOR SIMILARITY OR DISSIMILARITY MEASUREMENT

Generally the following data are suitable inputs for measuring the similarity between products in corrective methods:

the "product database",

the "utilities database" or

distances or similarities measured with Non-Conjoint-Measurement data (i.e. distances calculated with multidimensional scaling)

The first and probably simplest method to measure similarity is to use the "product database". Different specifications (attribute levels) mean dissimilar products and identical specifications mean similar or identical products.

One prerequisite for a main-effects Conjoint-Measurement study is: attributes used to describe products should be independent from each other. If this requirement is not met, the use of the additive model to calculate total utilities is not an appropriate method. In that case the researcher should pay attention to interaction effects.

Proceeding on the assumption that the attributes are independent, different product specifications actually mean dissimilar products and identical product specifications identical or similar products. However, like in other fields theory and practice are highly

contradictory. Although researchers try to design studies with independent attributes, you will rarely find these requirements met in practice. (e.g. if “brand” is one of the attributes you will certainly have dependent attribute since “brand” is a ‘medley’ composed of different attributes). On the other side, attributes of that kind are irreplaceable to describe the product. In that case it will be better to use a kind of “perceived” similarity data.

The example of the two minivans “Volkswagen Sharan” and “Ford Galaxy” in Germany illustrates this point. These two cars are identical except the brand/model-name. In other terms the cars do not differ in the product specifications. However, Volkswagen sells far more minivans than Ford. This may certainly be due to the better distribution system of Volkswagen, but also to the brand preference of Volkswagen in Germany opposite to Ford. If one would only consider the pure product specifications (i.e. without the criterion “brand”), both cars would be identical. A model based upon that information would predict the same market shares for both cars. Obviously both minivans are not considered as completely similar.

Another suitable class of data to measure product similarities is the “utilities database”. You can calculate similarity measures based on individual utilities as well as on an aggregate level (refer to section 5.1.2). On the aggregate level, a high variance of the utilities points to a perceived dissimilarity (this is, by the way, equivalent to the use of the product database plus considering the individual importance structure!).

Excursion: standardization of part worth utilities

The result from the estimation in Conjoint measurement are metrically scaled values for each attribute level on an individual basis. For each respondent the estimation algorithm uses a different scale. In order to compare utilities from different respondents the utilities have to be standardized. In other terms, utilities should base on a common neutral point and use identical units.

Different standardization formula are suitable (Gutsche 1994, p. 120 f.). e.g. the respective lowest utility value is used as zero point. The next step is to recalculate the utilities in order to produce comparable units:

- sum up the maximum utilities per attribute and rescale the sum to unity (Gutsche 1994, p. 121 f.) or
- the total utility of the most preferred product equals one (Backhaus et al. 1996, p. 520 f.).

The third suitable data input is additionally collected data apart from Conjoint Measurement. (i.e. the distances produced by a multidimensional scaling are a suitable data input). The distances in a mapping based on multidimensional scaling are interpreted as a dissimilarity measure and are transformed to similarities according to an appropriate rule.

There are two possible ways of data collecting for a multidimensional scaling. First, importance and perception data are collected: the products considered in the simulation are rated by the respondents. Additionally the respondents specify their individual importance structure.

The second method dispenses with the reference to the relevant attributes. The respondents rank product pairs according to their similarity (Green and Tull 1982, p. 433 f.).

The second method is advantageous because the perceived similarity is measured without placing the respondents before a (restrictive) set of attributes. The disadvantage of this method is the expense and difficult evaluation task. We know from experience that a high percentage of rankings are not consistent. Therefore this individual method, performing multidimensional scaling on a individual basis is not always suitable (especially if there is a high number of products considered). The researcher should at least take care of consistence measures (i.e. the coefficient of alienation, stimuli r^2). In addition the presented method is bound up with additional data collecting and analyzing.

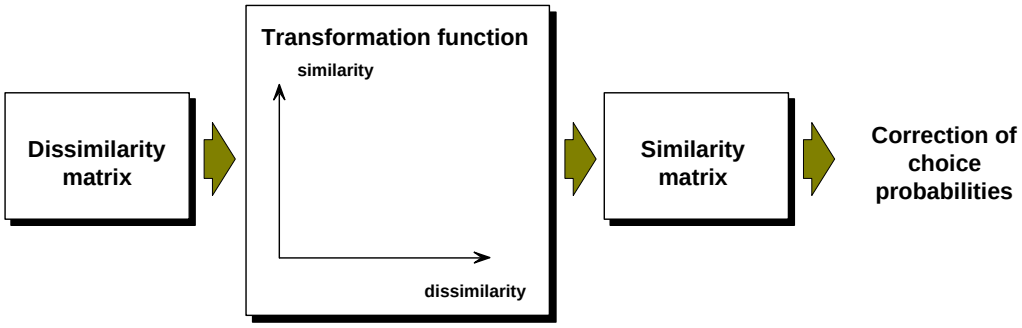
5. Approaches to minimize the IIA-bias

The following section discusses different ways to minimize the IIA-bias. The requirement catalog from section 3 serves as an evaluation scheme. The reflections start with the actual ACA/CBC correction method. Two modifications will be presented for this approach. Contrary to the ACA/CBC method, another approach uses part worth utilities as data input. A third set of "corrections" ("alternative models") will be presented in the last section of this section.

5.1 Post-hoc corrections

The "normal" flow of a post-hoc correction is as follows: first a distance matrix is computed. The elements of this matrix each indicate the dissimilarity between two products. A suitable formula then transforms distances to similarities. In the last step a correction is performed with the help of the calculated product similarities.

Fig. 3 General run of a post-hoc correction



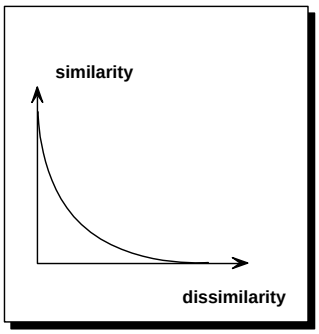
The two following correction methods differ in the used data input, in the transformation function and in the way the corrective factors are calculated.

5.1.2 The ACA/CBC correction

Short description. (cf. ACA Manual) Consider a scale for each attribute where the levels are coded (1, 2, 3, ...). First a dissimilarity matrix is computed. The dissimilarity of a pair of products is taken as the absolute difference between their codes, but with a maximum of 1. In other terms, if the levels between two products are identical the dissimilarity value is 0, otherwise it is 1.

Then the differences are summed up for all attributes (="total dissimilarities"). Two products differing by an entire level on each attribute are maximally dissimilar. Next, dissimilarities are converted to similarities. The first step in this direction is: total dissimilarities are rescaled by a constant so the maximum possible is 3.0. The actual transformation is performed by a negative exponential function. Minimum possible similarity now is .05 and maximum is 1. A further rescaling sets the minimum to 0 and the maximum to 1.

Fig. 4 Convex transformation function



The aim is to calculate a value indicating a product's similarity to all other products. Therefore column totals of the similarity matrix are calculated (="total similarities"). The actual correction consists in dividing shares of preference by the corresponding "total similarities". Total similarities range between 0 and 1. In other terms, where two or more products are identical and totally unlike any others, they divide among themselves the share that each product would have if the others were not present. Finally shares of preference are renormalized to have sum of unity.

The ACA/CBC correction is an "appropriate" procedure for taking similarities into account. This is illustrated in the following 3 product example:

Suppose, products 1 and 3 were identical and totally different from product 2. The similarity matrix then is:

$$S = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

"Total similarities" for product 1 and 3 equal 2 and equal 1 for product 2. In other terms, the correction leads to a reduction of one half in the share of preference for products 1 and 3.

The ACA/CBC correction suggests that evoked sets and product specifications are identical for each respondent. Thus it starts from a kind of aggregate level. The method can easily be converted to an individual approach. In that case one may consider individually composed choice sets, too.

The ACA/CBC correction does not take the individual importance structure into account. Consider the following example:

> Suppose, two products differ in only one of n attributes (n > 1). This attribute is totally irrelevant for an individual, the relative importance equals zero. In other terms the two products are still identical in the individual's opinion. Though the ACA/CBC correction calculates a similarity value < 1.

According to this example the individual importance structure should be taken into account (refer to section 5.1.2 (A)).

Criterion 6 requires that the worsening of a product must not lead to a increased share of preference. As described in section 3 this effect results from an "improvement" by the corrective method. This is the case when the product worsening has the effect that the product becomes more dissimilar to the other products (the "total similarity" is reduced for this product).

In this context the form of the transformation function should be reconsidered. The ACA/CBC correction uses a convex transformation function (negative exponential transformation). Thus an compensation (and thus an increase in the share of preference) will be likely, if the worsening product is relatively similar to the other products (this is the steep part of the transformation function). Slight changes in products will result in relatively big changes in the similarity values. In the flatter part of the transformation function an offset gets more improbable.

This observation suggests a discussion of alternative transformation functions particularly as the reasoning behind the convex functional form is not always clear: What causes a product change relatively bigger "similarity losses" to relatively similar products than to relatively dissimilar products? Alternatively one can use linear, stepwise or concave functions with a kind of "threshold level" between similar and dissimilar products. This would also help to meet criterion 7 ("'slight' product changes only lead to 'slight' changes in similarity and therefore only to minimal changes in the share of preference").

The amount of additional programming of the ACA/CBC correction in its basic form is relatively low. If the suggested modifications are programmed (the individual importance structures are taken into account and individual evoked sets are considered), the amount will increase considerably, but it can still be handled.

5.1.2 An alternative post-hoc correction

Contrary to the ACA/CBC correction another method suggested by Gutsche (Gutsche 1994, S. 150-152) uses the utilities database as data input. As discussed in section 4 the utilities database contains standardized utility values.

The reading of the respective chapter in Gutsche's "Produktpräferenzanalyse" becomes puzzling, because obvious index errors prevent the reader from authentically

interpreting the correction method, though the approach deserves consideration since it contains two interesting ideas:

- The dissimilarities are not only based on distances between individual utilities but also on the variance of the utilities considering all individuals. If the variance of an attributes' utility in a data sample is relatively high this will indicate a kind of perceived dissimilarity.
- As will be described in a later section the actual correction factor is calculated from a comparison between an "empirical relative similarity" and the expected value of the relative similarity.

The following example contains the detailed proceeding:

Example. Consider the purchase of a bicycle. Suppose the relevant attributes were color and price. There are two colors (red and blue) and two different prices ($500 and $1000). The idea is illustrated in a three individuals/three products-case. The considered products are:

Product 1: blue, $500

Product 2: red , $1000

Product 3: blue, $500

The utilities can be read from the following table:

	red	**blue**	**$1000**	**$500**
person 1	-1	1	-2	2
person 2	1.5	-1.5	-1	1
person 3	.5	-.5	-1.5	1.5

Preparatory to the application of the correction the utility values have to be standardized: the respective lowest utility is set to zero and the maximum utilities per attribute are summed up and rescaled to unity, which yields:

	red	**blue**	**$ 1000**	**$ 500**
person 1	0	.33	0	.67
person 2	.6	0	0	.4
person 3	.25	0	0	.75

In the next step a similarity matrix is calculated. It is not clear whether Gutsche means an individual or an aggregated method. That is why the approach is interpreted in both directions:

A. Individual interpretation

Consider e.g. individual 1. His/her utilities can be read from the following table:

	Attribute 1: color	**Attribute 2: price**
Product 1	.33	.67
Product 2	0	0
Product 3	.33	.67

The product similarities are then calculated as follows (only individual 1):

$$S_{12} = 1 - \left[|.33 - 0| + |.67 - 0|\right] = 0$$
$$S_{13} = 1 - \left[|.33 - .33| + |.67 - .67|\right] = 1$$
$$S_{23} = 1 - \left[|0 - .33| + |0 - .67|\right] = 0$$

In other terms products 1 and 3 are absolutely similar (identical) and products 1 and 3 differ completely from product 2. Now, suppose product 1 was red. The utility value for the attribute color is not .33 but 0. The similarities then are calculated as follows:

$$S_{12} = 1 - \left[|0 - 0| + |.67 - 0|\right] = .33$$
$$S_{13} = 1 - \left[|0 - .33| + |.67 - .67|\right] = .67$$
$$S_{23} = 1 - \left[|0 - .33| + |0 - .67|\right] = 0$$

This is plausible, because the price is more important to individual 1 than the color. Products 1 and 2 differ only in price and products 1 and 3 differ only in color. Because the price difference is more important to individual 1 than the difference in colors the similarity between products 1 and 2 is smaller than between products 1 and 3. The individual importance structure is taken into account.

The similarity matrix in the former case (product 1 is blue) is

$$S = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Finally the actual correction for product similarities is performed. Therefore column totals of the similarity matrix are calculated (analogous to the calculation in the

ACA/CBC correction) which yields “total similarities”. After that the “total similarity” is divided by the sum of the “total similarities” (sum of all matrix elements). This quotient (“empirical relative similarity”) indicates the impact of a product on the “total (market) similarity”.

In the example the expected value of the “relative similarity” equals $1/K$ (K = number of products). The correction is performed according to the following formula, e.g. for product 1 (SP = Share of preference):

$$SP_1^{corr} = SP_1^{old}\left[1-\left(\frac{S_1}{\sum_{k=1}^{3} S_k}-\frac{1}{K}\right)\right] = SP_1^{old}\left[1-\left(\frac{2}{5}-\frac{1}{3}\right)\right] = SP_1^{old}\cdot .933$$

The correction factor for product 1 equals .933. The analogous calculation for products 2 and 3 yield 1.133 and .933. This correction has to be performed for every individual. After that, shares of preference will be aggregated in the sample.

According to criterion 1 the above correction is another “appropriate” procedure for taking similarities into account. As demonstrated in the example the share of preference is reduced with those products that have relatively high “total similarities” and vice versa. Due to the correction formula the factors are considerably smaller than those of the ACA/CBC approach. In the limit, if two products (i.e. products 1 and 3) are identical and differ completely from another product (product 2), the shares of preference will not be reduced by .5 but by a value > .5.

This seems questionable at first. But critics argue that the reduction of one half in the share of preference is not appropriate because the effect of a “rich supply” is neglected. In other terms the existence of a rich product supply has a positive effect on the shares of preference. The above explained approach would take this opinion into account.

The correction method considers individual evoked sets. It does not prevent increasing shares in case of a product worsening although a linear transformation function is used. The amount of extra programming is relatively low.

B. Aggregated interpretation

According to Gutsche the dissimilarity between two products is not only based on distances between individual utilities but also on the variance of the utilities measured in the sample. The underlying idea is that a relatively high variance indicates a relatively high (perceived) dissimilarity and vice versa.

The products are "evaluated" as follows:

		Color	Price
Person 1	Product 1	.33	.67
	Product 2	0	0
	Product 3	.33	.67
Person 2	Product 1	0	.4
	Product 2	.6	0
	Product 3	0	.4
Person 3	Product 1	0	.75
	Product 2	.25	0
	Product 3	0	.75

The similarity between products 1 and 2 is calculated with the help of the following formula:

$$S_{1,2} = \sum_{k=1}^{2}(1-\frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3}|u_{i1k} - u_{j2k}|) - 1 = .13,$$

with

$i, j =$ individuals (i, j = 1, 2, 3)

$k =$ attributes (k = 1, 2)

$u =$ utility

The other similarities can be read from the similarity matrix below:

$$S = \begin{pmatrix} .70 & .13 & .70 \\ .13 & .73 & .13 \\ .70 & .13 & .70 \end{pmatrix}$$

Only if all individuals have the same importance structure and two products are identical the similarity between those products will yield 1. (This is the reason for diagonal

elements differing from 1. The variance indicates a perceived dissimilarity). The correction factors are calculated as in the same context above. They come to .9, 1.2 and .9.

The correction method described in the above example indeed considers the individual importance structures, but it does not allow one to model individual evoked sets. Comparisons between simulations with and without correction (individual and aggregate interpretation) show only little differences in the calculated shares of preference. In other terms the corrective factors have only little impact, the calculation of correction factors in the ACA approach appears more appropriate. The amount of extra programming is considerably high.

Excursion: criterion 6 reconsidered. The claim in criterion 6 (exclusion of an increase in the share of preference in case of product worsening) has not been addressed with any of the above methods and is one of the main problems that has to be further examined.

This excursion describes conditions under which the problem occurs. After that it will be shown that working at the individual level helps at least reduce the problem.

When does one need to watch out for the problem? Very generally, the problem will occur

- if the (direct) decline in the share of preference due to the product worsening is relatively small. This effect will be reinforced by a relatively low Logit-Model exponent. In such a case a product change leads to relatively small effects on the share of preference;
- if a product worsening clearly reduces the “total similarity”, e.g. if a product from a very homogeneous set of products is worsened, especially if a convex transformation function is used for correction purposes (cf. 5.1.2).

How does one best address the problem? First, the correction should be performed at the individual level. Only then it is proper to let a product get less share as it gets worse. When similarities are measured for a group, it might be proper to let a product get more share as it gets worse, since all respondents might not agree on what is “worse”. Second, similarities should be measured with utility-weighted differences.

Example. Suppose product 1 and 2 were identical and totally different from product 3. The considered products are

Product 1: red, $1000

Product 2: red, $1000

Product 3: green, $500.

Consider e.g. individual 1. The utilities can be read from the following table:

	Attribute 1: color	Attribute 2: price
Product 1	.5	0
Product 2	.5	0
Product 3	0	.5

The ACA share of preference model with correction for product similarity calculates 25% for product 1 and 2 respectively and 50% for product 3 (logit exponent = 1). Now , suppose product 2 was blue, i.e. a slight ("unimportant") product change:

	Attribute 1: color	Attribute 2: price
Product 1	.5	0
Product 2	**.4**	0
Product 3	0	.5

The ACA-Model then calculates the following shares of preference:

	Share of preference
Product 1	32.4%
Product 2	29.3%
Product 3	38.3%

In other terms the "improvement" via the corrective method compensates the worsening of the product. Now consider an alternative method (the following example presents a correction method composed of selected elements of the above cited methods). Similarities are measured with utility-weighted differences (cf. 5.1.2, A). The similarity matrix in this case is (actually a **linear** transformation function has been used):

$$\begin{pmatrix} 1 & .9 & 0 \\ .9 & 1 & .1 \\ 0 & .1 & 1 \end{pmatrix}$$

Now column totals of the similarity matrix result in a vector of total similarities. Then shares of preference are divided by the corresponding total similarities and then re-normalized to have sum of unity (ACA procedure). This model calculates the following shares of preference:

	Share of preference
Product 1	27.9%
Product 2	23.9%
Product 3	48.2%

Product 2's share of preference is reduced due to the product worsening. This method is more realistic than the ACA approach because it distinguishes between slight and significant product changes. Additionally it does not overvalue the slight similarity changes, because it uses a linear transformation function.

The correction in the above example helps avoid the problem of criterion 6. Additionally one could identify outliers (only "downward") with total utilities lower than n standard units from the mean and remove those products from the evoked set.

5.2 Alternative decision models

Up to now, correction methods have been proposed in which the established shares of preference of the Logit model were corrected by a similarity measure. There may also be alternative decision models to think about, which are easy to apply and which do not have the IIA-characteristic or which at least reduce their influence. In the following, three models are proposed which use as a substance models of the LUCE-family. However, the difference is that these models meet the conditions that apply for the LUCE models, i.e. to examine stimuli which are "alike as possible".

The idea will first be demonstrated by using an example of the automotive sector. After that, a general "cascade model" will be presented. Another example about cellular phone tariffs will form the conclusion.

5.2.1 Automobile example

This method was applied in a simulation model of the German and Italian compact class market elaborated for an automobile producer. The simulation model was programmed with a very high level of itemization, as, among others, the different shapes of car bodies and motor variants were also taken into consideration. That means, e.g. for the Volkswagen Golf, 21 different models were included in the simulation, and for the Fiat Punto, 29. A total of approximately 220 vehicles or models were included in the simulation (universal set). The adjoining classes were included in the analysis (subcompact and middle class). The range of vehicles went from compact cars as e.g. VW Polo, Fiat Punto with 45 PS and prices of approximately 17 TDM to representatives of the middle class as e.g. Audi A3 or Mercedes Benz C 180 with up to 150 PS and prices up to approximately 50 TDM.

In general, no respondent considers the full range of automobiles in his purchase decision. In the "universal set" of an individual there are a lot of automobiles which are

very dissimilar to the ideal car and consequently are not on the short list. In such a case, the use of the Logit model for the calculation of the individual share of preference would lead to results which would not represent the reality.

This is the reason why, in a first step, the "universal set" was reduced to an evoked set by means of answers to direct questions. Questions like the following were asked: "Would you buy a station wagon?" or: "Please indicate the range of prices for your car purchase!" or: "In what range of power should the car you would like to buy have?". Also the answers to the unacceptable-questions from ACA fall within this definition, e.g. "Which of the following brands do you not take into consideration *per se*?"

The consideration of the answers to these non-compensatory questions already limits considerably the individual choice set. e.g. only 43 % of all respondents accept a Japanese car, and only 46 % accept an Opel (GM). Only 71 % of the respondents would take into consideration to buy a compact car. To say it with other words, those cars are sorted out which in reality do not have any relation of substitution to the favorite car types.

Nevertheless, in these reduced sets there will be a high grade of heterogeneity in the seized similarity and so in the possible relations of substitution. The absolute extent of a similarity measure is not decisive for the consideration of the IIA-bias, but the relative similarity. Therefore the evoked set will be further limited to **only relatively similar** cars. There are some possibilities to accomplish this step:

- The first option is to consider only those cars which are similar to a "favorite car". For that purpose it is necessary to question about the favorite car in addition. In this study only those respondents were allowed who either had bought a car one year ago or earlier or who were planning to buy a car in the near future. The purchased car/brand or the car/brand which should be bought shortly was considered as "favorite car".

  All the above discussed patterns of data (product database, preference database and additional data) may be taken into consideration in order to determine the similarity to the favorite automobile. The weighted Euclidean distance of the product specification may be used as a measure of similarity or distance (one may as well use the distance between the part worth utilities or anything similar, cf. section 5.1).

  Also the expert knowledge of the analyst should be consulted in order to define the evoked set. Unplausible compositions of the evoked set may be reached by the setting of a maximum dissimilarity. So there will only be relatively similar cars in the evoked set. The application of the Logit model does not have the IIA problem, though. Various analysis have shown that post-hoc-corrections nearly do not have any effects if such a limited evoked set is applied.

- Another method which may lead to the same results as the one just discussed is the following: Only those n-cars which show the highest total utility are chosen out of the remaining elements of the evoked set. It is possible to fix the amount n or to consider only these cars which are no outlier "downward" (e.g. by means of the variance of the total utility measured). A "normal" logit analysis is carried out with the remaining automobiles. Also in this case the IIA-bias should not play an important role any more.
- A third possibility is to take one of each brand/model, actually the one which is most similar to the favorite, and add it to the evoked set (unless one brand/model is not acceptable). This reduces the bias which is created by the different amount of models of one brand/model. Again, all of the above mentioned methods may serve to determine the similarity. Also in this case a maximum limit of dissimilarity should be fixed. The condition for this proceeding is, though, that the brand really was cited as the most important characteristic by the respondent (cf. to this also the mobile radio example in section 5.2.3).

The idea of this proceeding is to limit the market, which from the buyer's point of view is heterogeneous, to a small amount of choices with only a small number of stimuli "alike as possible". The results of the models of the LUCE-family will then not be affected by the IIA-bias.

5.2.2 A "cascade model" by M. Kneller

A comparable idea is pursued by M. Kneller and his "cascade model". The proceeding is explained in the following schedule:

1. Compose for each individual a relatively narrow evoked set by means of direct questioning and unacceptable-questions (cf. description in section 5.2.1).
2. Find out the respectively best and worst product by means of the total utility.
3. Put every further product of the evoked sets in order either to the best or the worst product. The criterion of order is the sum of the squared distances of the standardized utilities. The result of the first step of this procedure is the splitting in two groups. cf. e.g. the following table:

Step	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6	Product 7	
1	1	0	1	1	0	0	0	1
2								
3								
4								

1 = is more similar to the best product of the group

0 = is more similar to the worst product of the group

Supposing that in the example product 1 might be the best and product 7 the worst product. The products 3, 4 and 8 are more similar to product 1 than to product 7 and will be classified with the first group. The products 2, 5 and 6 on the other hand are more similar to product 7 than to product 1 and therefore will be classified with the second group.

4. Strike an arithmetic mean upon the total utilities in the new groups and distribute the share of preference by means of the Logit model among the composed groups. In the example may result e.g. the following shares of preference:

	Group 1 (Products 1, 3, 4, 8)	**Group 0 (Products 2, 5, 6, 7)**
Share of preference	.6	.4

5. Repeat all the steps beginning with 2. Find out, then, one best and one worst product in each of the newly composed group. After that, classify the elements of one group with the respectively best or worst product of the group and calculate the shares of preference of the newly composed group. e.g.:

Step	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6	Product 7	Product 8
1	1	0	1	1	0	0	0	1
2	1	0	1	0	0	1	1	0
3								
4								

Example of reading: Product 3 which had been classified to the "better" group in step 1, is now as before more similar to product 1 than to the worst product of group 1. Product 4, on the other hand, is now more similar to the worst product of group 1.

The shares of preference will then be calculated by multiplication of the shares of preference in the single steps. e.g.:

	Group 1 (Products 1, 3, 4, 8)		**Group 0 (Products 2, 5, 6, 7)**	
Share of preference of step 1	.6		.4	
	Group 11 (Products 1,3)	**Group 10 (Products 4,8)**	**Group 01 (Products 6,7)**	**Group 00 (Products 2,5)**
Share of preference of step 2	.7	.3	.65	.35
Total share of preference of steps 1 and 2	**.42**	**.18**	**.26**	**.14**

The proceeding will be continued until there is a share of preference for each product.

The demonstrated model to analyze the share of preference divides the situation of decision of a buyer in "paired comparisons". The calculation of the share of preference in the case of only two alternatives in each step of the procedure is *per definitionem* not affected by the IIA-bias (hereto there would have to exist at least three alternatives).

5.2.3 An example with cellular phone tariffs by J. Engelke (a "nested logit" approach)

This example is taken from a study in the cellular phone branch. In this study a simulation model was established in order to prognosticate the market shares of individual mobile radio offers (tariffs). The criteria in this study are, among others, network providers, tariffs, coverage, quality of reception.

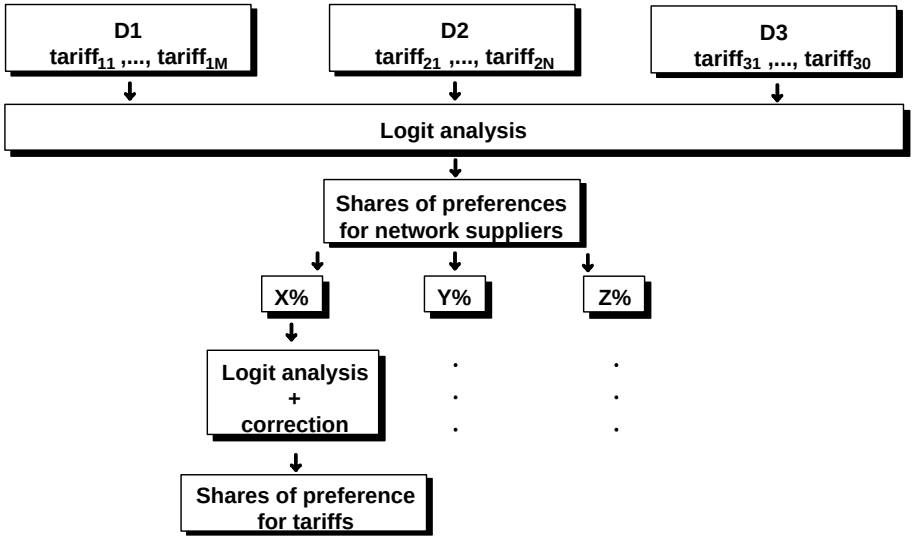
It was known from earlier studies in the cellular phone branch that the buyers and the potential buyers orient themselves first of all on the network suppliers, (i.e. before choosing a certain tariff they decide upon the network provider). In Germany there are three main suppliers: D1, D2, and E-Plus. There are several cellular phone offers (tariffs) by different providers for these network suppliers.

The first step in this model was to find out tariffs with the maximum utility of each network provider (first-choice model). A logit analysis was carried out with these three offers. By that, the choice probabilities of the three network suppliers are quasi determined. Herewith, the IIA problem is avoided, since only one representation of each network supplier will be included in the simulation.

In a second step, the ascertained shares of preference of the network suppliers are distributed among the individual tariffs. This happens by means of a "normal" logit analysis, in the course of which a post-hoc correction method is used in addition.

In a second step, the ascertained shares of preference of the network suppliers are distributed among the individual tariffs. This happens by means of a "normal" logit analysis, in the course of which a post-hoc correction method is used in addition.

Fig. 5 Formation of decision hierarchies in the cellular phone example



The formation of decision hierarchies happens in a way that the IIA-bias does not play any role in each level. In the same way, a decision hierarchy could be developed in the bus example which has been discussed in section 1. The decision between the means of transportation bus and car would happen on a first level. In a following step, the commuter would have to decide upon the red and the blue bus. This procedure is of course only applicable to a case in which the hierarchies are known (cf. Ben-Akiva and Lerman 1994, p. 54).

6. OUTLOOK

In an application oriented research, such criteria like simplicity of the model structure, easy application and speed of the analysis are especially important. The models of the LUCE-family match these criteria as decision models among the brand selection decision.

If such an analysis is in a way faulty (e.g. because of the IIA-bias), this instrument is not acceptable. The present essay is supplying a “tool box” which may be used to correct subsequently or to avoid from the beginning any distortions arising by the IIA-characteristic in the family of LUCE-models. If this is possible by means of the procedures described above, the application of the LUCE-models is still justified and given preference over the more “academic models” e.g. the generalized Logit or Probit-models, elimination-by-aspects, ... (cf. Hermann 1994 and Skiera 1995).

The present essay discusses and evaluates different approaches. Further research activities still have to elaborate the following points:

- Unambiguous methods have to be developed to measure the heterogeneity of the products. They may serve as a decision basis to decide if a correction procedure has to be applied or not. This aim could also be achieved by means of appropriate tests.
- The above mentioned methods could also be used in order to measure the reduction of dissimilarity by the correction procedure.
- The claim for a procedure which excludes an increase in the share of preference in case of product worsening, as it is mentioned in criterion 6, has to be further examined concerning the procedures pointed out in this essay.

CITED AND FURTHER LITERATURE:

Backhaus, Klaus, Bernd Erichson, Wulff Plinke und Rolf Weiber, Multivariate Analysemethoden, 7. Auflage, Berlin, 1994

Balderjahn, Ingo, Ein Verfahren zur empirischen Bestimmung von Preisresponsefunktionen, in: Marketing, Zeitschrift für Forschung und Praxis, I. Quartal 1991, Heft 1, p. 33-42

Balderjahn, Ingo, Marktreaktionen von Konsumenten, Berlin, 1992

Bechtel, Gordon G., Share-ratio of the nested multinomial logit model, in: Journal of Marketing Research, May 1990, p. 232-237

Ben-Akiva, Moshe E. und Steven R. Lerman, Discrete Choice Analysis: Theory and Application to Travel Demand, Cambridge, 1985

Currim, Imran S., Predicitve testing of consumer choice models not subject to independence of irrelevant alternatives, in: Journal of Marketing Research, May 1982, p. 208-222

Green, Paul E. und Donald S. Tull, Methoden und Techniken der Marketingforschung, Stuttgart, 1982

Gutsche, Jens, Produktpräferenzanalyse, Berlin, 1994

Herrmann, Andreas, Die Bedeutung von Nachfragemodellenfür die Planung marketingpolitischer Aktivitäten, in: Zeitschrift für Betriebswirtschaft, Octobre 1994, p. 1303-1325

Jain, Dipak C. and Frank M. Bass, Effect of choice set size on choice probabilities: An extended logit model, in: International Journal of Research in Marketing (1989), p. 1-11

Kamakura, Wagner A. and Rejandra K. Srivastava, Predicting choice shares under conditions of brand interdependence, in Journal of Marketing Research, November 1984, p. 420-434

Luce, R. D., Individual Choice Behavior, New York, 1959

Malhotra, Naresh K., The use of linear logit models in marketing research, in: Journal of Marketing Research, February 1984, p. 20-31

Mc Fadden, Daniel, Econometric models for probabilistic choice among products, in Journal of Business, vol. 53, no. 3, pt. 2, p. S13-S29

Skiera, Bernd, Implikationen des allgemeinen Probit-Modells für die Marketingplanung, in: Zeitschrift für Betriebswirtschaft, February 1995, p.191-198

Timmermans, Harry; Borgers, Aloys and Peter van der Waerden, Mother logit analysis of substitution effects in consumer shopping destination choice, in: Journal of Business Research 24, 1992, p. 177-189

COMMENT ON PAFFRATH

Jon Pinnell
MarketVision Research, Inc.

Probabilistic choice models such as BTL or logit are known to suffer from a limitation referred to as independence from irrelevant alternatives, or IIA. While the phrase sounds at least innocuous (if not beneficial) the effects can produce widely biased share estimates. IIA is sometimes referred to as the constant-odds principle.

IIA is a problem when two products in a choice set are more similar than other products in the same choice set. This is commonly illustrated with the red bus/blue bus example. Imagine the probability of driving to work is 0.90 and of taking the blue bus is 0.10. Now imagine that a red bus service is introduced, identical to the blue bus in every way but color which does not influence choice. Logit/BTL models would estimate the resulting shares to be: drive 0.8181, blue bus 0.0909, red bus 0.0909. Note that the ratio between drive and the blue bus remained constant at 9:1 (constant odds). The probability of taking a bus, however, increased from 0.10 to 0.1818—suggesting that no matter how bad a product, shares will always increase with additional choice options. This unintuitive result stems from logit/BTL assuming that new products will always steal share from existing products in proportion to their existing share. Rather, new products should steal share disproportionately from "similar" products.

Currently, the ACA and CBC simulators both provide post hoc computational adjustments to reduce the problems of IIA. The "adjustment for product similarity" creates a matrix of product similarities and scales the resulting shares by the inverse of the similarities. To illustrate this point using the previous example, assume the car is maximally different from the two buses, but the buses are identical to each other. The product similarities would therefore be:

car	1	(similar only to itself)
blue bus	2	(similar to itself and one other product – the red bus)
red bus	2	(similar to itself and one other product – the blue bus)

The unadjusted shares are scaled by the inverse of these similarities, and repercentaged as shown below:

Option	Unadjusted Share	Similarity		Adjusted Share	Repercentaged Adjusted Share
Car	0.818181		0.81818	0.900	
Red Bus	0.090912		0.04545	0.050	
Blue Bus	0.090912		0.04545	0.050	
				0.90908	1.000

At least in this example, the results are far more appealing, with the two buses jointly accounting for a 0.10 probability. In practice, however, the Sawtooth adjustment for similarity has been found to have several limitations.

First, product changes which make a product worse but also more unique frequently produce increased shares of preference. For example, assume three products have many similar features including price. If one product increases its price slightly it should be less attractive (have a lower share of preference). With the adjustment for similarity, however, the changed product is more unique and will benefit from that change—many times with an increased share.

Second, attribute level differences are treated equally, without regard for the importance of the attribute. Two products might be identical on several attributes that have no influence on choice and very different on one attribute that substantially influences choice. These products would be viewed as very nearly identical when in fact they are very different to the decider.

Third, occasionally very minor changes in product specification can have substantial impact on resulting shares. In fact, the Sawtooth Software manuals provide a warning for these counterintuitive findings.

The previous paper presents examples of these and other problems in dealing with IIA violations in choice simulations. The author has taken the stance of a good method is an easy method, which has some merit, but I am also reminded of a quote from H. L. Menken, “For every complex problem there is a solution that is simple, neat and wrong.”

Sticking with the easy methods, for now at least, I believe the author has presented two key points in adjusting for similarity:

First, similarity should be determined based on attribute importances.

Second, similarity should be determined at the level of the individual.

Both would represent modifications to the current Sawtooth Software approach.

I also believe that the author has missed an easy method that works well in several instances. Let’s consider the red bus/blue bus scenario again. We have found that our

shares of preference for the bus options are biased upward because the bus alternative is included in the consideration set twice. The ACA/CBC correction was demonstrated above. Another approach to correcting the bias would be to include the car option in the consideration set twice and ignore the correction for similarity. This approach will produce more intuitive and more stable results than using the correction for similarity. This is a useful approach, for instance, when one wants to conduct sensitivity analyses on a product that is similar to at least one other product.

The author also includes nested logit as an easy method. It is not clear that nested logit should be included as an easy method or not, but it should be included as a potentially dangerous method. The hierarchy structure as discussed is specified by the researcher. Resulting shares of preference can be highly susceptible to errors in specifying this hierarchy. When specified correctly, though, the computations aren't overly complex.

Excluded by the author as a complex method is probit. Probit is most likely to offer the long term solution to IIA limitations encountered in logit models. The computations behind probit are intensive, involving multiple integration, and software programs are limited in availability/accessibility for most users. Software will inevitably be more available in the future, and efficient computational approximations are likely to emerge.

Probit is generally superior to logit because of the assumptions relating to the products being simulated. Logit assumes simulated products to be uncorrelated. In actuality, the correlations are not required to be zero, but they must be nearly equal. Therefore, in a two product logit simulation, IIA is never a problem. In simulations of more than two products but equal product similarities (correlations), IIA is again not a limitation. Only in instances with more than two products and different similarities (correlations) does IIA produce biased results. In probit models, simulated products can have any level of varying correlation without negatively impacting the resulting shares of preference.

The one method that the author presents that I am most intrigued by is the cascade model. In this model, the best and worst products are identified and each remaining product is assigned as more like one or the other. Then, one simulation is conducted on two "constructed" products. The two products represent the average of the products most like the best product and the average of the products most like the worst product. Then, the process is repeated forming another two groups of products, separately within the best products and within the worst products. This is repeated until each product forms its own group. The final shares of preference are the multiplicative products of the shares from each step. In this way, we form a tree, but the tree is based on product similarities. Also, since at each stage we only have two products, violations of IIA are not a concern.

For all of the methods, though, I am interested how different the simulated shares actually are.

THE NUMBER OF CHOICE ALTERNATIVES IN DISCRETE CHOICE MODELING

Jon Pinnell and Sherry Englert
MarketVision Research, Inc.

ABSTRACT

When designing a discrete choice experiment, researchers must decide how much information to collect from each respondent, balancing the need for information against the burden on the respondent. Increasing the number of alternatives in each task provides greater statistical efficiency, but respondents may have more difficulty processing all of the information. This paper reports the results from three experimental studies that varied the number of alternatives within and between respondents. We conclude that respondents are capable of responding accurately to choice tasks with a relatively large number of concepts. In fact, we provide evidence that it is probably advantageous to use a number of alternatives per task greater than two.

BACKGROUND

As researchers, we frequently make trade-offs between the precision of the research we design and our respondents' ability and willingness to provide accurate responses in the way we hope. Unfortunately, in our attempt to collect information that is more precise, we may inadvertently decrease the quality of the measurement itself. The purpose of the current work is to investigate the sometimes conflicting objectives of statistical efficiency and respondent burden in discrete choice studies.

Recent works have outlined a number of approaches to learn more from each respondent in a discrete choice interview. We group these approaches into two categories:

1. Collect more information from each respondent by asking more questions
2. Collect more insightful information from each respondent by carefully constructing the choice tasks but not asking additional questions

The details of these approaches are outlined below:

1. Collect more information from each respondent by asking more questions

Number of Tasks

Discrete choice studies typically ask respondents to provide choices from six to twelve choice sets or tasks. One approach to increasing the number of choice observations is to ask each respondent to evaluate more tasks. That is, one can either have 200 respondents give 6 choices or 100 respondents give 12 choices. While it is clear that the number of choices provided is equal in both cases, it isn't immediately clear if these provide equivalent results or represent a good practice.

Louviere (1993) reports on two studies which each identified slight reliability decreases among ratings tasks with a relatively large (32) number of profiles. More recently Brazell and Louviere (1997) have found that the utilities from late tasks differ from early tasks by only a multiplicative constant and that data quality increases to a point, around 40 to 60 tasks, and then degrades. This finding of scale differences indicates that late tasks and early tasks differ in the amount of "noise" in the data, but that the underlying utility structures are the same. Even more substantial are the findings from Johnson and Orme (1996). In a meta-analysis of five commercial studies that included 15 or more choice tasks, not one showed a decrease in reliability in the second half of the tasks relative to the first. In fact, all but one of the five showed a slight increase—one produced no change in reliability. Late tasks also had a higher correlation with the pooled model (based on all tasks). Johnson and Orme, however, did show a "shift" in attribute importance, specifically between price and brand. It remains unclear whether earlier or later choices are more valid.

While these results are generally comforting, the fact that late tasks become more alike and seem to differ from early tasks might indicate the occurrence of respondent learning or respondent simplification. Respondent learning certainly takes place as evidenced by a monotonic decrease in time taken per task throughout a choice exercise. However, respondent simplification—paying differential attention to particular attributes due to boredom or fatigue—would be a troublesome finding.

Overall, though, it appears that respondents are capable of providing reliable choices for up to 20 tasks, and maybe more. It appears this is a reasonable way to increase the number of observations, and therefore the relative efficiency in choice studies.

Second Choices

An alternative approach of asking more questions involves asking more questions for each choice task. Typically, respondents are presented with a choice task of three to five concepts and asked to select the one they would purchase or that they most prefer. Unfortunately, all the researcher learns is which one the respondent most likes. No information is provided about the remaining concepts. Some researchers have wanted to eliminate this waste by having respondents give both a first and second choice or even rank-order the concepts—indicate their first choice, then second choice, and so on until they have indicated their rank-order preference for all concepts shown in a task.

Several researchers have investigated the area of second choice and probing depth. Pinnell and Huber (1996) investigated the effect on first choices and found that respondents who knew they were being asked to provide a full rank order might have provided slightly better first choices. However, the authors also found that second choices look different than first choices, contribute very little explanatory power, and increase cost (time) by about 15%. In the same meta-analysis mentioned above, Johnson and Orme strongly confirmed a consistent bias in second choices. They found that second choices had a smaller scale (indicating more "noise" in the choices). More problematic, the authors also found that interior levels of attributes violated the linear form that would indicate congruence with first choices. The interior levels were consistently biased upward. The authors were unable to explain the source of the bias. Through computer simulations, however, they did determine the effect was psychological (at the level of the respondent) and not algorithmic (based on logit or similar assumptions).

Walsh and Schmittlein (1997) also confirmed the occurrence of a bias in second and third choices. They explored utilities independently estimated from first, second, third and fourth choices in predicting actual choices. This predictive ability is a more meaningful test than seeing that the utilities simply look different (but might predict identically). The authors found that first choices predicted first choices well, predicted second choices less well, and predicted third choices even less well. Similarly, utilities developed from second choices predicted second choices better than first or third choices. They too were unable to provide an explanation for this bias in second or third choices.

Overall, it appears that asking second choices is not only costly from a time perspective but adds a consistent bias to choice predictions. We conclude that asking second choices is an ineffective way to increase efficiency in choice studies and has a deleterious side effect.

Both previous alternatives constitute collecting more information from each respondent by asking more questions, either by having the respondent indicate more first choices or by probing more deeply. An alternative approach to increasing the amount of information collected is to collect the same number of choices but make sure that each choice provides more information.

2. Collect more insightful information from each respondent by carefully constructing the choice tasks but not asking additional questions

Utility Balance

One approach to making each choice more informative is to equalize, as nearly as possible, the choice probabilities of each alternative. By eliminating dominated concepts, every choice provides a greater opportunity for respondents to provide insight into their utility structure. 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 compared to "easy" (or dominated) pairs. In ACA, this design criterion is considered second to level balance. As it turns out, utility balance is frequently at-odds with other design criteria. Utility balance in discrete choice modeling must strike a compromise between three other design criteria: orthogonality, level balance, and minimal overlap.

Utility balance can be traced back to before the days of computer based interviewing. Specifically, Thurstone's law of comparative judgment develops scaling based on the discriminal process between stimuli, such that unlike objects are placed far apart on a scale.

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% fewer respondents. Their approach requires prior knowledge of the utilities. It appears that even quite misspecified utilities are better than a null assumption of all $\beta s = 0$. Therefore, it can be inferred that some utility balance is better than no utility balance. However, the task of developing even misspecified priors might be difficult. And the work isn't over then, even the authors describe the process of identifying an efficient task, given the priors, as tedious.

We view the largest limitation of the Huber and Zwerina utility balance is that its application is today specific and limited to fixed-design choice tasks. To explore the value of utility balance, even in randomized designs, Huber, Zwerina, and Pinnell (1996) conducted a within subject analysis of existing data. The authors divided each person's choices into those with the most and least randomly developed utility balance. Separate models were estimated for each set of tasks, pooling across respondents. By balancing within each respondent, the analysis identifies the relative benefit from utility balance. The results indicated that although utility balance allows equal precision with 30% fewer respondents, it increases the time to complete a task by only 7%.

In essence, the effect of utility balance is to make the choices maximally difficult to produce the most information from the respondent, provided the respondent can deal with

the added complexity of the task. We conclude utility balance to be a useful approach to making each respondent choice more meaningful. In fact, the implementation in CBC would be welcome, even if it means giving up the ability to analyze by "counting".

We believe that validation is required, though, to ensure that the results from balanced choices are more predictive of in-market behavior. That is, we lack validation to ensure that balance is not promoting respondent simplification.

Number of Alternatives per Task

Utility balance seems to work well because it makes tasks more difficult. An alternative method to produce more information from each choice task might be to have each respondent indicate their first choice from a larger choice of options or concepts. This also will make tasks more difficult, but in a different way.

Imagine a simple example of identifying a favorite product out of a set of six. Two extremes are possible to identify the winner. First, a respondent could be presented with pairs of products, eliminating "losing" products until there was a single winner. Alternatively, a respondent could be presented with the six products and asked to pick one. In the first case, after five pairwise comparisons (and assuming no intrasensitivities), we would have a winner. In the second case, one question provides the winner. If a respondent answers all five questions reliably, the first alternative provides more information (analogous to second choices), but would clearly take longer.

This logic can be applied to discrete choice modeling as well. Most people analyze discrete choice studies by using multinomial logit (MNL). Multinomial logit is as concerned with which concepts are *not* chosen as it is with which concept is chosen. Therefore, one approach to increasing the statistical efficiency of a choice design, while not asking more questions, is to have respondents select their choice from a larger choice set.

To help make this reasoning more concrete, think about the number of pairwise inequalities created by choice sets of different size. We can use our earlier example of picking a favorite brand from a set of six.

	Pairwise Comparison	All Product "Portfolio"
TASK:	Pick one	Pick one
SET:	a b	a b c d e f
WINNER:	a	a
INFERENCES:	a > b	a > b a > c a > d a > e a > f
INFORMATION INDEX: (# of inequalities created)	1	5

It would be expected that the portfolio task—picking a favorite from a set of six—would be a more difficult respondent task. However, unless the task of picking a favorite brand from a set of 6 is more costly by a factor of five than picking a favorite from a set of two, it seems like a beneficial approach.

Again, from a discrete choice standpoint, this would suggest that choice designs should include more alternatives per task. Statistical efficiency, then, could be equalized with fewer respondents or fewer tasks.

Very little research has been focused on this issue. Louviere and Woodworth (1983) include a brief discussion of the efficiency of alternative designs and conclude, based on the coefficient of variation of the choice probability, that pairs produce less efficient designs.

Bunch, Louviere, and Anderson (1991) report simulation results for the statistical efficiency of a number of design strategies, some of which vary between 2 and 9 alternatives. However, their results are primarily focused on differences between design strategies and not number of alternatives. But more importantly, their results deal with expected efficiencies based on computer simulations.

Both papers conclude that paired comparison choice tasks are less efficient from a design perspective, reinforcing our expectations as developed above.

Among many conjoint and choice researchers, however, there is a concern that as tasks become too burdensome (include too many concepts) respondents will have difficulty responding in a reliable fashion. Therefore, the belief has implicitly been that simple choice tasks are no worse in practice than more complex tasks, even if knowingly less efficient.

The purpose of this paper is to compare how respondents, not machines, deal with the added complexity from the increase in number of alternatives per task.

We report the findings from three different studies, each of which included an experimental choice modeling component. In each study, respondents were randomly assigned to treatments which varied the number of alternatives they saw or the order in which the alternatives were presented. We will detail the research designs below.

EMPIRICAL DATA

The empirical findings draw from three independent datasets. These datasets varied in the number of respondents, the number of attributes, the number of levels per attribute, the audience, and the product category.

Two of the three studies were commercial applications of choice modeling. The first study was among approximately 300 health benefits managers and the second was among 250 consumer respondents for a consumer durable. In both studies, respondents were presented with nine choice tasks consisting of 3 pairs (2 two alternatives per task), 3 triples (three alternatives), and 3 quads (four alternatives). Each task also included a default choice or "None" option. Half of the respondents saw the 9 tasks in this progressive order and half saw them in reverse order—balancing any order effect across the extremes in the number of alternatives treatment. Our primary interest was in the extremes in the number of alternatives, the pairs versus the quads. Therefore, the analysis from this study was conducted by pooling the pairs across both order treatments and separately pooling the quads across both order treatments.

The third study was conducted from a MarketVision Research experimental research budget. This study represents 260 respondents split between three experimental cells. Two of the cells are of direct interest in this investigation. In the first cell, respondents indicated 12 first choices from seven alternatives plus a default and then responded to 8 first choices from pairs. Respondents in the second cell indicated 12 first choices from pairs and then 8 first choices from tasks of seven alternatives. The blocks of discrete choice questions were separated by a series of profiling questions.

All three studies relied on randomized choice designs and computer-aided self-interviewing.

EVALUATION CRITERIA

Several criteria are used to report our findings. They are broken into three categories of cost criteria, congruence criteria, and efficiency criteria. Each is introduced below.

Cost Criteria

To evaluate the "goodness" of an approach, we must consider what we are forced to give up to get the responses. In terms of choice studies, the greatest costs are time, measurement error, and use of the default choice.

Time is the measure of cost to the respondent to read the concepts, consider the options, and provide a response. We evaluate time in terms of median response time in seconds for each choice task.

Measurement error has multiple components. The first is random error, which degrades the predictive validity of a particular approach. As such, we will consider random error the complement of predictive ability in the next section. Conversely, systematic error, as would be expected by processes like respondent simplification, will affect predictive ability, but should be considered separately and as a cost because it won't cancel itself out with large samples. Even respondent simplification can manifest itself in several ways. We specifically consider respondent simplification in two ways. The first way is a comparison of attribute importances between results when the number of alternatives is varied. This form of respondent simplification will also be considered as a congruence criterion, and will be discussed in more detail elsewhere. The other form of respondent simplification, and the one studied more frequently in conjoint-related fields, relates to position bias. Therefore, the conjoint corollary to errors of primacy and recency is discussed as a cost of simplification.

In the third study, we actually requested respondents to recall the level of each attribute that was in their most recently selected alternative. We report a proven recall measure which indicates what portion of time the claimed selection matched the actual selection. This measure is the complement of simplification.

The third cost, and the most controversial, is the use of the default alternative. We believe that many factors influence respondents' use of the default alternative. One of the many factors is task difficulty. After their meta-analysis of several commercial choice studies, Johnson and Orme conclude that task difficulty does not influence the use of a default. We posit, however, that tasks composed of varying numbers of alternatives could cause respondents to use a default alternative in different ways, not in the economic sense as commonly suggested, but in the psychological decision avoidance sense, as suggested

by Huber and Pinnell (1994) and supported by Tversky and Shaffir (1992) and Dhar (1992, 1997). Our reason for introducing None usage is that as the usage of the default increases (for any reason), the efficiency of randomized choice designs decreases. We would not propose that excluding the default choice alternative is the appropriate response, but we would rather have relatively low usage of the default rather than relatively high usage.

Congruence Criteria

In addition to the relative costs of alternative methods, we also consider the relative merits of each approach. We are interested in the similarity of the results produced by the alternate treatments. Specifically, we consider attribute importances, utilities (both before and after accounting for possible scaling differences), and predictive ability in cross-task comparisons.

Efficiency Criteria

In keeping with the established norm (Kuhfeld, Tobias and Garratt 1994, Bunch, Louviere and Anderson, 1994, and Huber and Zwerina 1995), we evaluate the statistical efficiency of designs in terms of D-efficiency. Specifically, we consider relative D-efficiency of zero-centered (utility neutral) attributes in a design matrix.

FINDINGS—COST CRITERIA

Choice Times

The first criterion to evaluate is the time it takes respondents to answer choice tasks of different sizes. Response times are consistently halved after the first three or so tasks, and most of the respondent's learning has occurred by the ninth or tenth task.

Recall that in the first two studies, respondents indicated their first choice from pairs and quads as either the first, second, and third tasks or seventh, eighth, and ninth tasks. The order of presentation was balanced across respondents. In this way, we can explore the average difference (ratio) of times between pairs and quads. The response times are summarized in the following table:

Median time per task (in seconds)
Averaged across six tasks

	STUDY	
	ONE	TWO
2 Alternatives	17	9
4 Alternatives	24	12
Ratio	1.43	1.28

It would appear that the additional alternatives (four instead of two) adds roughly one third of the time of pairs alone. The difference between the two studies is relatively large, but it would appear the quads provide three times the information (based on number of inequalities presented) for only marginal cost increase.

In these two examples, both the number of respondents and number of tasks was somewhat limited. In the third study, respondents indicated far more choices, allowing greater analysis. The following table is aligned by number of tasks. Recall that respondents would have switched columns halfway through the exercise. That is, those respondents who first answered pairs ended answering sevens.

Median time per task (in seconds)
Comparison by Number of Alternatives by Task Order

STUDY THREE

	Number of Alternatives		Ratio
Task	2	7	7:2
First Set of Tasks			
1	13	27	2.08
2	10	20	2.00
3	10	16	1.60
4	10	14	1.40
Average (first 4)	**10.75**	**19.25**	**1.79**
5	11	18	1.64
6	8	14	1.75
7	10	14	1.40
8	8	12	1.50
Average (second 4)	**9.25**	**14.5**	**1.57**
9	7	13	1.86
10	8	13	1.63
11	10	15	1.50
12	8	11	1.38
Average (third 4)	**8.25**	**13.0**	**1.58**
Second Set of Tasks			
13	13	14	1.08
14	9	9	1.00
15	7	8	1.14
16	6	7	1.17
Average (first 4)	**8.75**	**9.5**	**1.09**
17	8	9	1.13
18	6	6	1.00
19	6	7	1.17
20	6	7	1.17
Average (second 4)	**6.5**	**7.25**	**1.12**

In analyzing the table, it initially appears that choices made from seven alternatives take about sixty percent longer than choices from two alternatives. Note that there is some variation of the ratios among the first twelve tasks, suggesting that learning occurs

at a slightly faster rate among the treatment with seven compared to the pairs. In analyzing tasks 13 through 20, however, the ratios behave far differently. In tasks 13 and 14, it would appear there is no difference between the time required for two or seven alternatives.

We were not surprised that people could learn "more" on how to answer sevens relative to pairs, and therefore decrease the ratio between the two treatments. We were somewhat surprised that in the second set of tasks, the pairs and the sevens were so similar in their times.

Decrease in Relative Time

	Number of Alternatives	
	2	7
Task 1 through 4	10.75	19.25
Task 5 through 8	9.25	14.50
RATIO	**0.86**	**0.75**
Task 5 through 8	9.25	14.50
Task 9 through 12	8.25	13.00
RATIO	**0.89**	**0.90**
Task 13 through 16	8.75	9.50
Task 17 through 20	6.50	7.25
RATIO	**0.74**	**0.76**

If respondents could learn how to answer sevens at a quicker rate (as is shown in the first ratio above), we would have expected the time for the sevens the second time around (beginning with task 13), to produce a time spike (as the pairs partially did), but also for the times to decrease more quickly—neither of which happens.

We offer two conjectures to this seeming quandary:

One, it is possible that the random assignment of individuals worked against us and we ended up with one group of people who could read and process their choices more quickly than the other group.

Two, alternatively, it is possible that the respondents who first evaluated tasks with two alternatives per task developed processing heuristics that were different from those respondents who first evaluated tasks with seven alternatives per task. Then, when the respondent task changed (from two to seven alternatives), their heuristics first prompted too simplistic a response relative to the other respondents.

Systematic Measurement Error

The second cost criterion we evaluate is systematic measurement error. While the number of potential sources of error is huge, we specifically consider one systematic error—respondent simplification. Two specific respondent simplification schemes will be considered.

One form of simplification is based on attribute importance. In some instances, respondents dealing with too burdensome a task will focus only on attributes they view as important. That is, respondents will pay relatively more attention to important attributes and relatively less importance to unimportant attributes.

The second form of simplification is based on attribute position or order. This form of simplification manifests itself as would the traditional question and response order effects of recency and primacy. The effects of response order bias are detailed in Schuman and Presser (1981) and their impact on conjoint methods is discussed by Johnson (1981, 1989) and Chrzan (1994). Johnson reported the results for two full-profile ratings-based hold-out experiments while Chrzan reports the results for choice-based pairs. In all three instances, attribute order effects are seen impacting attribute importances.

While simplification has a negative connotation, it is not clear that simplification based on attribute importance is bad, and would be far less dangerous than attribute position effects. In fact, it might better represent actual purchase decisions where shoppers are potentially dealing with more information than they can process.

The third study under consideration includes a direct and explicit measurement of respondent simplification. Our attempt to measure simplification was to ask a subset of respondents, after two specific choice tasks, a series of follow-up questions dealing with simplification. After making their choice, respondents were asked for each attribute: a) if they recalled the particular level in the concept just selected, and if so, b) what was that level.

The questionnaire was computer administered so the respondent had no way of knowing that the follow-up series of questions was coming. Also, since previous evidence had suggested that the first couple of tasks tend to produce different results than later tasks, the follow-up questions were delayed until the fourth and tenth choices.

Since our attributes had varying numbers of levels, the direct comparison between attributes within a treatment cell is probably not meaningful. However, the comparison *within* an attribute *between* treatments is meaningful. The following table represents the average proven recall of selected level by attribute for each treatment cell:

Proven Attribute Level Recall
Comparison by Number of Alternatives

		Number of Alternatives	
Attribute	Levels	2	7
1	6	25	40
2	5	27	40
3	5	15	31
4	5	18	34
5	3	30	45
6	3	23	23
7	4	29	23
Avg.		23.9	35.1
Ratio (relative to pairs)			1.47

It is somewhat surprising that the tasks with seven concepts consistently perform better on this measure than the tasks with only two concepts. Evaluating choice tasks with only two concepts and recalling the seven appropriate levels (one for each attribute) as compared to the seven unselected levels would seem more easily accomplished than recalling the seven selected levels out of the 42 levels not chosen.

Again, to this puzzle we offer a conjecture. The attributes studied ranged between three and six levels. Note that in the scenario with two concepts per task, it is impossible for a respondent to see all possible levels of an attribute. It is perfectly likely to see all levels when seven concepts are shown. We hypothesize that respondents in the pairs are making relative decisions between the levels shown while respondents in the sevens are making more absolute determinations of the entire range of the consideration set. Since each respondent evaluated many sets of alternatives, the aggregate parameter estimates from MNL were able to provide appropriate measures of intra-attribute distances even in the pairs. However, the differences in the proven recall do cause us to question the advisability of using pairs when the number of levels per attribute is large or relatively unfamiliar to respondents.

We can evaluate the average proven recall for each treatment based on the number of levels in an attribute. Here, multiple attributes with the same number of levels have been averaged.

Proven Attribute Level Recall
Comparison by Number of Alternatives

	Number of Alternatives		
Levels	2	7	Ratio
3	27	34	1.26
4	29	33	1.14
5	20	35	1.75
6	25	40	1.60
AVG.	27.75	35.50	
MAD	3.00	2.25	

This table supports our previous conjecture that the effect is related to the number of attribute levels being studied. This table is also interesting in that we see the seven treatment produce relatively more stable results (based on the Mean Absolute Deviation MAD) while the pairs are less stable.

It is unclear what generalizations, if any, can be made from this one dataset. With many possible explanations and few observations, the effect of any one source of variation cannot be identified with much accuracy. However, evidence supporting attribute importance simplification would be far more comforting than attribute position simplification. At least this one dataset suggests caution with the use of pairs in the presence of a large number of attribute levels.

None Usage

The third cost criterion relates to the use of the default or none option. It is generally accepted that including the default option is worthwhile and improves the validity of the parameter estimates (Olsen and Swait), even if decreasing efficiency in randomized designs. The reduction in efficiency should be equivalent to a comparable reduction in sample size or number of tasks. For our purposes, however, the question we are to answer is do respondents use the None option similarly in choice tasks with varying numbers of alternatives.

In randomized choice designs, the best concept in a set of four (quads) is likely better than the best product in a set of two (pairs). As such, the economic hypothesis would suggest lower use of the default.

However, if choices were made purely at random, we would also expect quads to have lower default use than pairs.

At the same time, it could be argued that since quads require more reading and processing, they form a more difficult respondent task, and might increase the use of the default alternative if the decision avoidance hypothesis is believed, especially later in the exercise.

The results from the first two studies' (pooled) default usage is shown below.

	When Pairs Come First	When Quads Come First
Pairs	0.21	0.30
Quads	0.11	0.12

It is interesting that the quads produce the same default usage regardless of position. The pairs, however, do behave as expected with pairs coming later in the task having higher default usage.

Exploring the default usage in the third study, we can actually look at the differences in default usage between early and late tasks.

Use of Default Alternative
By Number of Alternatives and Task

	Number of Alternatives 2	7	Ratio 2:7
Task 1-4	13.7	9.8	1.40
Task 5-8	19.4	14.7	1.32
Task 9-12	21.9	19.0	1.15
Task 13-16	9.4	11.4	0.82
Task 17-20	13.8	12.8	1.08
Ratio of none usage			
First 4 to Second 4	1.42	1.50	
First 4 to Third 4	1.60	1.93	

Again, it is not immediately clear what to make of these findings. It does appear, however, that the rate of change in the none usage differs based on the number of alternatives. One plausible hypothesis as to the increase in none usage throughout tasks is that respondents, as they learn the quality of concepts available, are refining their expectations of acceptable. We would expect respondents to be more likely to see exceptional products in the sevens treatment rather than in pairs. But at the same time, since the number of levels of each attribute is less than the number of concepts, respondents can determine the range of possible products, and therefore the best product—no refinement should be expected.

Overall, the cost criteria have raised a number of issues concerning the use of pairs, or at least identified differences between pairs and other treatments. Specifically, it appears that respondents are processing concepts differently in pairs and simplifying their choice heuristics. This is particularly interesting when we consider that evaluating pairs takes about two-thirds of the time of processing seven alternatives and about three-fourths of the time of processing fours. That is, the cost in time of having more alternatives per choice task is relatively small.

It is possible that even with these differences, different treatments in the number of concepts per task produce the same utilities and choice predictions. The similarity or difference of these results is considered in the next section, congruence criteria.

FINDINGS—CONGRUENCE CRITERIA

Predictive Validity—Cross Task Validation

Probably the most straightforward way to evaluate the similarity of two sets of logit utilities is to identify how similarly each predicts choices. That is, we evaluate how well logit utilities developed from pairs, for example, predict respondents' actual choices from quads. In both of the first two datasets, we are severely constrained by number of observations, precluding the ability to conduct this test over random sample replicates. The hit rate of the data set used in logit model estimation will not form an upper bound on predicted hits (Wittink and Johnson), but it is very important to recall that the estimation data set has fully capitalized on chance.

We evaluate predictive validity by reporting a "hits" measure, which is the proportion of times the actual choice is the same as the predicted choice.

It is also somewhat unclear how these hits should be evaluated. Choices are a fallible criterion and contain a fair amount of noise. But more importantly, if one of the treatments is systematically biased, as pairs might appear to be, should competing utilities be deemed poor for not reproducing that bias? We think not.

STUDY ONE

Pairs predicting Pairs	.766	Overfit (model development)
Quads predicting Pairs	.739	
DIFFERENCE	.027	
RATIO	.965	
Quads predicting Quads	.537	Overfit (model development)
Pairs predicting Quads	.515	
DIFFERENCE	.022	
RATIO	.959	

It should be pointed out that if we are just looking at hit rates, pairs seem to do better. However, the level of chance is higher as well. To evaluate the level of chance, let us first exclude the differential rate of none usage from pairs and quads, and then determine what the level of chance would be.

	All Choices	Less Nones		Number of Alternatives	Chance Level
Pairs	100	-25	= 75	÷ 2	= 38
Quads	100	-11	= 89	÷ 4	= 22

The fits in cross-task prediction do nearly as well as the hit rate from which the logit model was developed. Initially, this indicates that either model does a nearly equal job of producing predicted choices. We were intrigued by this similarity (ratios around 0.96). This might indicate that either models are equally good at predicting choices, or it might indicate that the two models are equally good at predicting something, but that they might not be predicting the same thing.

To investigate this point a little further, we determined the proportion of tasks in which independent models from the pairs and quads agreed on which concept should be the most preferred.

Agreement among pairs presented:	.894
Agreement among quads presented:	.813

We were surprised by how much lower these proportions were compared to the ratios of hits. We believe congruence above these ratios is indicative of systematic heterogeneity between the treatments.

The results from the second study resemble the first.

STUDY TWO

Pairs predicting Pairs	.682	Overfit (model development)
Quads predicting Pairs	.646	
DIFFERENCE	.036	
RATIO	.947	
Quads predicting Quads	.443	Overfit (model development)
Pairs predicting Quads	.434	
DIFFERENCE	.009	
RATIO	.980	

In the third study, we are able to conduct this analysis using sample replicates. We repeated the process using three independent random splits of the data. The respondents from each treatment were randomly split into two subsamples. In this way we have predictions between sample subgroups both between treatments and also within treatments. The following results show the average of the three repeated sample splits.

STUDY THREE

Pairs predicting Pairs	.695	(Within replicates only)
Pairs predicting Pairs	.632	(Between replicates only)
Sevens predicting Pairs	.662	
Sevens predicting Sevens	.393	(Within replicates only)
Sevens predicting Sevens	.378	(Between replicates only)
Pairs predicting Sevens	.343	

We were not surprised that sevens predicted sevens better than pairs did. However, we were intrigued to see that sevens predicted pairs even better than pairs predicted pairs. We believe most conjoint researchers would have assumed that cross-task comparisons will do less well than within-task comparisons. Here, however, it would appear that the sevens do no worse predicting pairs than independently estimated utilities from pairs. To test this finding, we use a rather conservative test.

We calculated the number of correct hits for pairs predicting pairs and sevens predicting pairs and compared the difference in the number of correct hits at the level of the respondent. We found that sevens correctly predicted 6.55 of the 12 pairs and pairs correctly predicted 6.28 of the 12 pairs. That difference is statistically significant ($t = 1.98$).

We also calculated the number of correct hits for pairs predicting sevens and sevens predicting sevens and compared the difference in the number of correct hits at the level of the respondent. Not surprisingly, we found that sevens predict sevens better than pairs ($t = 3.69$). Sevens correctly predicted 4.01 of the 12 sevens , while pairs correctly predicted 3.39 of the 12 sevens.

Attribute Importance

The second criterion to determine if the difference in number of alternatives per task is providing different answers is the similarity of attribute importances. Here we evaluate attribute importance as is commonly done in conjoint methods, percentaging the range of an attribute's utilities against the sum of the ranges. Looking at the first study, we see the following trend in importances.

Logit Attribute Importances
By Attribute Importances

Attribute Position	2	4	Ratio 2:4
5	0.051	0.067	0.76
2	0.136	0.179	0.76
3	0.216	0.209	1.03
4	0.266	0.255	1.04
1	0.331	0.289	1.15

In visually inspecting this relationship, we see the hint of a non-linear trend where important attributes are more important in pairs. To investigate this statistically, we predicted the attribute importances from pairs based on the importance from quads as well
as that importance squared. The t-ratios from that run are shown below, and support the existence of a non-linear relationship.

	t-ratios
Attimp (4)	2.32
AttimpSQRD (4)	3.10
Intercept	N.S. (excluded)

We can look at the same relationship in the second study.

Logit Attribute Importances
By Attribute Importances

Attribute Position	2	4	Ratio 2:4
1	0.078	0.039	2.00
2	0.079	0.083	0.95
5	0.195	0.129	1.51
4	0.150	0.189	0.79
6	0.203	0.249	0.82
3	0.294	0.311	0.95

Here the congruence is much less strong between the two data sets, and we don't initially see a non-linear effect. This is further demonstrated by a non-significant square term in the following regression.

	t-ratios
Attimp (4)	3.42
AttimpSQRD (4)	N.S.
Intercept	N.S. (excluded)

Two things are worth pointing out with this analysis.

First, the regression results change drastically by removing one variable from the analysis. The attribute in position 5 is a clear leverage point from the previous table and actually behaved rather sporadically (suffering from extreme reversals) in both treatment cells. Removing that variable and reconducting the analysis we see the following regression results.

	t-ratios
Attimp (4)	N.S.
AttimpSQRD (4)	7.30
Intercept	8.59 (included)

Second, and probably more interestingly, the average number of levels per attribute in the first study was over 4 but was exactly 3.0 in the second study. In fact, this study had two 2-level attributes, two 3-level attributes and two 4-level attributes. It is even more interesting the congruence of attribute importance ratios by the number of levels.

Logit Attribute Importances
By Number of Levels

Attribute Position	Number of Levels	2	4	Ratio 2:4
2	2	0.079	0.083	0.95
3	2	0.294	0.311	0.95
1	3	0.078	0.039	2.00
4	3	0.150	0.189	0.79
6	4	0.203	0.249	0.82
5	4	0.195	0.129	1.51

Since there were no two level attributes in either the first or third study and pairs were always included as one of the treatments, this analysis cannot be replicated in either dataset.

Examining the relationship between attribute importances on the third dataset, we see the same non-linear relationship.

Logit Attribute Importances
By Attribute Importances

Attribute Position	2	4	Ratio 2:4
6	0.053	0.026	2.04
1	0.062	0.051	1.22
4	0.120	0.150	0.80
3	0.140	0.163	0.86
5	0.150	0.151	0.99
7	0.213	0.209	1.02
2	0.261	0.250	1.05

Conducting a regression as above, we see a similar result with a strong non-linear component and a non-zero intercept.

	t-ratios
Attimp (4)	N.S.
AttimpSQRD (4)	3.34
Intercept	3.32 (included)

Relatively more important attributes become even more important in pairs, but don't necessarily go through the origin. We do see huge non-linear effects, which suggest respondent simplification or some mental reweighting in pairs.

Since attribute importances, derived in this way, are not independent measurements, we should also consider other measures to distinguish differences between attribute importances among the treatment cells. The issue of attribute importance in aggregate logit models is confounded by heterogeneity that might not be entirely defeated by respondents' random assignment to treatment cells. This random source of variation is on top of any systematic heterogeneity that might come about as a result of differing numbers of alternatives in choice tasks.

Attribute Reweighting

One solution that captures both sources of variability is the reweighting of individual level utilities. The assumption in conjoint simulators (including ACA's) is that the weights for each of the attributes are unity. It has been shown that non-uniform weights can improve the ability of individual level ratings utilities to predict choices (Huber and Pinnell, 1995; Pinnell, 1994; Huber, Wittink, Johnson, and Miller, 1992).

The first study also included a ratings based conjoint task (ACA) immediately prior to the choice-based task. From this task, we have individual level utilities that can be used to re-estimate a logit model. In this model, though, we are no longer solving for dummy or effects coded levels. Rather, we are solving for a multiplicative weight, developed simultaneously for all respondents, that best predicts their actual choices, effectively maximizing the likelihood of the following term:

$$\frac{e^{(\Sigma \beta X)}}{\Sigma\, e^{(\Sigma \beta X)}}$$

where: X represents the individual level utilities for the shown level from each ACA attribute

β represents the reweight coefficient for each attribute

The cell entries in the following table are indexed within each column so that an attribute that remains equally important in its raw and reweighted form will have a coefficient of 1.00. It should be pointed out that the original ACA study included approximately 18 attributes, only five of which were included in the choice-based task. The attribute importances have been rescaled to sum to 100 for just these five attributes, but were calculated at the individual level using ACA utilities.

Individual-Level Utility Reweighting Coefficients

Ind. Level Attrib Imp.	Number of Alternatives per Task 2	3	4
28.9	1.41	1.23	1.18
22.6	1.34	1.14	0.98
20.3	1.06	1.32	1.18
15.4	1.02	0.83	1.06
12.8	0.17	0.49	0.62

Interestingly, in this analysis, the most important attribute is always weighted up, regardless of the number of alternatives. Conversely, the least important attribute is always weighted down, again regardless of the number of alternatives. Both effects are most severe, however, in the pairs, relative to the quads.

Utilities

After predictive ability and attribute importances, the third congruence criterion is the similarity of the utilities. To compare independent utility estimates, we first must test the congruence of the multiplicative scale factor, as discussed in Swait and Louviere (1993). In keeping with their notation, we conducted the tests shown below. We have arbitrarily scaled logit utilities from the pairs treatment to unity and solved for a relative scale factor for the other treatment. All tests are performed at $\alpha = .05$.

Results of Test of Scale Parameter

$H_0: \beta_2 = \mu_4\beta_4$

	μ_2	μ_4	
STUDY ONE	1.0	1.14	Fail to reject
STUDY TWO	1.0	1.41	Fail to reject

$H_0: \beta_2 = \mu_7\beta_7$

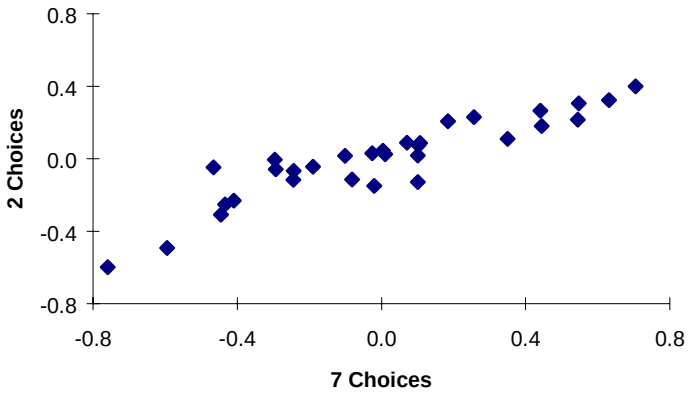
STUDY THREE	REJECT H_0	Utilities are not the same

This table indicates that the study three utilities derived from pairs are not the same as the utilities derived from the sevens. However, in studies one and two we fail to reject the null that they are the same except for a multiplicative scaling constant. The fact that the scale is relatively larger for the quads relative to the pairs indicates that the pairs are "noisier" and that is reflected in logit utilities closer to zero. Based on the differences we have seen between the utilities in the first two studies, however, we question whether the power of the scale test above is sufficient with our limited number of observations.

We will focus our attention on the congruence between utilities in the third study. While the sample sizes are relatively small in the third study as well, we were able to ask far more relevant tasks of each respondent for each treatment, effectively quadrupling the number of tasks evaluated relative to either of the first two studies.

The following graph shows the logit utility estimates derived from the pairs and the sevens. The horizontal axis represents the utilities from the choice tasks with seven alternatives. The vertical axis represents the utilities derived (independently) from the choice tasks with two alternatives per task.

Parameter Estimates
All Levels



Even though we rejected the hypothesis $\beta_2 = \mu_7\beta_7$, we can still make an inference about the relative noise of the two treatments by investigating the scale effect. If we look only at the range of positive utilities (greater than zero) from the sevens, we see a maximum utility of 0.705. The corresponding utility from the pairs is 0.400, loosely implying a scale effect of 1.8. Examining the five most extreme levels, we see similar strong findings.

Sevens	Pairs	Scale
0.705	0.400	1.76
0.631	0.322	1.96
0.548	0.305	1.79
0.545	0.215	2.53
0.444	0.180	2.47

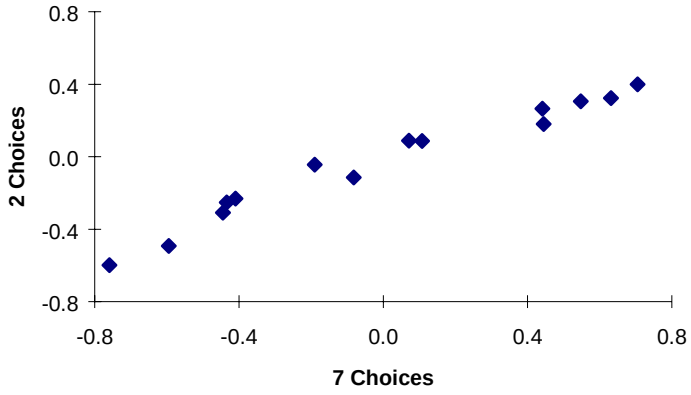
If we fit a least squares line through the positive utilities (based on sevens), the slope is 1.47. In fact, if we fit the line through just the top ten points, the slope is 1.62.

However, if we turn our attention to the negative range of utilities, we see a sharply different effect. The slope of a line fit through the negative utilities is 0.93.

It appears that there is a non-linear trend in the relationship between the two sets of utilities, and it appears most strong in the very extreme negative utility values. We have previously seen that in all three datasets there was a non-linear effect between attribute importances derived from pairs and attribute importances derived from tasks with more alternatives. Therefore, it shouldn't be surprising that the utilities (which are used to create the importances) differ as well. However, there are many ways in which the utilities could differ.

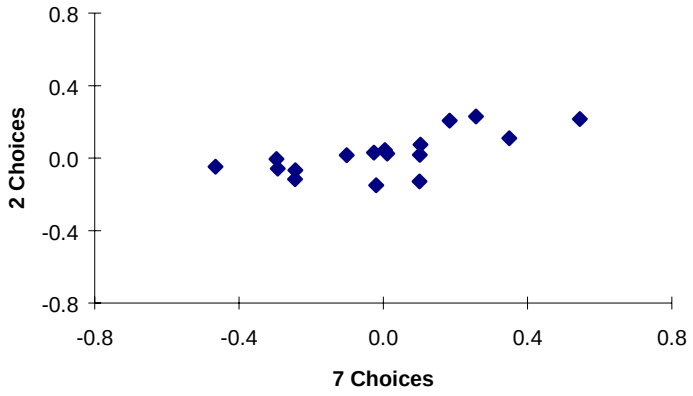
Initially, let us examine a graph, similar to the previous one, but this time only for the most and least preferred level within each attribute.

Parameter Estimates
Extreme Levels



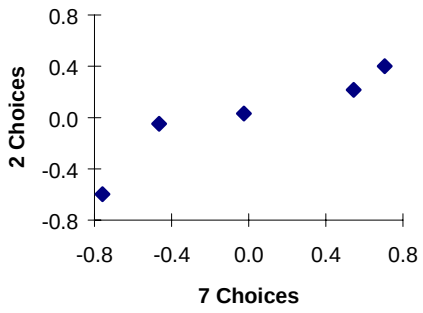
In this instance we see a very similar relationship to that when all levels are considered. This chart, however, presents a cleaner picture of the non-linearity. For completeness, we also show the similar chart for only interior levels.

Parameter Estimates
Interior Levels

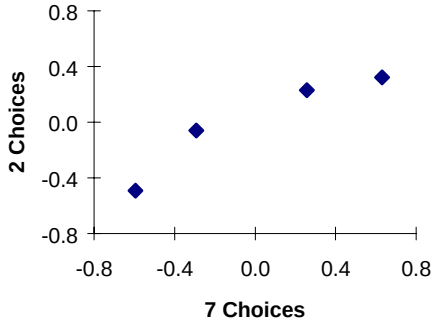


To explore this possible "extremeness" effect in pairs from another perspective, let us consider comparisons among specific attributes. Shown below are similar graphs for a five-level attribute and a four-level attribute from the third study.

Parameter Estimates
All Levels - One Attribute

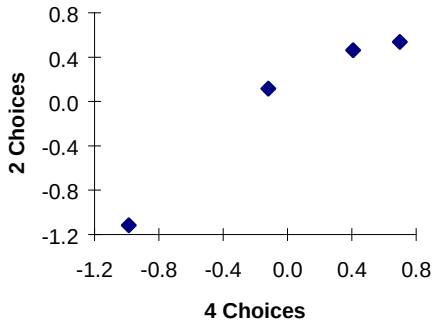


Parameter Estimates
All Levels - One Attribute

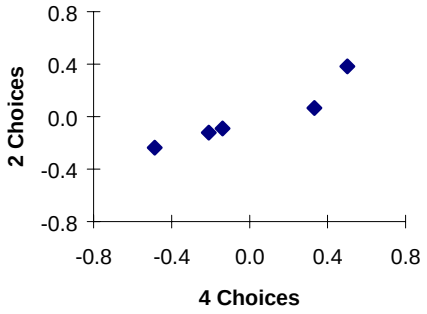


These confirm our expectation of loss aversion in the pairs, and one provides some support for a positivity effect. Exploring these relationships in attributes from the first two studies, we find similar support.

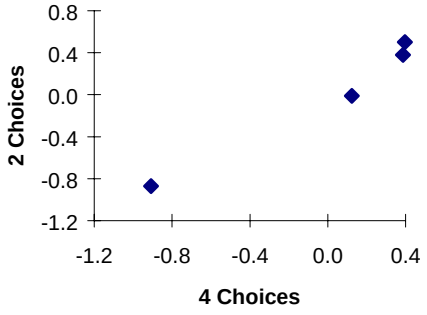
Parameter Estimates
All Levels - One Attribute



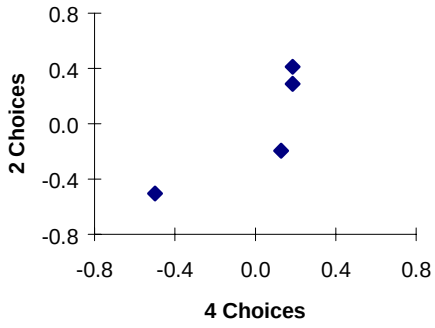
Parameter Estimates
All Levels - One Attribute



Parameter Estimates
All Levels - One Attribute



Parameter Estimates
All Levels - One Attribute



We can only conclude that respondents process pairs differently, specifically relating to the extremes of an attribute.

FINDINGS—EFFICIENCY CRITERIA

So far, we have discussed the benefits in terms of efficiency of a larger number of alternatives per task, but have not substantiated that position. Using a randomized choice design will involve a slight decrease in efficiency relative to a true orthogonal, level-balanced design. If we would have had a level balanced and orthogonal design for both the pairs and the sevens, we would have seen a ratio of 6.0 in relative D-efficiency (sevens are six times better). As in Huber and Zwerina, we consider relative D-efficiency of zero-centered (utility neutral) attributes in a design matrix. We find that our sevens provide 540 percent of the pairs' D-efficiency, calculated this way.

More concretely, we can evaluate the difference between the utility estimates and their standard errors between the two models. We calculate a simple average of the absolute values of all model parameters from the pairs and the sevens. We also calculate an average of the standard errors of the estimates from each model.

	Avg. Absolute. Parameter	Avg. Standard Error	Inferred Avg. Absolute t-ratio
Pairs	0.166	0.0875	1.89
Sevens	0.263	0.0658	4.00

We see that in the sevens we have a larger parameter by an average of 59% and a smaller standard error by an average of 25%. Therefore, if we were to calculate an inferred t-ratio across these averages we see that we have more than twice the signal-to-noise ratio in the sevens as in the pairs.

CONCLUSIONS

We reported our findings according to three sets of criteria: cost, congruence, and efficiency.

The cost of an approach can be defined by the following:

- the time it takes a respondent to evaluate the alternatives presented and provide a choice,
- systematic measurement error (specifically, respondent simplification),
- and use of the none alternative.

Although it takes about 33% more time for respondents to evaluate four alternatives than two and about 60% more time to evaluate seven alternatives than two, we collect far more information from fours and sevens.

Pairs show problems with systematic measurement error, specifically respondent simplification. When asked whether they recall the level of an attribute of the concept chosen, respondents correctly identify the level chosen *more* often with sevens than they do with pairs. Independent measures of attribute importance have a stronger relationship with the proven recall for sevens than pairs. Attribute position seems to have an effect on recall for pairs.

In studies with two and four alternatives, respondents use the none alternative less often with fours than with pairs. The highest incidence of none usage is with respondents who see pairs after fours (40%) possibly indicating that respondents are unable to find an acceptable concept among pairs after seeing fours. When respondents see pairs and sevens, the none alternative is used approximately 30% more often in pairs than in sev-

ens. Again, those respondents who see sevens first and then pairs use the none alternative more often in pairs than they do in sevens.

The congruence criteria include:

- predictive validity,
- similarity of attribute importances,
- and similarity of utilities.

Predictive validity is evaluated by calculating the percent of choices correctly predicted using utilities independently estimated from tasks with differing numbers of alternatives. In two of the three studies, sample size limitations prevented us from looking at independently developed utilities. In the third study, however, we were able to conduct the analysis among sample replicates to develop independent estimates. As we expected, sevens predicted sevens better than pairs. We were surprised to find that sevens predicted pairs better than pairs. This result is confirmed at the respondent level—comparing the mean number of hits predicted by pairs and sevens.

The relationship between attribute importances derived from pairs and those derived from four or seven alternatives is non-linear. That is, important attributes are more important in pairs.

We found that the utilities estimated in the first two studies for pairs and quads differ only by a scale parameter. The scale parameter (applied to the utilities estimated from quads) is larger than unity in both studies indicating more noise among the utilities estimated from the pairs. With limited sample sizes in the first two studies, we question the power of the test for scale differences. In the third study we found that utilities based on the pairs and sevens are not the same. We can infer a scale value for sevens of approximately 1.5 for most of the utilities estimated. However, among the least preferred levels estimated in the pairs, we observe a non-linear relationship indicating a loss aversion effect. We confirm this effect in specific attributes from all three studies. There is also some evidence of a positivity effect within specific attributes.

We evaluate the efficiency of the design both theoretically and empirically. Empirically, we find that the sevens have larger parameter, by an average of 59%, and a smaller standard error, by an average of 25%.

RECOMMENDATIONS

While we have examined only three studies and the base sizes in each are somewhat limited, our findings caution against the use of pairs. Our data show that pairs are processed differently, have lower predictive validity, are less stable, and don't save much time relative to larger choice tasks.

So, what is the right number of alternatives per task?

The appropriate number of alternatives should be determined based on the number of levels in the attributes being studied. For example, if your choice study has only four level attributes, then we would recommend including at least four alternatives per choice task. Having exactly four alternatives per task can insure no level overlap in presented concepts. Level overlap in presented concepts will decrease the information provided by each choice.

REFERENCES

Brazell, Jeff and Jordan Louviere (1997), "Helping, Learning, and Fatigue: An Empirical Investigation of Length Effects in Conjoint Choice Studies," Presentation at INFORMS Marketing Science Conference, Berkeley, CA.

Bunch, David, Jordan Louviere and Don Anderson (1983), "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes," Working Paper, Graduate School of Business, University of California, Davis.

Chrzan, Keith (1994), "Three Kinds of Order Effects in Choice-Based Conjoint Analysis," *Marketing Letters*, 5:2, 165-172.

Dhar, Ravi (1992), "Investigation Context and Task Effects on Deciding to Purchase," Ph.D. Dissertation, University of California, Berkeley.

Dhar, Ravi (1997), "Context and Task Effects on Choice Deferral," Marketing Letters, 8:1, 119-130.

Huber, Joel and David Hansen (1986), "Testing the Impact of Dimensional Complexity and Affective Differences of Paired Concepts in Adaptive Conjoint Analysis," in *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 (1995), "Consistent Differences between Experimental Choice and Ratings-Based Trade-offs, Presentation at INFORMS Marketing Science Conference, Sydney, New South Wales, Australia.

Huber, Joel and Jon Pinnell (1994), "The Impact of Set Quality and Decision Difficulty on the Decision To Defer Purchase." Working Paper, Fuqua School of Business, Duke University.

Huber, Joel, Dick Wittink, Richard M. Johnson, and Richard Miller (1992), "Learning Effects in Preference Tasks: Choice-Based versus Standard Conjoint," in *Proceedings of the Sawtooth Software Conference*. Ketchum, ID: Sawtooth Software, 275-282.

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

Huber, Joel, Klaus Zwerina and Jon Pinnell (1995), "Are Utility Balanced Choice Designs Really More Efficient?" Presentation at INFORMS Marketing Science Conference, Sydney, New South Wales, Australia.

Johnson, Richard M. (1981), "Problems in Applying Conjoint Analysis," Presentation at Conference on Analytic Approaches to Product and Marketing Planning, Vanderbilt University.

Johnson, Richard M. (1987), "Adaptive Conjoint Analysis," in *Proceedings of the Sawtooth Software Conference*. Ketchum, ID: Sawtooth Software, 253-65.

Johnson, Richard M. (1989), "Assessing the Validity of Conjoint Analysis," in *Proceedings of the Sawtooth Software Conference*. Ketchum, ID: Sawtooth Software, 273-280.

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.

Kuhfeld, Warren, Randall Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, Vol 21 (November), 545-557.

Louviere, Jordan (1993), "Conjoint Analysis for Large Number of Attributes," Presented at AMA Advanced Research Techniques Forum.

Louviere, Jordan and George Woodworth (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data." *Journal of Marketing Research*, Vol. 20 (November), 350-67.

Olsen, Douglas and Joffre Swait (1994), "The Importance of Nothing," Working Paper.

Pinnell, Jon (1994), "Multistage Conjoint Methods to Measure Price Sensitivity." Presentation at AMA Advanced Research Techniques Forum, Beaver Creek, CO.

Pinnell, Jon and Joel Huber (1997), "Number of Choice Alternatives in Discrete Choice Experiments," Presentation at INFORMS Marketing Science Conference, Berkeley, CA.

Pinnell, Jon and Joel Huber (1996), "The Effectiveness of Second Choices in Experimental Choice Studies," Presentation at INFORMS Marketing Science Conference, Gainesville, FL.

Schuman, Howard and Stanley Presser (1981), Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Content. New York: Academic Press.

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

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

Tversky, Amos and Eldar Shaffir (1992), "Choice Under Conflict: The Dynamics of Deferred Decisions," *Psychological Science*, 3.6 (November), 358-361.

Walsh, John and David Schmittlein (1997), "Using Choice-Based Conjoint Analysis for Individual Level Predictions: An Empirical Investigation of Choice Set, Choice Task, and Model Estimation Factors," Presentation at INFORMS Marketing Science Conference, Berkeley, CA.

Wittink, Dick and Richard M. Johnson (1992), "Estimating the Agreement Between Choices Among Discrete Objects and Conjoint-Ratings-Based Predictions After Correcting for Attenuation," Working Paper, Cornell University.

EXTENSIONS TO THE ANALYSIS OF CHOICE STUDIES

Thomas L. Pilon [1]
TRAC, Inc.

Most choice studies have made use of "standard" analysis, without attention to differential cross elasticities or unequal competitive effects among brands. This paper will present results from a large client-sponsored data set demonstrating how suitably designed choice studies can also be used to measure differential cross effects among brands. This can lead to more accurate simulators of market behavior, as well as "maps" which graphically portray the extent of competition among brands.

CHOICE VS CONJOINT

Conventional conjoint analysis may lead to biased estimates of price sensitivity. In particular, price sensitivity may be systematically understated (Luery, 1990). Also, most types of conjoint analysis are limited in terms of the number of brands or SKUs that are included in the study. In the beer study that is described below, there were 42 brands included. Also, there were five major pack types for most brands: 6-pack cans, 6-pack bottles, 12-pack cans, 12-pack bottles, and cases of 24 cans. Furthermore, there would have been many more brands and pack types (cases of 24 bottles, 18 packs, 30 packs, etc.) if the budget would have allowed it. In recent years, the marketing research community has discovered that these limitations of conventional conjoint analysis can be circumvented through the use of choice studies. In fact, it is probably appropriate to say that choice based conjoint analysis is the "tool of choice" for pricing studies in the mid-nineties.

It is the purpose of this paper to discuss and show a few extensions to the standard analysis of choice data. The extensions all have to do with the derivation and analysis of cross effects, also known as cross "elasticities". One extension is the derivation of the cross effect matrix itself. Another extension involves rescaling the cross effect matrix so that it can be portrayed in a multidimensional scaling type map. A third extension shows the improvements to standard conjoint simulators that result when cross effects are included in the simulator.

Before describing the data and the choice study, definitions of elasticities and cross-elasticities and a brief review of the marketing literature will be provided.

[1] The author wishes to acknowledge both theoretical and computational contributions from Bryan Orme and Rich Johnson, both of Sawtooth Software.

ELASTICITIES AND CROSS-ELASTICITIES–BACKGROUND

It has long been established in the economic literature that the price actions of one product (or brand) affect the sales (or share) of other products. Econometricians refer to this as "degree of substitutability." This substitutability can be quantified in terms of "elasticities" and "cross-elasticities. A price elasticity (hereinafter referred to as elasticity) can be expressed algebraically as:

$$\frac{\% \Delta S_A}{\% \Delta P_A}$$

where the numerator is the change (Greek letter delta) in sales for brand A, S_A, as a percentage of the original sales volume of brand A and the denominator is the change in price of brand A, P_A, as a percentage of the original price of brand A.

A price cross elasticity (hereinafter referred to as cross elasticity) can be expressed algebraically as:

$$\frac{\% \Delta S_B}{\% \Delta P_A}$$

where the numerator is the change in sales volume for brand B as a percentage of the original sales of brand B and the denominator is the change in price of brand A as a percentage of the original price of brand A.

The use of price elasticities and cross-elasticities in pricing studies is not new as summarized by Rao (1984) in a review of over 200 pricing studies. Recent scholarly articles in this domain include Reibstein and Gatignon (1984) who demonstrated the importance of using elasticities and cross elasticities in product line pricing, and Cooper (1988) who used maps to portray how brands influence competing brands (more on this later). Other recent studies include Krishnamurthi, Raj, and Sivakumar (1995), Cooper, Klapper and Inoue (1996) and Guiltinan and Gunlach (1996), Gupta, Chintagunta, Kaul, and Wittink (1996) and Richard, Allaway, Berkowitz and D'Souza (1996).

Discussions of price elasticity models have appeared in the practitioner literature as well. Luery (1990) describes the evolution of conjoint analysis and the use of cross-effects in simulators. Smallwood (1991), Datoo (1994) and Mohn (1995) all provide easy to understand introductions to the uses of price elasticities and cross elasticities in choice and conjoint models. Wyner, Benedetti and Trapp (1984) provide a very readable paper on price elasticity choice models.

Any review of the applied pricing literature would be remiss without mention of Nagle (1987) and Monroe (1990); both provide excellent comprehensive discussions of applied pricing. Finally, for more rigorous discussions of pricing issues see Devinney (1988) and for an advanced discussion of market response models see Hanssens, Parsons, and Schultz (1990).

THE BEER STUDY

In 1994, a beer manufacturer commissioned a study to learn more about the effects of pricing in their industry. Although the study was conducted in 10 major markets in the U.S., the data presented below are from one market only. The name of that market will remain unmentioned for proprietary reasons. Over 1,400 choice interviews were conducted. On average, each respondent completed slightly less than 20 choice tasks; thus the data consist of nearly 28,000 choice tasks.

The data were collected using Sawtooth Software's Ci3 computerized interviewing program. Due to the large number of brands (42) and other complexities of the study, it was not possible to use Sawtooth Software's Choice Based Conjoint program. Respondents were asked to choose five brands from a list of 42 brands that they would most likely buy or consider buying in a certain situation at a certain type of outlet. Based on these selections, Ci3 then configured a choice screen (see Appendix A–Ci3 Choice Screen 1 for an example). Given a matrix of these brands and pack types and randomly chosen prices within each combination of brand and pack type, respondents were asked which brand they would choose. After all the information about the other brands was removed from the screen, respondents were asked which pack type they would choose (see Appendix A—Ci3 Choice Screen 2). Finally, after choosing the pack type, respondents were asked how many units of that brand/pack type they would purchase at the price that was shown (see Appendix A—Ci3 Choice Screen 3). The combination of the three screens described above represented one task.

ELASTICITIES AND CROSS-ELASTICITIES

To calculate the elasticities, the log of the number of units of a brand chosen (adjusted for pack type size) was regressed against the log of the brand's price. The double-log transformation is commonly employed by econometricians (Johnston, 1984) because it corresponds to the assumption of a constant elasticity between brand and price and the simple application of linear methods to the logarithms of the variables directly produces an estimate of that elasticity (the βs are the elasticities). The cross elasticities were calculated in the same manner: the log of the number of units of a brand chosen was regressed against the log of the competitive brand's price. Of course, only tasks which included both brands could be used in this case.

The elasticities and cross-elasticities were each calculated independently in a series of bivariate regressions. Since the Ci3 was randomized, and therefore essentially orthogonal, the effects could be investigated separately without loss of information. Also, trying to estimate all the effects in one model would have required too many coefficients to estimate reliably at one time.

Running the 144 (12 elasticity and 132 cross-elasticity) regressions yield the cross elasticity matrix shown in Table 1. The diagonal elements are individual brand elasticities. Elasticities are usually negative because price and choice volume typically move in opposite directions within a given brand (a brand decreases its price and its choice volume increases). For example, the elasticity of Miller Lite (MIL) is -2.10, thus, if Miller Lite increases its price by 1 percent, its choice volume would drop by 2.10 percent (assuming all other things were held constant, which of course, they never are). The off-diagonal elements are cross elasticities. Cross-elasticities are usually positive because the price of one brand and the choice volume of competing brands typically move in the same direction (a brand decreases its price and the choice volume of competing brands decreases). For example, the 1.18 in the Budweiser (Bud) row and the Coors column indicates that if the price of Bud was increased by 1 percent, Coors choice volume would increase by 1.18 percent.

Table 1–Cross Elasticity Matrix

	Bud	**BudL**	**Mich**	**MichL**	**MGD**	**MGDL**	**MlL**	**Coors**	**CoorL**	**Hein**	**Molsn**	**Sam**
Bud	**-3.80**	1.70	1.87	1.11	1.68	0.50	0.70	1.18	0.61	0.90	0.60	0.60
BudL	1.50	**-3.34**	1.30	1.60	1.10	0.77	0.97	0.80	1.00	0.90	0.40	0.51
Mich	1.10	0.68	**-3.18**	1.00	0.92	0.29	0.38	0.74	0.35	0.12	0.23	0.56
MichL	0.70	1.14	0.92	**-3.53**	0.46	0.52	0.63	0.33	0.72	0.08	0.19	0.51
MGD	0.90	0.70	0.97	0.50	**-3.42**	0.96	0.45	0.73	0.42	0.09	0.22	0.64
MGDL	0.20	0.37	0.19	0.33	0.64	**-2.20**	0.32	0.18	0.29	0.04	0.14	0.31
MlL	0.60	1.26	0.66	1.32	0.90	0.92	**-2.10**	0.54	1.07	0.12	0.17	0.32
Coors	0.70	0.44	0.81	0.37	0.83	0.27	0.33	**-2.30**	0.54	0.06	0.18	0.45
CoorL	0.40	1.27	0.55	1.30	0.75	0.71	1.02	0.80	**-2.20**	0.14	0.13	0.51
Hein	0.60	0.62	0.85	0.86	0.75	0.41	0.46	0.50	0.52	**-0.42**	0.37	1.18
Molsn	0.40	0.34	0.44	0.40	0.53	0.38	0.18	0.27	0.18	0.09	**-0.96**	0.77
Sam	0.29	0.22	0.34	0.35	0.40	0.31	0.11	0.26	0.19	0.09	0.23	**-2.85**

It is not surprising to see the relatively large elasticities in the Bud and Bud Light rows because they are by far the two largest-selling brands (see Table 3 below). Suppose that there are just two brands in the market place. Brand L sells 100,000 units a year and Brand S sells 50,000 units a year. If Brand L lowers its price and increases its sales by 1 percent (gains 1,000 units), and the smaller brand does not react (and the market is in some type of equilibrium), then Brand S will lose a 1,000 units in sales—but 1,000 units is 2 percent of Brand S's sales.

PRODUCT MAPS

While cross-elasticity matrices are informative, they do not offer ready access to the big picture of the relative price sensitivities among brands. Careful study of the cross-elasticity matrix, along with a subjective adjustment for the size effect discussed above, reveals that the "most similar" brands have the largest cross-elasticities. Both the desire to see the "big picture" and the difficulty of subjectively adjusting elasticities inspired the development of an algorithm that would remove brand size effects and rescale the matrix so that it was amenable to the arsenal of mapping techniques that market researchers have developed.

The effect of brand size can be removed from the elasticities so that the "elasticities" are not percentage changes, but instead are proportional to absolute changes. Rescaling the rows of the matrix to be of the same size removes the effect of the size of the "active" brand (brand making the price changes). Rescaling the columns of the matrix to be of the same size removes the effect of the size of the "passive" brand (brands affected by the price change). Iteratively rescaling the rows and columns until they converge removes both effects and makes the final result independent of whether the rows or columns were rescaled initially. After the process converges, the matrix is much more symmetric, but not exactly so. Because most mapping programs require a symmetric matrix as input, the elements on each side of the diagonal were averaged[2].

Finally, the matrix was "standardized" by dividing each element by the square root of the product of the diagonal elements. This "standardization" removes the arbitrary scaling so that products have unit similarity with themselves. The final similarities matrix is presented in Table 2. See Appendix B for the details of the step by step process of getting from the cross-elasticity matrix to the similarities matrix.

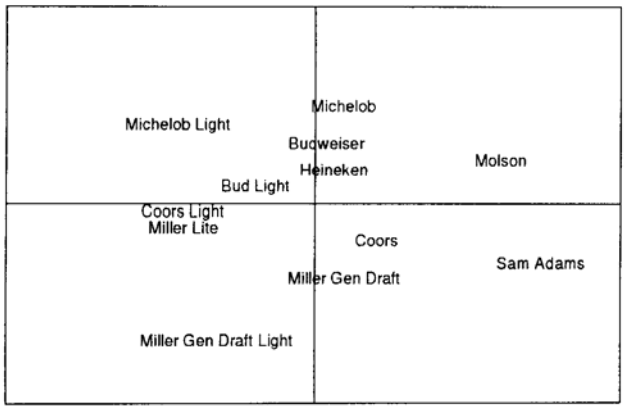
[2] Using a special case of three mode factor analysis on the original asymmetric cross elasticity matrix, Cooper (1988) was able to derive two sets of brand positions, one which portrays how brands exert influence over the competition and the other which portrays how brands are influenced by others.

Table 2—Similarities Matrix

	Bud	BudL	Mich	MichL	MGD	MGDL	MlL	Coors	CoorL	Hein	Molsn	Sam
Bud	**-1.00**	0.45	0.42	0.25	0.34	0.11	0.23	0.31	0.17	0.68	0.26	0.13
BudL	0.45	**-1.00**	0.29	0.40	0.26	0.20	0.42	0.21	0.42	0.76	0.21	0.11
Mich	0.42	0.29	**-1.00**	0.29	0.29	0.09	0.19	0.29	0.17	0.28	0.18	0.13
MichL	0.25	0.40	0.29	**-1.00**	0.14	0.15	0.34	0.12	0.35	0.23	0.15	0.13
MGD	0.34	0.26	0.29	0.14	**-1.00**	0.29	0.24	0.28	0.20	0.23	0.19	0.16
MGDL	0.11	0.20	0.09	0.15	0.29	**-1.00**	0.25	0.10	0.21	0.14	0.16	0.12
MlL	0.23	0.42	0.19	0.34	0.24	0.25	**-1.00**	0.19	0.49	0.26	0.12	0.08
Coors	0.31	0.21	0.29	0.12	0.28	0.10	0.19	**-1.00**	0.29	0.19	0.15	0.13
CoorL	0.17	0.42	0.17	0.35	0.20	0.21	0.49	0.29	**-1.00**	0.29	0.11	0.12
Hein	0.68	0.76	0.28	0.23	0.23	0.14	0.26	0.19	0.29	**-1.00**	0.30	0.30
Molsn	0.26	0.21	0.18	0.15	0.19	0.16	0.12	0.15	0.11	0.30	**-1.00**	0.26
Sam	0.13	0.11	0.15	0.13	0.16	0.12	0.08	0.13	0.12	0.30	0.26	**-1.00**

Again, the goal of the above exercise was to rescale the cross elasticity matrix into a matrix from which we could produce a map that shows the relative degree of price sensitivity among brands. The similarities matrix in Table 2 can now be subjected to various types of metric and non-metric multi-dimensional scaling techniques. The map in Figure 1 was produced using the Systat Multidimensional Scaling (MDS) routine which is non-metric. A Kruskal loss function that produces results comparable to the well known Bell Labs' KYST was used. Seventy-six percent of the variance is explained by the two dimensions. Examination of the Shepard Diagram, which is a scatterplot of distances between points in the MDS plot versus the similarities that were input, indicated that there was a good fit.

Figure 1 - MDS Map of Similarities Data in Table 2



The Young loss function, which is designed to produce results comparable to ALSCAL (available in SPSS Professional Statistics 7.5), produced very similar results. With the assumption that there is a linear relationship between distances and similarities, a principal components or factor analysis method could have been used. Note, for methods that require a full matrix (such as principal components), the signs of the diagonals should be reversed. Usually, but not necessarily, multidimensional scaling can fit an appropriate model in fewer dimensions than can principal components, so MDS was chosen. See Pilon (1989, 1992) for applied comparisons of results obtained from alternative perceptual mapping techniques or see Green, Carmone, and Smith (1989) for a much more detailed discussion. Also, the chapter on perceptual mapping in Hair, Anderson, Tatham and Black (1995) contains an excellent readable discussion.

The map has face validity in that all of the light beers are together (see Figure 2). Also, Figure 3 shows that all the pairs of companion brands (the regular and light beers of the same brand) are relatively close to together. Miller Lite is not really a companion brand to Miller Genuine Draft and Miller Genuine Draft Light but it is in the "Miller" neighborhood. In general, the horizontal axis can be interpreted as a "lightness/ heaviness" dimension while the vertical axis can be interpreted as a "manufacturer" dimension.

Figure 2 - MDS Map with Light Beers in Shaded Box

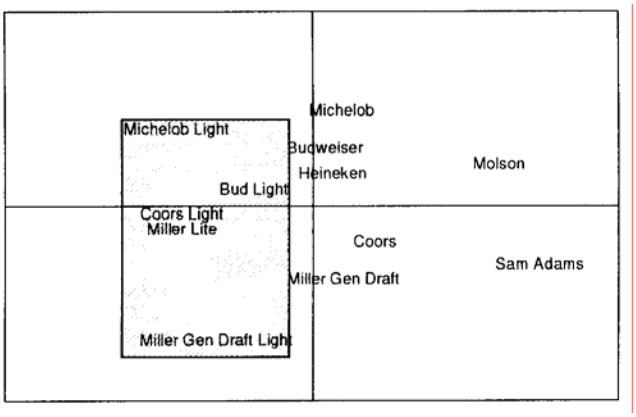
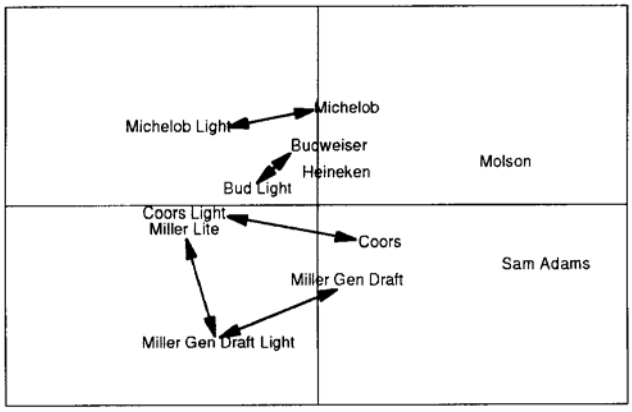


Figure 3 - MCD Map with Companion Brands Connected



If these maps can be believed, they have very important pricing ramifications. Seemingly, brands may compete with their companion brands as much as their competitive brands. When a brand decreases its price, it seems that it would take as many customers away from the companion brand as it would from its competitive brands.

CROSS ELASTIC SIMULATORS

With these observations in mind, a simulator was built that included these cross elastic effects. Most conjoint simulators, especially those derived from main effects only conjoint models, allocate the share given up by a brand that raises its price proportionately (to share) across all the other brands in the simulator. In many (if not most) cases, this proportionate allocation is not an accurate representation of how markets actually respond.

One of the major difficulties in conjoint analysis is overcoming the limitations of the independence of irrelevant alternatives (IIA) problem. The best explanation that I have seen of this problem is:

> The basic idea of IIA is that the ratio of any two products' shares should be independent of all other products. This sounds like a good thing, and at first, IIA was regarded as a beneficial property.
>
> However, another way to say the same thing is that an improved product gains share from all other products in proportion to their shares; and when a product loses share, it loses to others in proportion to their shares. Stated that way, it is easy to see that IIA implies an unrealistically simple model. In the real world, products compete unequally with one another, and when an existing product is improved, it usually gains most from a subset of products with which it competes most directly.
>
> Imagine a transportation market with two products, cars and red busses, each having a market share of 50%. Suppose we add a second bus, colored blue. An IIA simulator would predict that the blue bus would take share equally from the car and red bus, so that the total bus share would become 67%. But it's clearly more reasonable to expect that the blue bus would take share mostly from the red bus, and that total bus share would remain close to 50%. Indeed, the IIA problem is sometimes referred to as the "red bus, blue bus problem." (Johnson, 1997)

By incorporating the cross elasticities from Table 1 above into a simulator, the IIA problem is greatly alleviated. In the Cross Elasticity Simulator, the coefficients from each column of the Bud row of Table 1 were applied independently. Specifically, the volume of Bud was reduced by 3.8%, the volume of Bud Light was increased by 1.7%, ..., and the volume of Sam Adams was increased by 0.6%. As a final step, the resulting shares were rescaled to sum to 100. In the Standard IIA Simulator, only the elasticity of Bud was applied. The share that Bud gave up was allocated proportionately across the other brands' shares. Again, the resulting shares were rescaled to sum to 100. Table 3 shows how the results differ from a simulator that includes the main effects (elasticities) only. Note that the magnitude of Bud's percent loss is much less with the Standard IIA Simulator than with the Cross Elasticity Simulator. Also, note the variation in the % Gain with the Cross Elasticity Simulator as opposed to the constant % Gain with the Standard IIA

Simulator. While "truth" is not known, the Cross Elasticity Simulator results seem to coincide more with what one would expect. If one believes the cross elasticity matrix above, then one would believe the Cross Elasticity Simulator's results more so than the Standard IIA Simulator's results.

Table 3—Simulation Results from a 1% increase in Bud's Price

	Base Case Market Share	Standard IIA Simulator Share	% Gain/ Loss	Cross Elasticity Simulator Share	% Gain/ Loss
Bud	31.80%	30.96%	-2.69%	30.73%	-3.46%
BudLgt	16.60%	16.80%	1.21%	16.96%	2.13%
Michelob	1.36%	1.37%	1.21%	1.39%	2.30%
MichelobL	1.72%	1.74%	1.21%	1.75%	1.56%
MillerGD	8.06%	8.16%	1.21%	8.23%	2.11%
MillerGDL	3.80%	3.84%	1.21%	3.83%	0.97%
MillerLite	17.63%	17.85%	1.21%	17.84%	1.16%
Coors	2.63%	2.66%	1.21%	2.67%	1.63%
CoorsLgt	14.90%	15.09%	1.21%	15.07%	1.07%
Heineken	0.85%	0.86%	1.21%	0.87%	1.36%
Molson	0.41%	0.41%	1.21%	0.41%	1.06%
SamAdm	0.24%	0.24%	1.21%	0.24%	1.06%
	100.0%	100.0%		100.0%	

DISCUSSION

Choice studies like the one described above have several advantages over conjoint studies. Specifically, they allow for unique price levels and price effects (utilities) for each brand. As was shown above, they also allow for the calculation of cross-effect matrices. These cross effects can be graphically portrayed in various types of perceptual maps. When cross effects are incorporated into simulators, they yield more believable results than traditional simulators and they also alleviate the IIA problem that has plagued conjoint simulators since their inception.

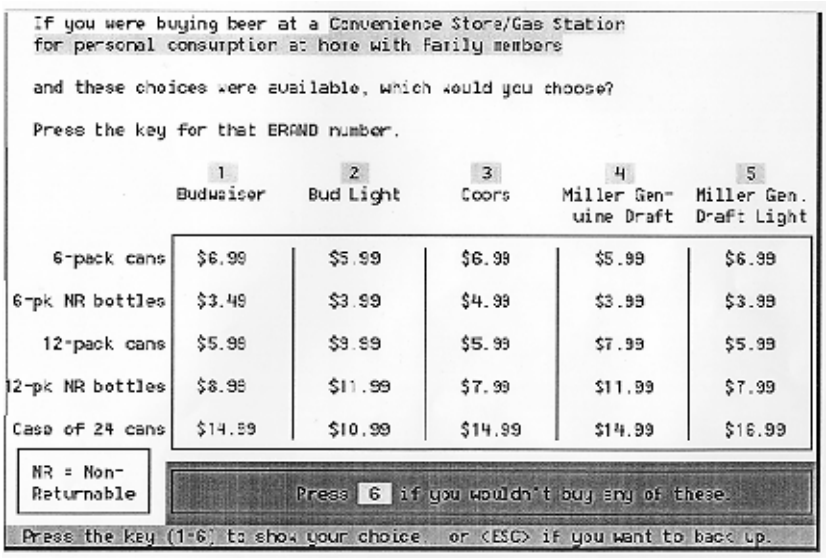
However, one problem with this type of study is the focus is clearly on price. Although we tried to disguise the price focus of the study by varying the situation and outlet type, it did not take respondents long to realize that we were playing pricing games. We may have made them overly sensitive to price. It would have been better to have a few other attributes and fewer brands and pack types and price points for each brand/pack type.

Another problem with this type of study is that various types of statistical anomalies may occur. Cross-effects can be very small or negative requiring smoothing, and repercentaging results of simulators so that they add to 100 can sometimes create reversals.

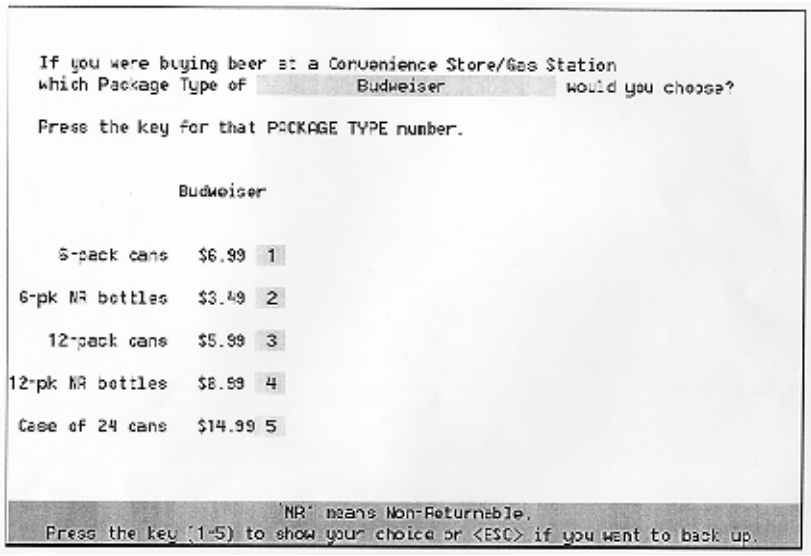
Finally, I think it would be very useful to find a simpler way than Cooper (1988) to show both how a brand is affected by other brands and how a brand affects others brands, rather than simply rescaling to remove the size factor and then averaging the two effects as was done above.

Other methods that are commonly used for pricing studies have problems, as well. Discrete Choice Models add other attributes and varying them across scenarios deflects the undue emphasis on price and decreases response bias, but brings back the IIA problem. Mother logit models do allow for cross-effects other than strictly proportionate share draws, but are complex and are not available to most researchers.

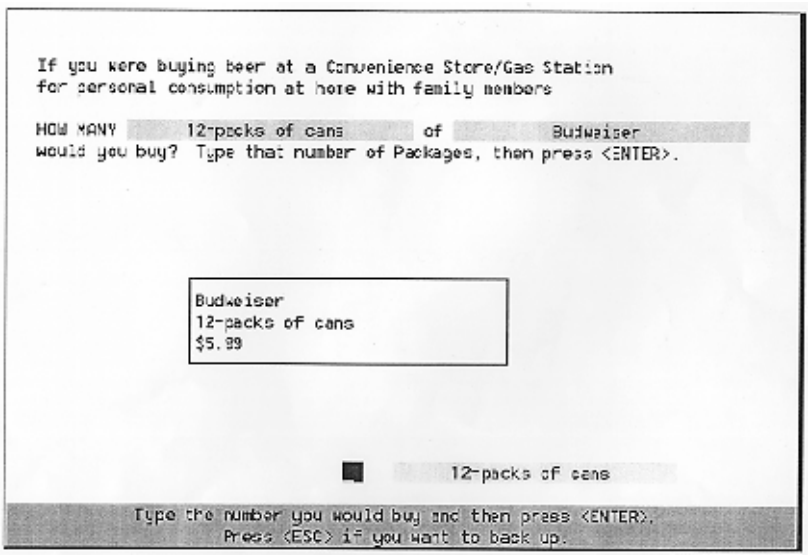
Appendix A - Ci3 Choice Screen 1



Appendix A - Ci3 Choice Screen 2



Appendix A - Ci3 Choice Screen 3



Appendix B - Iterative Rescaling of Cross Elasticity Matrix to Similarities Matrix

Original Cross Elasticity Matrix:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-3.80	1.70	1.87	1.11	1.68	0.50	0.70	1.18	0.61	0.90	0.60	0.60	11.45
BudL	1.50	-3.34	1.30	1.60	1.10	0.77	0.97	0.80	1.00	0.90	0.40	0.51	10.85
Mich	1.10	0.68	-3.18	1.00	0.92	0.29	0.38	0.74	0.35	0.12	0.23	0.56	6.36
MichL	0.70	1.14	0.92	-3.53	0.46	0.52	0.63	0.33	0.72	0.08	0.19	0.51	6.19
MGD	0.90	0.70	0.97	0.50	-3.42	0.96	0.45	0.73	0.42	0.09	0.22	0.64	6.58
MGDL	0.20	0.37	0.19	0.33	0.64	-2.20	0.32	0.18	0.29	0.04	0.14	0.31	3.01
MIL	0.60	1.26	0.66	1.32	0.90	0.92	-2.10	0.54	1.07	0.12	0.17	0.32	7.89
Coors	0.70	0.44	0.81	0.37	0.83	0.27	0.33	-2.30	0.54	0.06	0.18	0.45	4.97
CoorsL	0.40	1.27	0.55	1.30	0.75	0.71	1.02	0.80	-2.20	0.14	0.13	0.51	7.57
Hein	0.60	0.62	0.85	0.86	0.75	0.41	0.46	0.50	0.52	-0.42	0.37	1.18	7.12
Molsn	0.40	0.34	0.44	0.40	0.53	0.38	0.18	0.27	0.18	0.09	-0.96	0.77	3.98
Sam	0.29	0.22	0.34	0.35	0.40	0.31	0.11	0.26	0.19	0.09	0.23	-2.85	2.79
Off Diag Col Sum:	7.39	8.74	8.89	9.14	8.95	6.04	5.55	6.32	5.90	2.63	2.86	6.35	

Rescale rows by dividing every element in the row by its row sum from the table above:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-0.33	0.15	0.16	0.10	0.15	0.04	0.06	0.10	0.05	0.08	0.05	0.05	1.00
BudL	0.14	-0.31	0.12	0.15	0.10	0.07	0.09	0.07	0.09	0.08	0.04	0.05	1.00
Mich	0.17	0.11	-0.50	0.16	0.14	0.05	0.06	0.12	0.05	0.02	0.04	0.09	1.00
MichL	0.11	0.18	0.15	-0.57	0.07	0.08	0.10	0.05	0.12	0.01	0.03	0.08	1.00
MGD	0.14	0.11	0.15	0.08	-0.52	0.15	0.07	0.11	0.06	0.01	0.03	0.10	1.00
MGDL	0.07	0.12	0.06	0.11	0.21	-0.73	0.11	0.06	0.10	0.01	0.05	0.10	1.00
MIL	0.08	0.16	0.08	0.17	0.11	0.12	-0.27	0.07	0.14	0.02	0.02	0.04	1.00
Coors	0.14	0.09	0.16	0.07	0.17	0.05	0.07	-0.46	0.11	0.01	0.04	0.09	1.00
CoorsL	0.05	0.17	0.07	0.17	0.10	0.09	0.13	0.11	-0.29	0.02	0.02	0.07	1.00
Hein	0.08	0.09	0.12	0.12	0.11	0.06	0.06	0.07	0.07	-0.06	0.05	0.17	1.00
Molsn	0.10	0.09	0.11	0.10	0.13	0.10	0.04	0.07	0.04	0.02	-0.24	0.19	1.00
Sam	0.10	0.08	0.12	0.13	0.14	0.11	0.04	0.09	0.07	0.03	0.08	-1.02	1.00
Off Diag Col Sum:	1.18	1.34	1.31	1.35	1.44	0.92	0.84	0.92	0.91	0.32	0.45	1.03	

Rescale columns by dividing every element in the column by its column sum from the table above:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-0.28	0.11	0.12	0.07	0.10	0.05	0.07	0.11	0.06	0.25	0.12	0.05	1.11
BudL	0.12	-0.23	0.09	0.11	0.07	0.08	0.11	0.08	0.10	0.26	0.08	0.05	1.14
Mich	0.15	0.08	-0.38	0.12	0.10	0.05	0.07	0.13	0.06	0.06	0.08	0.09	0.98
MichL	0.10	0.14	0.11	-0.42	0.05	0.09	0.12	0.06	0.13	0.04	0.07	0.08	0.99
MGD	0.12	0.08	0.11	0.06	-0.36	0.16	0.08	0.12	0.07	0.04	0.07	0.09	1.01
MGDL	0.06	0.09	0.05	0.08	0.15	-0.80	0.13	0.07	0.11	0.04	0.10	0.10	0.97
MIL	0.06	0.12	0.06	0.12	0.08	0.13	-0.32	0.07	0.15	0.05	0.05	0.04	0.94
Coors	0.12	0.07	0.12	0.06	0.12	0.06	0.08	-0.50	0.12	0.04	0.08	0.09	0.94
CoorsL	0.04	0.13	0.06	0.13	0.07	0.10	0.16	0.11	-0.32	0.06	0.04	0.07	0.96
Hein	0.07	0.06	0.09	0.09	0.07	0.06	0.08	0.08	0.08	-0.18	0.12	0.16	0.96
Molsn	0.08	0.06	0.08	0.07	0.09	0.10	0.05	0.07	0.05	0.07	-0.54	0.19	0.94
Sam	0.09	0.06	0.09	0.09	0.10	0.12	0.05	0.10	0.08	0.10	0.19	-0.99	1.06
Off Diag Col Sum:	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Appendix B - Iterative Rescaling of Cross Elasticity Matrix to Similarities Matrix (cont)

Rescale rows by dividing every element in the row by its row sum from the table above:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-0.25	0.10	0.11	0.06	0.09	0.04	0.07	0.10	0.05	0.22	0.11	0.05	1.00
BudL	0.10	-0.20	0.08	0.10	0.06	0.07	0.09	0.07	0.09	0.23	0.07	0.04	1.00
Mich	0.15	0.08	-0.39	0.12	0.10	0.05	0.07	0.13	0.06	0.06	0.08	0.09	1.00
MichL	0.10	0.14	0.11	-0.43	0.05	0.09	0.12	0.06	0.13	0.04	0.07	0.08	1.00
MGD	0.11	0.08	0.11	0.06	-0.36	0.16	0.08	0.12	0.07	0.04	0.07	0.09	1.00
MGDL	0.06	0.09	0.05	0.08	0.15	-0.82	0.13	0.07	0.11	0.04	0.11	0.10	1.00
MIL	0.07	0.13	0.07	0.13	0.08	0.14	-0.34	0.08	0.16	0.05	0.05	0.04	1.00
Coors	0.13	0.07	0.13	0.06	0.12	0.06	0.08	-0.53	0.13	0.04	0.09	0.09	1.00
CoorsL	0.05	0.13	0.06	0.13	0.07	0.11	0.17	0.12	-0.33	0.06	0.04	0.07	1.00
Hein	0.07	0.07	0.09	0.09	0.08	0.06	0.08	0.08	0.08	-0.19	0.12	0.17	1.00
Molsn	0.09	0.07	0.09	0.08	0.10	0.11	0.06	0.08	0.05	0.08	-0.57	0.20	1.00
Sam	0.08	0.06	0.09	0.09	0.09	0.11	0.05	0.09	0.07	0.09	0.18	-0.94	1.00
Off Diag Col Sum:	1.01	1.02	1.00	1.00	1.01	1.01	1.00	1.00	1.01	0.95	0.99	1.02	

Rescale columns by dividing every element in the column by its column sum from the table above:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-0.25	0.10	0.11	0.06	0.09	0.04	0.07	0.10	0.05	0.23	0.11	0.04	1.01
BudL	0.10	-0.20	0.08	0.10	0.06	0.07	0.09	0.07	0.09	0.24	0.07	0.04	1.01
Mich	0.15	0.08	-0.39	0.12	0.10	0.05	0.07	0.13	0.06	0.06	0.08	0.09	1.00
MichL	0.10	0.14	0.11	-0.43	0.05	0.09	0.12	0.06	0.13	0.04	0.07	0.08	1.00
MGD	0.11	0.08	0.11	0.06	-0.36	0.16	0.08	0.12	0.07	0.04	0.08	0.09	1.00
MGDL	0.06	0.09	0.05	0.08	0.15	-0.82	0.13	0.07	0.11	0.04	0.11	0.10	1.00
MIL	0.07	0.13	0.07	0.13	0.08	0.13	-0.34	0.08	0.16	0.05	0.05	0.04	1.00
Coors	0.13	0.07	0.13	0.06	0.12	0.06	0.08	-0.53	0.13	0.04	0.09	0.09	1.00
CoorsL	0.05	0.13	0.06	0.13	0.07	0.11	0.17	0.12	-0.33	0.06	0.04	0.07	1.00
Hein	0.07	0.07	0.10	0.09	0.08	0.06	0.08	0.08	0.08	-0.20	0.12	0.16	1.00
Molsn	0.09	0.07	0.09	0.08	0.10	0.11	0.06	0.08	0.05	0.08	-0.58	0.20	1.00
Sam	0.08	0.06	0.09	0.09	0.09	0.11	0.05	0.10	0.07	0.10	0.18	-0.91	1.00
Off Diag Col Sum:	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Rescale rows by dividing every element in the row by its row sum from the table above (convergence achieved):

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam	Off Diag Row Sum
Bud	-0.25	0.10	0.11	0.06	0.09	0.04	0.06	0.10	0.05	0.23	0.11	0.04	1.00
BudL	0.10	-0.20	0.08	0.09	0.06	0.07	0.09	0.07	0.09	0.24	0.07	0.04	1.00
Mich	0.15	0.08	-0.39	0.12	0.10	0.05	0.07	0.13	0.06	0.06	0.08	0.09	1.00
MichL	0.10	0.14	0.12	-0.43	0.05	0.09	0.12	0.06	0.13	0.04	0.07	0.08	1.00
MGD	0.11	0.08	0.11	0.06	-0.36	0.16	0.08	0.12	0.07	0.04	0.08	0.09	1.00
MGDL	0.06	0.09	0.05	0.08	0.15	-0.82	0.13	0.07	0.11	0.05	0.11	0.10	1.00
MIL	0.07	0.13	0.07	0.13	0.08	0.13	-0.34	0.08	0.16	0.05	0.05	0.04	1.00
Coors	0.13	0.07	0.13	0.06	0.12	0.06	0.08	-0.54	0.13	0.04	0.09	0.09	1.00
CoorsL	0.05	0.13	0.06	0.13	0.07	0.11	0.17	0.12	-0.33	0.06	0.04	0.07	1.00
Hein	0.07	0.07	0.10	0.09	0.08	0.06	0.08	0.08	0.08	-0.20	0.12	0.16	1.00
Molsn	0.09	0.07	0.09	0.08	0.10	0.11	0.06	0.08	0.05	0.08	-0.58	0.20	1.00
Sam	0.08	0.06	0.09	0.09	0.09	0.11	0.05	0.09	0.07	0.09	0.18	-0.91	1.00
Off Diag Col Sum:	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Appendix B - Iterative Rescaling of Cross Elasticity Matrix to Similarities Matrix (cont)

Average corresponding off-diagonal elements:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam
Bud	-0.25	0.10	0.13	0.08	0.10	0.05	0.07	0.11	0.05	0.15	0.10	0.06
BudL	0.10	-0.20	0.08	0.12	0.07	0.08	0.11	0.07	0.11	0.15	0.07	0.05
Mich	0.13	0.08	-0.39	0.12	0.11	0.05	0.07	0.13	0.06	0.08	0.09	0.08
MichL	0.08	0.12	0.12	-0.43	0.05	0.09	0.13	0.06	0.13	0.07	0.07	0.08
MGD	0.10	0.07	0.11	0.05	-0.36	0.15	0.08	0.12	0.07	0.06	0.09	0.09
MGDL	0.05	0.08	0.05	0.09	0.15	-0.82	0.13	0.06	0.11	0.05	0.11	0.11
MIL	0.07	0.11	0.07	0.13	0.08	0.13	-0.34	0.08	0.16	0.07	0.05	0.04
Coors	0.11	0.07	0.13	0.06	0.12	0.06	0.08	-0.54	0.12	0.06	0.08	0.09
CoorsL	0.05	0.11	0.06	0.13	0.07	0.11	0.16	0.12	-0.33	0.07	0.05	0.07
Hein	0.15	0.15	0.08	0.07	0.06	0.05	0.07	0.06	0.07	-0.20	0.10	0.13
Molsn	0.10	0.07	0.09	0.07	0.09	0.11	0.05	0.08	0.05	0.10	-0.58	0.19
Sam	0.06	0.05	0.09	0.08	0.09	0.11	0.04	0.09	0.07	0.13	0.19	-0.91

Divide each element by the square root of the product of the two corresponding diagonal elements to get Final Similarity Matrix:

	Bud	BudL	Mich	MichL	MGD	MGDL	MIL	Coors	CoorsL	Hein	Molsn	Sam
Bud	-1.00	0.45	0.42	0.25	0.34	0.11	0.23	0.31	0.17	0.68	0.26	0.13
BudL	0.45	-1.00	0.29	0.40	0.26	0.20	0.42	0.21	0.42	0.76	0.21	0.11
Mich	0.42	0.29	-1.00	0.29	0.29	0.09	0.19	0.29	0.17	0.28	0.18	0.13
MichL	0.25	0.40	0.29	-1.00	0.14	0.15	0.34	0.12	0.35	0.23	0.15	0.13
MGD	0.34	0.26	0.29	0.14	-1.00	0.29	0.24	0.28	0.20	0.23	0.19	0.16
MGDL	0.11	0.20	0.09	0.15	0.29	-1.00	0.25	0.10	0.21	0.14	0.16	0.12
MIL	0.23	0.42	0.19	0.34	0.24	0.25	-1.00	0.19	0.49	0.26	0.12	0.08
Coors	0.31	0.21	0.29	0.12	0.28	0.10	0.19	-1.00	0.29	0.19	0.15	0.13
CoorsL	0.17	0.42	0.17	0.35	0.20	0.21	0.49	0.29	-1.00	0.29	0.11	0.12
Hein	0.68	0.76	0.28	0.23	0.23	0.14	0.26	0.19	0.29	-1.00	0.30	0.30
Molsn	0.26	0.21	0.18	0.15	0.19	0.16	0.12	0.15	0.11	0.30	-1.00	0.26
Sam	0.13	0.11	0.15	0.13	0.16	0.12	0.08	0.13	0.12	0.30	0.26	-1.00

REFERENCES

Cooper, Lee G. (1988). "Competitive Maps: The Structure Underlying Asymmetric Cross Elasticities." *Management Science*, vol.34(6) (June), 707-723.

Cooper, Lee G., Akihiro Inoue (1996). "Building Market Structures from Consumer Preferences." *Journal of Marketing Research,* vol.33 (August), 293-306.

Cooper, Lee G., Daniel Klapper, and Akihiro Inoue (1996). "Competitive-Component Analysis: A New Approach to Calibrating Asymmetric Market-Share Models." *Journal of Marketing Research*, vol.33 (May), 224-238.

Datoo, Bashir A. (1994). "Measuring Price Elasticity." *Marketing Research,* vol.6(2) (Spring), 30-34.

Devinney, Timothy M. (1988). *Issues in Pricing.* Lexington: Lexington Books.

Green, Paul E., Frank J. Carmone, and Scott M. Smith (1989). *Multidimensional Scaling: Concept and Applications.* Boston: Allyn & Bacon.

Guiltinan, Joseph P., and Gregory T. Gundlach (1996). "Aggressive and Predatory Pricing: A Framework for Analysis." *Journal of Marketing*, vol.60 (July), 87-102.

Gupta, Sachin, Pradeep Chintagunta, Anil Kaul, and Dick R. Wittink (1996). "Do Household Scanner Data Provide Representative Inferences from Brand Choices: A Comparison with Store Data." *Journal of Marketing Research,* vol.33 (November), 383-398.

Hair, Joseph F. Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1995). *Multivariate Data Analysis.* 4th ed. New Jersey: Prentice-Hall.

Hanssens, Dominique M., Leonard J. Parsons, Randall L. Schultz (1990). *Market Response Models: Econometric and Time Series Analysis.* Boston: Kluwer Academic Publishers.

Johnson, Rich (1997), "*Getting the Most from CBC–Part 2*", Sawtooth Software Technical Paper.

Johnston, J. (1984), *Econometric Methods*, 3rd Edition. New York: McGraw-Hill Publishing Company.

Krishnamurthi, Lakshman, S.P. Raj, and K. Sivakumar (1995). "Unique Inter-Brand Effects of Price on Brand Choice." *Journal of Business Research,* vol.34, 47-56.

Leury, David A. (1990). "How to Predict Market-Share Sensitivity to Price Changes." *Journal of Pricing Management*, Summer, 1990.

Mohn, N. Carroll (1995). "Pricing Research for Decision Making." *Marketing Research,* vol.7(1) (Winter), 11-19.

Monroe, Kent B. (1990). *Pricing: Making Profitable Decisions. 2nd ed.* New York: McGraw-Hill.

Nagle, Thomas T. (1987). *The Strategy & Tactics of Pricing*. New Jersey: Prentice-Hall.

Pilon, Thomas L. (1989). "Discriminant versus Factor Based Perceptual Maps: Practical Considerations." *Sawtooth Software Conference Proceedings,* 166-182.

Pilon, Thomas L. (1992). "A Comparison of Results Obtained from Alternative Perceptual Mapping Techniques." *Sawtooth Software Conference Proceedings,* 163-178.

Rao, Vithala R. (1984), "Pricing Research in Marketing: The State of the Art," *Journal of Business* (January).

Reibstein, David J., and Hubert Gatignon (1984). "Optimal Product Line Pricing: The Influence of Elasticities and Cross-Elasticities." *Journal of Marketing Research,* vol.21 (August), 259-267.

Richard, Michael D., Anthony W. Allaway, David Berkowitz, and Giles D'Souza (1996). "Capturing Competitive, Cannibalistic, and Variety-Seeking Influences on Market Share: An Asymmetric Modeling Approach." *Journal of Applied Business Research,* vol.12(3), 108-119.

Smallwood, Richard (1991), "Using Conjoint Analysis for Price Optimization." *Sawtooth Software Conference Proceedings,* 157-162.

Wyner, Gordon A., Lois H. Benedetti, and Bart M. Trapp, "Measuring the Quantity and Mix of Product Demand." *Journal of Marketing*, Winter 1984, 101-109.

SOFTWARE REFERENCES

Ci3 System
Choice Based Conjoint (CBC)
both by
Sawtooth Software, Inc.
502 South Still Road
Sequim, WA 98382
(360) 681-2300

P-STAT version 2.19
by
P-STAT, Inc.
230 Lambertville-Hopeville Rd.
Hopewell, NJ 08525
(609) 466-9200

SPSS Professional Statistics 7.5
SYSTAT 6.0 & 7.0
both by
SPSS, Inc.
444 North Michigan Avenue
Chicago, IL 60611
(312) 329-2400

COMMENT ON PILON

Bryan Orme
Sawtooth Software, Inc.

This is a nice paper on a couple of accounts. First, Tom presents real data for a familiar category and brands. He is to be congratulated for his initiative to get these data released. It's much more appealing than looking at data labeled as "Brand A," "Brand B," etc. Secondly, this paper scores high on the "Gee Whiz" scale for demonstrating creative ways to make the most of choice data. I expect that after reading this paper many of us will revisit our CBC data sets to pan for cross-elasticity gold.

COMPUTATIONAL NOTE

One can calculate cross-elasticities for standard choice data sets using the log of choice probability (from aggregate counts tables) as the dependent variable, the log of price as the independent variable, and as many observations as price levels measured. For example, counting the percent of times Pepsi was chosen when Coke was offered at five different prices might result in the following aggregate counts table:

Effect of Coke Price Changes on Pepsi Choice Probability

Coke's Price	Pepsi's Choice Probability
$1.40	.23
$1.60	.24
$1.80	.27
$2.00	.29
$2.20	.33

Taking the natural log of each column and regressing Pepsi's choice probability on Coke's price results in a beta (cross-elasticity) of 0.798.

CROSS-ELASTICITY SIMULATOR

I created a cross-elasticity simulator using the approach above for a synthetic CBC data set. The data set had known utilities and normally distributed error with s.d. of 1.0. It included 300 respondents, 20 tasks each, and a 3^3 design with utilities of 1, 0, -1.

While experimenting with the cross-elasticity simulator, I noted that it seemed to work well as long as price changes were modest. Modeling brands (especially with small shares) toward the extremes of the price range proved less stable. With the cross-elasticity simulator as presented here, share is bounded on the upper end at 100%, but is not necessarily bounded by 0 on the downside. For example, the lowest share brand in my example had an elasticity of slightly greater absolute value than -2. Specifying it at the highest price (a 50% increase in price in my model), resulted in a negative share. Including the intercept term from the log-log regressions will control against negative shares. Indeed, there are a number of transformations that could be employed for predicting the preliminary share of a brand before adjustments by cross-elasticities which could relax the assumption of linearity and bound shares by 0 and 100.

SIMULATION AND IIA ISSUES

The extensions to choice analysis that Tom presents result from attempts to overcome weaknesses in the logit model and its IIA property. Maximum utility rule conjoint simulators (First Choice Model in ACA and CVA vernacular) are immune to IIA problems, but require individual-level data. Each respondent contributes a single vote to the product with the highest utility. Since respondents cannot split their vote or cast two votes, maximum utility rule simulators cannot artificially inflate share for like products.

Maximum utility rule simulations on individual-level data can also reveal self- and cross-elasticities, even when the data are based on main-effect designs. If respondents preferring Brand X are more price sensitive in general than individuals who prefer Brand Y, simulations will reveal differences in price sensitivity between brands. Indeed, one could conduct a similar analysis as Tom presents based on market simulations from an ACA or card-sort conjoint. However, I expect that calculating both types of elasticities is more direct, realistic and powerful from choice data.

As a final note, including cross-elasticities in a choice simulator does not solve all of the IIA problems. The initial share estimates in the base case can have IIA biases (but Tom side-stepped this by using a secondary source for base case shares). I'd suggest, space permitting, representing each brand once in each choice task to minimize the opportunity for IIA violations with respect to brands and to maximize the ability to measure cross-elasticities.

RESPONDENTS' BEHAVIOUR IN COMPLEX CHOICE TASKS; A SEGMENTATION-BASED AND INDIVIDUAL APPROACH

Marco Hoogerbrugge
SKIM Analytical

1 INTRODUCTION

Most conjoint packages have the interesting property that they estimate utility values for each respondent separately. There are two occurrences when this property is most valuable.

In the first place when we are conducting a pilot study to test the validity of the survey design. Then we have a limited number of respondents of which we can study the utility values individually.

Apart from pilot studies we do not analyze individual conjoint results. The second reason why the individual utilities are still interesting is that we make an indirect use of them, because we use them as input for multivariate techniques. Most common in this respect is a cluster analysis. This has even become so common that Sawtooth Software developed a special cluster analysis program for clustering ACA utility values, and they called it CCA, Convergent Cluster Analysis.

Today I am going to talk about Choice-Based Conjoint, CBC. In the last years it has received a lot of attention—last year's ART Forum devoted almost half of their time to this technique—and most scientists think favorably of it. However, compared to most other conjoint methods it has lacked one property: it cannot determine individual utility values. This is more or less an automatic result of the Choice-Based methodology: it gathers only 0's and 1's as input and this is from a statistical point of view not very precise. In theory this problem could be solved by offering a respondent some dozens of choice tasks but in practice you will understand that will not work. That is why Sawtooth Software has—so far—not even tried to build in an estimation of individual utility values into their CBC program, it can now only calculate aggregate utility values. At the same time this has also been an advantage, because now you are allowed to offer as few choice tasks per respondent as you want, you can even limit it to one choice task per respondent.

Yet, I was not confident with aggregate data. I will show you two examples why not, one example based on utility values and one example based on simulations.

Sheet 1:

	CBC aggregate utility values
brand 1	-0.25
brand 2	0.5
brand 3	-0.25

Conclusion: brand seems not to be important

Suppose you could derive utility values for clusters:

	cluster 1	cluster 2
brand 1	0.5	-1
brand 2	1.5	-0.5
brand 3	-2	1.5

Conclusion: brand is really important, but there is heterogeneity

Sheet 2

CBC simulation results, based on aggregate data

Base case

	share of choice
brand 1, $ 70, good quality	60%
brand 2, $ 90, excellent quality	40%

Scenario: new product introduction by brand 3

	share of choice
brand 1, $ 70, good quality	36%
brand 2, $ 90, excellent quality	24%
brand 3, $ 60, reasonable quality	40%

The increase in share of choice of the new variation from 0% to 40% is subtracted from the existing products, proportional to their original share of choice. But this is completely against intuition! We would expect that in reality the gain of brand 3 would go mainly at the cost of brand 1. And why would we expect that? Because, based on these figures, we assume that in reality two clusters will exist:

- one cluster which cares primarily about price: they chose brand 1 in the base case and will probably divide between brand 1 and 3 in the scenario;
- and another cluster which cares primarily about quality: they chose brand 2 in the base case and will probably continue to do so.

To summarize, aggregate data are to a certain extent dangerous to use, because you may draw wrong conclusions about the importance of heterogeneous attributes like brand; and also because aggregate data cope sometimes very poorly with cross-elasticities.

So about 1.5 years ago I developed a program myself to calculate the counts per attribute level per respondent, and I run a standard cluster analysis on these counts data. Of course this is not the most elegant procedure because in this way the results are getting dependent on the accidental choice tasks that each respondent has got. But it was the best solution available for me at that time.

At the same time I kept in touch with Sawtooth Software. They recognized my problem and started to work on it. Meanwhile they even have developed two different programs which provide cluster solutions based on CBC data. One program is called Latent Class, which is sold as an add-on to CBC, the other program which is actually somewhat older but has not been commercialized yet, is called K-logit.

Both programs try to find the optimal cluster solution by an iterative process. This process can be described as follows:

Sheet 3

initial solution:
(random) assignment of respondents to clusters
⇓
calculation of utility values per cluster
first iteration: ⇓
re-assignment of respondents to clusters
⇓
re-calculation of utility values per cluster
second iteration: ⇓
re-assignment of respondents to clusters
⇓
re-calculation of utility values per cluster
and so on.

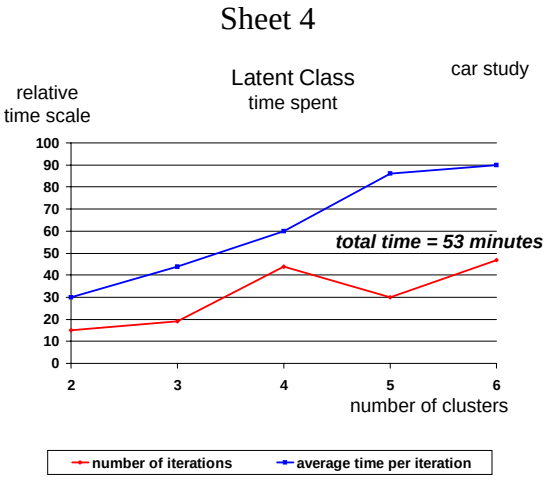
The main difference between the two programs is the fact that K-logit assigns each respondent uniquely to one cluster, while Latent Class calculates for each respondent the probability that he/she belongs to every cluster. In other words, Latent Class provides more detailed information, but at the same time it is no surprise that the calculations take more time. This is then the outline of my talk today, I will discuss the following topics:

- the calculation time of both methods
- the quality / interpretation of the solutions
- the added value of Latent Class: individual probabilities

I'll discuss these topics by showing examples from two studies. This is a small sample size based on which I will draw conclusions, and these conclusions may be overruled in the future when more experience is available with these methods.

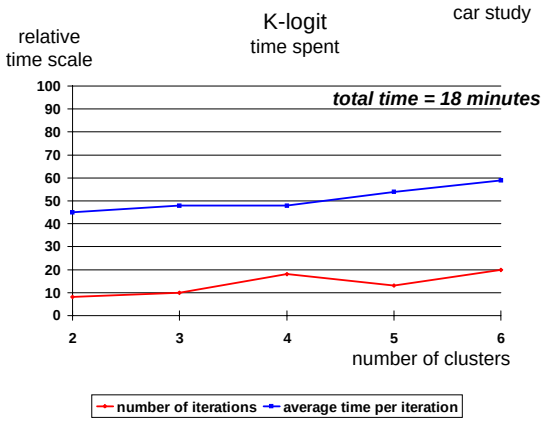
The two different studies are about beers and about cars, and they are about equally complex: they both contain 4 attributes and about 20 attribute levels. The main difference between the two studies is the number of respondents, it is almost 2000 in case of the beer study and some 450 in case of the car survey. In the example I will restrict myself to solutions from 2 to 6 clusters because we have experienced that we have too few choice tasks per respondent to be able to run Latent Class with a higher number of clusters than six. With K-Logit we could still run 10 clusters, so this is already an important difference.

2 CALCULATION TIME



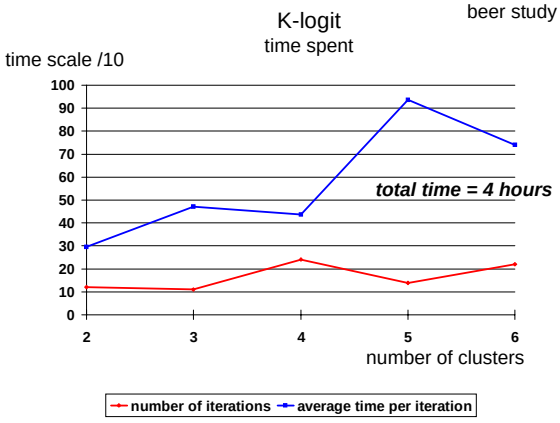
With the car study we see that Latent Class uses a number of iterations which is about proportional to the number of clusters. Also the time for each iteration is about proportional to the number of clusters. So if you multiply, the total time is proportional to the square of the number of clusters.

Sheet 5

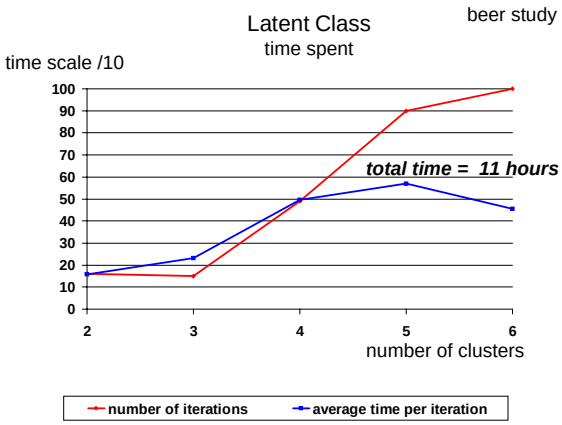


On the next slide the similar chart is shown for K-logit. Both curves, with the number of iterations and the time per iteration, are much flatter. In other words, the calculation time of K-logit is not so much dependent on the number of clusters. With two clusters, K-logit and Latent Class are equally fast, but as the number of clusters increases, K-logit is getting much more efficient than Latent Class. Consequently the total time for K-logit is also much more favorable than for Latent Class: 18 minutes versus 53 minutes, on a Pentium 133.

Sheet 6



With the beer survey, with 4 times as many respondents, something similar happens. Because it takes a lot more time per iteration, I rescaled the time axis in the figure by a factor 10. The number of iterations in the figure is not rescaled, that is just the absolute figure. We see that K-logit keeps requiring a very low number of iterations, regardless of the number of clusters. The time per iteration is much longer than in the previous example (think of the factor 10) and now it is also more clearly increasing.



With Latent Class the number of iterations is accelerating much more (again), it even reaches the system maximum of 100 with the 6 cluster solution. Also the time per iteration increases initially, but after 4 clusters the time per iteration seems to stabilize. However, this is still by far not good enough to compete with K-logit. With 6 clusters the time is 75 x 22 for K-logit and 45 x 100 for Latent Class, so K-logit is still three times as fast. That is by the way also the general conclusion, for all clusters together it takes 4 hours with K-logit and 11 hours with Latent Class.

And, remember, not only in this example is K-logit three times as fast but also in the previous example where we had 18 versus 53 minutes. When we compare the two examples, we can also draw the conclusion that a sample size which is four times as big, causes a calculation time which is more than 12 times as long. In other words, there is an almost square relationship between the number of respondents and the calculation time, both for K-logit and for Latent Class.

3 THE QUALITY / INTERPRETATION OF THE SOLUTIONS

3a With three clusters

It is difficult to give an absolute opinion about the quality of the cluster solutions. But a starting point is that we can compare the interpretation of the solutions. You should know first that a certain K-logit or Latent Class solution is not necessarily *the* optimal solution. The program may converge to a *local* optimum. If you haven't heard about this term, you should imagine that the programs are trying to find the highest top of a mountain range in the fog. Because of the fog they can do no more than just keep going higher every step and when they can't go any higher they assume that this is the top. But it may be *a* top instead of *the* top. Therefore, when we run for a certain number of clusters, we should actually run it several times with different starting points each time and then hope that at least one of the times the program really reaches the optimal solution.

So we can make three types of comparisons: we can compare the Latent Class solutions with each other, we can compare the K-logit solutions with each other and we can compare the best Latent Class solution with the best K-logit solution.

Fortunately in both examples it appears that with *three clusters*, all Latent Class runs converge to the same solution. With so few clusters as three we may draw a preliminary conclusion that Latent Class converges to *the* global optimum rather than to *a* local optimum. The same applies to K-logit.

Furthermore with three clusters, it appears that the cluster sizes and the utility values of the clusters are nearly the same for K-logit and for Latent Class. In addition, not very surprisingly, Latent Class classifies almost every respondent uniquely in one cluster just as K-logit does per definition. To be more exact, the average maximum clustership probability in Latent Class is around 95%. The general conclusion is that with three clusters Latent Class and K-logit produce the same.

Sheet 8

Car study, 3 clusters.

	K-logit cluster 1	Lclass cluster 1	K-logit cluster 2	Lclass cluster 2	K-logit cluster 3	Lclass clustr 3
size	37%	37%	37%	36%	26%	27%
pack A	-.14	-.14	-.03	-.06	.03	.03
pack B	-.26	-.22	-.09	-.08	.12	.06
pack C	.01	-.02	-.08	-.10	-.13	-.06
pack D	.07	.08	.13	.12	-.03	.00
pack E	.01	-.03	-.16	-.09	-.16	-.21
pack F	.30	.33	.22	.21	.17	.19
design I	.17	.15	-.01	.00	.10	.05
design II	-.18	-.17	-.02	.00	-.07	-.08
design III	.01	.02	.00	-.02	.10	.14
design IV	.00	.00	.04	.02	-.13	-.11
Chrysler	-1.46	-1.40	.27	.26	-1.45	-1.30
Buick	-.05	-.04	1.04	1.07	.33	.28
Ford	2.16	2.08	-.80	-.86	-.27	-.29
Chevrolet	-.65	-.64	-.50	-.47	1.38	1.31
$ 23,000	.53	.54	.97	.95	.17	.21
$ 24,000	.34	.34	.90	.86	.43	.51
$ 25,000	.09	.10	-.04	-.02	.43	.38
$ 26,000	-.20	-.23	-.60	-.56	-.17	-.20
$ 27,000	-.77	-.75	-1.23	-1.21	-.86	-.91

Sheet 9

Beer study, 3 clusters.

	K-logit cluster 1	Lclass cluster 1	K-logit cluster 2	Lclass cluster 2	K-logit cluster 3	Lclass clustr 3
size	38%	39%	32%	32%	30%	30%
4-pack	.40	.37	.30	.30	.22	.23
6-pack	-.40	-.37	-.30	-.30	-.22	-.23
Heineken	3.27	3.12	1.87	1.83	.88	.89
Tuborg	1.53	1.49	1.04	1.02	-.73	-.77
Corona	1.27	1.14	1.14	1.12	.06	.09
Budweiser	-.57	-.39	.14	.13	-1.25	-1.39
Palm	1.17	.91	.91	.90	-.85	-.79
Foster	-.02	-.11	-.72	-.69	.80	.81
Michelob	-1.88	-1.79	-1.50	-1.45	.15	.16
Grolsch	-2.22	-1.98	-1.21	-1.15	.48	.49
budget brand	-2.56	-2.41	-1.67	-1.70	.47	.50
with alcohol	1.43	1.37	-.27	-.26	.83	.82
alcohol free	-1.43	-1.37	.27	.26	-.83	-.82
price 1	.05	.05	-.04	-.03	.12	.12
price 2	-.06	-.06	.04	.05	-.03	-.03
price 3	.03	.05	.07	.06	.12	.12
price 4	-.05	-.06	.01	.01	-.10	-.10
price 5	.03	.02	-.08	-.09	-.11	-.10

3b With six clusters

With six clusters I won't bother you with all the detailed numbers, I'll just describe the clusters by their size and main characteristics. The car study results in the following three replications.

Sheet 10

K-logit

cluster description	size	utility value of lowest price
Replication 1		
1. Ford	28%	0.41
2. price sensitive	17%	2.67
3. Chevrolet	17%	0.00
4. Buick	13%	0.17
5. somewhat price sensitive somewhat package sensitive	15%	1.35
6. somewhat price sensitive and prefer Chrysler	9%	1.04
1. Replication 2		
2. Ford	26%	0.26
3. very price sensitive but also prefer Ford	18%	3.07
4. Chevrolet	16%	0.00
5. price sensitive, Buick or Chevrolet	15%	1.99
6. Buick	13%	0.40
7. somewhat price sensitive and prefer Chrysler	12%	0.92
1. Replication 3		
2. extremely price sensitive	20%	6.03
3. Ford	19%	0.32
4. Buick	18%	0.29
5. Chevrolet	18%	0.00
6. Ford (or Buick), somewhat price sensitive	14%	1.31
7. heterogeneous, slight preference for Ford	11%	0.48

As you can see, the replications hardly have any overlapping results. The only cluster they have in common is the completely price-insensitive Chrysler cluster. According to K-logit the second replication is the best one, because it has the highest χ^2 value. Well, we just have to believe that, but at the same time we have to realize that—with three such different replications—we may not have reached the global-optimal replication yet.

Sheet 11

Latent Class

cluster description	size	utility value of lowest price
Replication 1		
1. price sensitive	27%	1.92
2. Ford	22%	0.55
3. Buick or Chrysler	15%	0.51
4. Ford (or Buick)	14%	0.82
5. Chevrolet, definitely no Ford	13%	0.00
6. Chevrolet, definitely no Chrysler	9%	0.00
Replication 2		
1. price sensitive	27%	1.88
2. Chevrolet	18%	0.00
3. Buick or Chrysler	16%	0.46
4. Ford, somewhat package sensitive	15%	1.78
5. Ford	14%	0.00
6. Ford (or Buick)	11%	0.92
Replication 3		
1. Ford	34%	0.46
2. Chevrolet	16%	0.00
3. price sensitive, Buick or Chevrolet	15%	1.60
4. very price sensitive	14%	4.90
5. Buick	14%	0.38
6. Chrysler	8%	2.26

With Latent Class, the first and the second replication are quite similar. The main difference is that the first replication shows two Chevrolet clusters (which are combined in the second replication) and the second replication shows two Ford clusters (which are combined in the first). So it would seem easy to conclude that we should rather base our analysis on five clusters, namely the cross-section of the two replications. However, Latent Class indicates that the *third* replication is actually the best one, with the highest log-likelihood value.

The conclusion is that this example does not give a clue whether K-logit or Latent Class is better. The only thing one might say is that a breakdown in six clusters is apparently too much with this data, and K-logit and Latent Class both provide good information about this fact by offering different results per replication.

In the other example we have better results. Let’s first check K-logit.

Sheet 12

K-logit

cluster description	size
Replication 1	
1. with alcohol, 1 Heineken, 2 Tuborg, 3 Corona	29%*)
2. heterogeneous, slight preference for budget brands	20%
3. with alcohol, 1 Heineken, 2 Corona/Foster/Tuborg	15%*)
4. heterogeneous, slight preference for Palm/Corona/Heineken, slight preference for alcohol-free	13%
5. Tuborg or Heineken	12%
6. 1 Heineken, Corona or Palm	11%
Replication 2	
1. 1 Heineken 2 Tuborg 3 Corona/Palm, with alcohol	30%**)
2. Heineken/Tuborg/Corona	27%
3. 1 budget brands 2 Heineken and various B-brands, with alcohol	12%
4. with alcohol, Heineken/Foster/Corona	11%*)
5. with alcohol, no strong brand preference (though first place for Heineken)	11%*)
6. heterogeneous, slight preference for Foster	8%
Replication 3	
1. with alcohol, 1 Heineken 2 Tuborg/Corona	33%*)
2. 1 budget brands 2 various B-brands 3 Heineken	16%
3. with alcohol, no strong brand preference (though first place for Heineken)	15%
4. Heineken, Corona or Palm	3%
5. Heineken or Tuborg	12%
6. heterogeneous, slight preference for alcohol free	12%

*) with a high value for the “NONE” option
**) with a very high value for the “NONE” option

The largest cluster has always the same interpretation. Furthermore the Heineken/Tuborg cluster and the Heineken/Corona/Palm cluster re-appear every time (though in the second replication they are combined), as well as the alcohol (/Heineken) cluster, the budget brand cluster and the heterogeneous “rest cluster”. Especially the first and third cluster are look-alikes and they also appear to have a similar χ^2 value, which is clearly higher than for the second replication.

With Latent Class the replications are even exactly equal with two replications and nearly equal with the third. So I can suffice with showing you one uniform description of the Latent Class clusters:

Sheet 13

Latent Class

cluster description	size
Replication 1 / 2 / 3	
1. with alcohol, 1 Heineken 2 Tuborg/Corona	36%*)
2. heterogeneous, slight preference for budget brands, slight preference for with alcohol	18%
3. with alcohol, 1 Heineken (2 all others)	13%
4. Heineken/Corona/Palm	12%
5. heterogeneous, slight preference for Heineken, slight preference for alcohol free	11%
6. Heineken or Tuborg	11%

*) with a high value for the "NONE" option

If you compare this solution with K-logit, you can see that it is also close to the first and third K-logit replications. So my overall conclusion from these two examples is that:

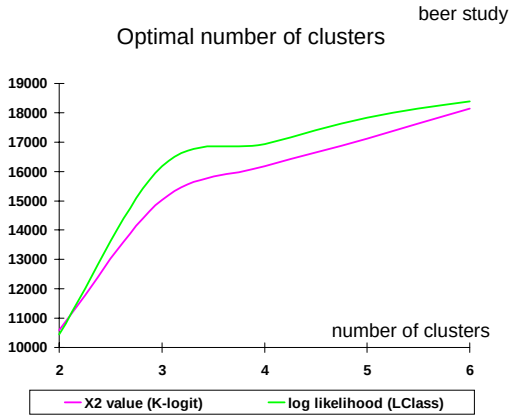
- if K-logit provides poor results, Latent Class does it also;
- if K-logit provides good results, Latent Class provides even better results.

3c The optimal number of clusters

The best way of determining the optimal number of clusters is in my view to check the consistency of replications. However, some people use an easier way and check the slope of the goodness-of-fit measure of the model when increasing the number of clusters (in case of K-logit it is the slope of the χ^2 value, in case of Latent Class the slope of the log-likelihood value). The reasoning behind is that if the goodness-of-fit measure does not improve a lot anymore when increasing the number of clusters, it does not make sense to use such a high number of clusters.

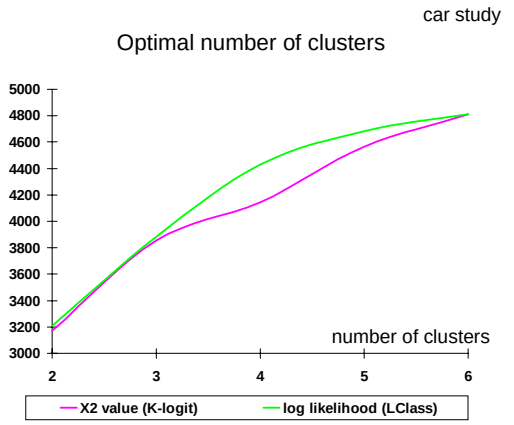
In the following graphs I have rescaled the log-likelihood value to make it comparable with the χ^2 value.

Sheet 14



In case of the beer study, where we had consistent results even with six clusters, the graph of the χ^2 value and log-likelihood value both suggest that three clusters would be optimal. There is no difference between K-logit and Latent Class in this respect and in addition I am inclined to conclude that checking the slope is not an appropriate method to determine the optimal number of clusters for both.

Sheet 15



In case of the car survey, the graphs do not give any clue about the optimal number of clusters, or we can also say that apparently two clusters is sufficient. Again there is no difference between K-logit and Latent Class.

4 THE ADDED VALUE OF LATENT CLASS: INDIVIDUAL PROBABILITIES

It is obvious that Latent Class gives more information about an individual respondent: it does not provide an absolute cluster membership (as K-logit does) but rather a probability distribution across the clusters for each individual. I will devote most of my attention to the six-cluster solution of the beer survey. That is because with the three-cluster solution nearly all respondents have a very high probability belonging to one single cluster. To be more precise, 87% of the respondents have a probability of 90% or more and 73% of the respondents even have a probability of 99% or more belonging to one single cluster. So the three-cluster solution is not so interesting to check.

The six-cluster solution has "only" 70% of respondents who have a probability of 90% or more belonging to one cluster. Actually that is still quite a lot, and it implies more or less that K-logit which classifies all respondents uniquely into one cluster, is not such a bad model. In addition still 40% of the respondents have a probability of 99% or more.

The question for me is now: does the information provided by Latent Class through the utility values of the clusters plus the probability of one respondent belonging to clusters, provide sufficient information about the choice behavior of that respondent? I can easily check this by running Choice-Based Conjoint for that one respondent. Although with so few observations results often tend to plus or minus infinite, that makes little difference in the interpretation.

First I list a few respondents belonging almost uniquely to cluster 6, which is the Heineken/Corona/Palm cluster. The average utility values for these brands are +2.

Sheet 16

Respondents who uniquely belong to cluster “Heineken-Corona-Palm”

	total cluster	resp1	resp2	resp3	resp 4	resp 5
4-pack	.24	.29	.03	.51	1.55	.13
6-pack	-.24	-.29	-.03	-.51	-1.55	-.13
Heineken	*2.19*	*8.86*	*-2.29*	*5.27*	*6.79*	*7.71*
Tuborg	-1.25	-5.66	-2.43	-1.30	-1.31	-.21
Corona	*2.23*	*5.90*	*-1.30*	*4.64*	*3.27*	*-.68*
Budweiser	-.04	-5.52	-.59	.54	-4.91	-2.45
Palm	*2.30*	*-3.24*	*10.03*	*8.34*	*6.14*	*9.99*
Foster	-1.29	-4.24	-1.15	-2.00	-2.69	-3.59
Michelob	-1.36	2.89	-1.70	-2.91	2.79	-5.14
Grolsch	-1.06	-0.70	1.38	-6.67	-7.06	-2.42
budget brand	-1.73	1.72	-1.95	-5.91	-3.02	-3.21
with alcohol	-.04	-2.81	-.03	2.42	-.17	1.22
alcohol free	.04	2.81	.03	-2.42	.17	-1.22
price 1	.08	2.14	.54	1.32	2.57	.00
price 2	.08	2.14	.22	1.32	2.57	.00
price 3	.08	-.62	.22	.44	-1.71	.00
price 4	-.09	-1.83	-.49	-1.54	-1.71	.00
price 5	-.14	-1.83	-.49	-1.54	-1.71	.00
NONE	0.64	-2.45	3.78	-1.84	3.58	-2.79

Two out of the five respondents (nos 3 and 4) match exactly the profile of the total cluster. Of course the utility values are more extreme, but the message is clear: these three brands are on top.

The other three respondents have contradictory values for certain attribute levels but are more extreme than the total cluster with respect to one or two attribute levels which have positive values only in this cluster. The first respondent has high values for Heineken and Corona (though not for Palm); because Corona is not so positive in any of the other clusters, it is uniquely classified in this cluster. Something similar happens with the second respondent, she has an extremely high value for Palm (but not for Heineken or Corona); because Palm is positive only in this cluster, she is uniquely classified in this cluster.

In short, even if a respondent is classified uniquely in one cluster, we cannot draw conclusions for this single respondent. Latent Class is not sufficient and we still need to calculate individual utility values if we are interested in this individual.

On the other hand, when a respondent is clearly classified in two clusters or more, then the individual results are quite similar to the weighted average of the clusters to which this individual belongs. It is difficult for me to show this in detail, because it is almost impossible to find respondents who share the same clusters; there are so many combinations possible. But anyway, here the Latent Class figures are really useful when you actually want to know the individual results. Unfortunately, as I said earlier, only a minority of respondents are classified in more than one cluster.

5 CONCLUSIONS

The main purpose of my paper is the comparison between K-logit and Latent Class. They have the following in common:

- The calculation time is proportional almost to the square of the number of respondents. Especially with many respondents it is recommendable to start with a limited number of clusters and check if that is sufficient.
- They do not necessarily reach the *global*-optimal solution. However, there is evidence (no proof) that they do reach it when you have 3 clusters or less.
- As they may not reach the global-optimal solution, you need to run multiple replications with each number of clusters.

But they are different in the following items:

- K-logit takes three times less calculation time than Latent Class, when you run 2 to 6 clusters. This result holds regardless the number of respondents.
- The calculation time of Latent Class is proportional to the number of clusters; the calculation time of K-logit is less than proportional. With a high number of clusters it is especially recommendable to use K-logit.
- If K-logit provides poor results, Latent Class does it as well. But if K-logit provides good results, Latent Class provides even better results.
- Latent Class gives more insight into individual choices since it is capable of calculating probabilities per cluster per respondent. However, in practice the majority of respondents are still assigned to one cluster with a probability of more than 90%.

INDIVIDUAL UTILITIES FROM CHOICE DATA: A NEW METHOD

Richard M. Johnson
Sawtooth Software, Inc.

BACKGROUND:

Researchers are becoming increasingly interested in choices. One of the main reasons to ask respondents for choices, as opposed to rankings or ratings, is that choice tasks more closely mimic what respondents actually do in the market place. However, choices are an inefficient way to obtain preference information. Before making each choice the respondent must study several product profiles. The answer only indicates which alternative is preferred, providing no information about intensity of preference, reasons for preference, or which alternative might be preferred if the one chosen were not available.

Because choices provide relatively little information from each respondent, choice data are most often analyzed by first aggregating data from all respondents. This necessarily assumes that all respondents are essentially similar, since aggregate methods cannot distinguish between real differences among respondents who have unique preferences and random response error.

Recently, there have been several new approaches to recognizing heterogeneity in choice data:

1) Latent class methods, such as that employed by DeSarbo, Ramaswamy, and Cohen (1995) and implemented in Sawtooth Software's CBC Latent Class Module, accommodate individual differences by recognizing multiple segments. However, these methods still assume that the individuals in each group are homogeneous. This assumption is at variance with most researchers' intuition, so latent class methods may understate the true amount of variety among individuals.

2) Zwerina and Huber (1996) collected desirability ratings for attribute levels, and then constructed an efficient choice design for each individual which permitted estimation of individual utilities. This was done by constructing choice tasks in which respondents had to consider alternatives that were nearly equally attractive. They showed that it is possible to obtain individual utilities from choice data.

3) Hierarchical Bayes methods were studied with full profile conjoint data by Lenk, DeSarbo, Green, and Young (1996), and with trade-off conjoint data by Allenby and Ginter (1995) and Allenby, Ginter, and Arora (1997). In all three cases, hierarchical Bayes analysis was able to provide reasonable estimates of individual utilities. Neither study dealt with choice data, but it seems likely that similar results would have been achieved if they had. Hierarchical Bayes methods may turn out to be the best way of analyzing choice data, but they are so intensive computationally that their widespread adoption may have to await faster computers.

This paper introduces a simple way of extending latent class analysis or other clustering techniques to estimate individual utilities, thus avoiding the assumption of homogeneity within class.

INDIVIDUAL UTILITIES FROM LATENT CLASS:

The latent class model assumes that each individual belongs to one and only one class. When applied to choice data, it estimates a set of utilities for each class, as well as the probability that each individual belongs to that class. For example, latent class analysis of a hypothetical data set might yield results like those in the first two tables:

Table 1
Latent Class Utilities
(Hypothetical Data Set)

	Utilities	
	--Group--	
Brand	1	2
A	0.5	1.0
B	-0.5	-1.0
Size		
Small	2.0	-1.0
Medium	-1.0	0.0
Large	-1.0	1.0

Table 1 shows utilities identifying the preferences of two groups. Both prefer Brand A, but the first group prefers the small size and the second prefers the large size. Also, size is relatively more important for group 1, and brand is relatively more important for group 2.

Table 2
Individual Probabilities of Group Membership
(Hypothetical Data Set)

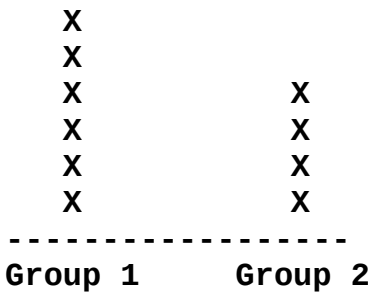
	--Group--	
Individual	1	2
1	.99	.01
2	.98	.02
3	.80	.20
4	.75	.25
5	.60	.40
6	.55	.45
7	.49	.51
8	.30	.70
9	.10	.90
10	.01	.99

The latent class model makes no provision for individual utilities, since it assumes that each individual belongs to one group or the other. The only information provided about individuals is the estimated probability that each individual belongs to each group. Table 2 provides probabilities for 10 individuals, sorted in order of likelihood of belonging to group 1. The first six individuals are more likely to belong to group 1, and the last four are more likely to belong to group 2.

If we were to estimate individual utilities while remaining true to the latent class model, we would estimate each individual's utilities as equal to those of the group for which that individual has highest probability. Note that individual 7 would be given the group 2 utilities, even though estimated to be almost equally likely to belong to group 1.

Figure 1 is a picture of the distribution of respondents in space according to the latent class model. The two groups of relative sizes 60% and 40% are shown ***concentrated*** at two points on a line.

Figure 1
Two Latent Classes



Latent class users have developed a heuristic way of estimating individual utilities that, although inconsistent with the underlying model, has nonetheless seemed intuitively useful. That is to estimate individuals' utilities by using their probabilities of belonging to each group as weights applied to that group's utilities. For example, the utilities for individual 7 would be estimated by taking .51 times the group 1 utilities plus .49 times the group 2 utilities. This produces a unique estimate of each individual's utilities.

Figure 2
Distribution of Individuals,
Estimated By Probability Weighting

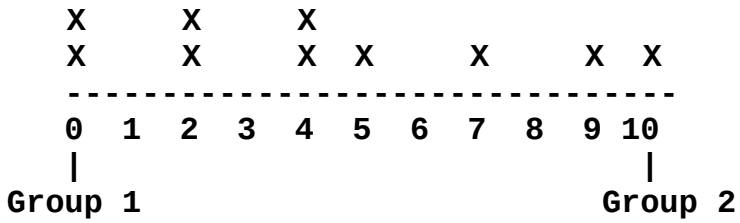
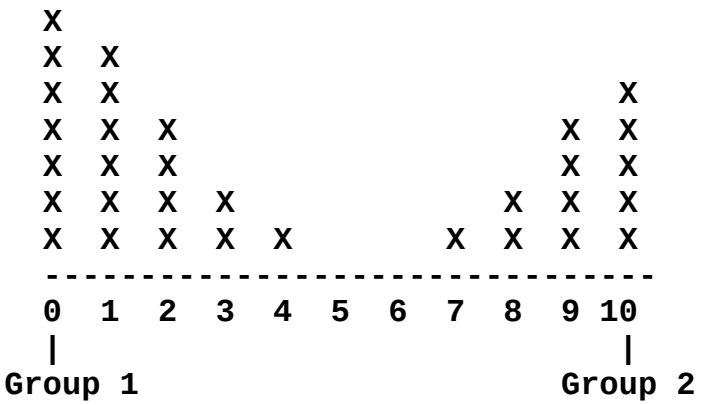


Figure 2 shows individuals distributed on the line separating the two groups. Since everyone's utility is expressed as a weighted combination of the two groups' utilities, all individuals lie in a one-dimensional space. Those with very high probabilities of belonging to group 1 are on the far left, and those with very high probabilities of belonging to group 2 are on the far right. The individual who was nearly 50/50 in probability is in the middle. A distribution like this seems to make more sense than assuming all individuals to be concentrated at two points, but it leads to an unintuitive result: because the weights are probabilities, all individuals lie ***between*** the two groups' locations.

If we had a large number of individuals who fell into two relatively distinct groups, we might see a U-shaped distribution of individuals, with most of them at the two ends of the distribution, as in Figure 3.

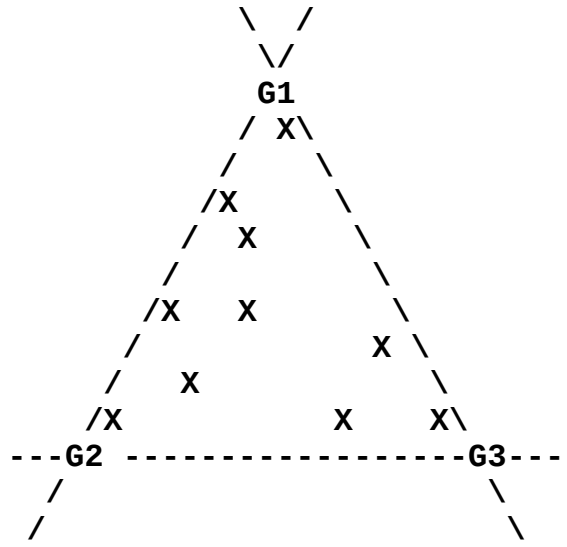
Figure 3
A U-Shaped Distribution of Individuals



This is sharply at odds with the intuition of most of us, who might expect to see something more like two overlapping normal distributions, with individuals distributed on ***both*** sides of the points describing each class.

But if we use probabilities as weights, individuals ***must*** lie within the convex hull of the configuration defined by the groups. With three rather than two groups, all points would lie in a plane, and would be concentrated within the triangle defined by the three groups, as in Figure 4.

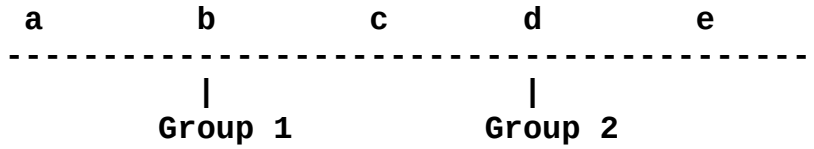
Figure 4
Individual Estimates With 3 Groups With Probability Weighting



PERMITTING NEGATIVE WEIGHTS:

We can let individuals lie outside of the convex hull, and therefore conform more clearly to our intuition, merely by relaxing the non-negativity constraint on the weights. Consider the case of two groups on a line:

Figure 5
Individual Positions With Possibly Negative Weights



In Figure 5 each individual has a position corresponding to his/her weights for the two groups.

Individual b, has a weight of 1.0 for group 1 and 0.0 for group 2, so is positioned at the same point as group 1.

Individual d has weights of [0.0, 1.0], so is positioned at the same point as group 2.

Individual c has weights of [0.5, 0.5], so is positioned midway between the two groups.

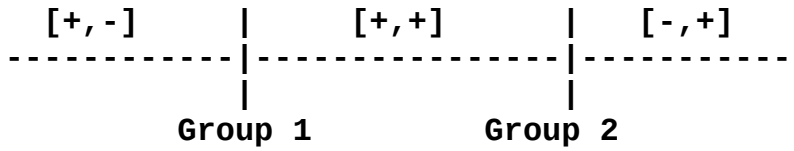
Notice that for each of these individuals, the sum of weights is 1.0, which is a requirement for any point located on the line. Individuals a and e, which lie outside of the domain between the two groups, have both positive and negative weights:

Individual a has weights of 1.5 for group 1 and -0.5 for group 2. This translates into a position to the left of group 1.

Individual e has weights of 1.5 for group 2 and -0.5 for group 1. This translates into a position to the right of group 2.

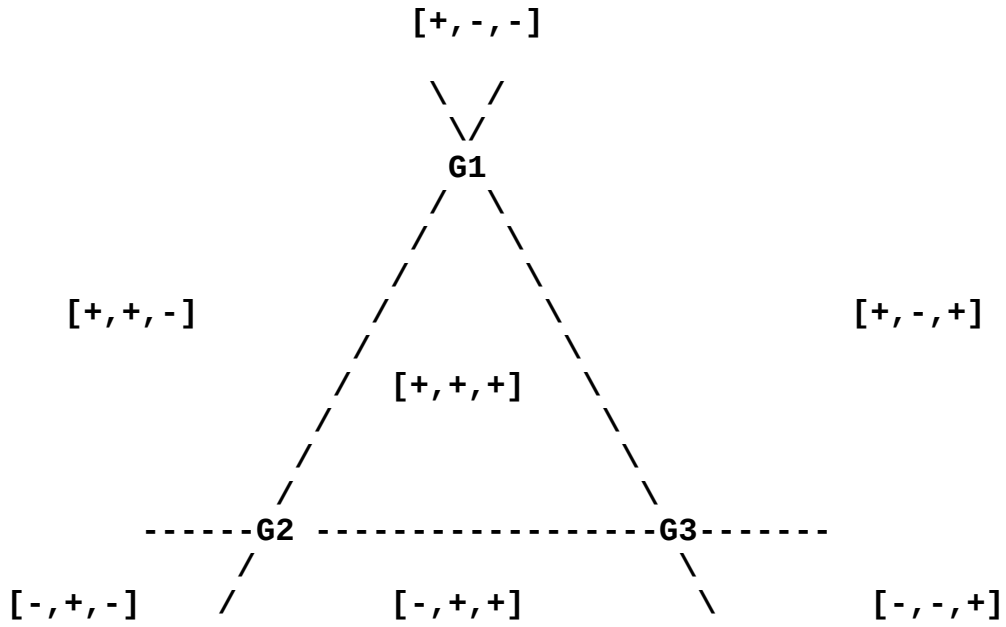
By permitting negative weights, we open up the possibility that individuals could be distributed in a symmetric way around the group's point. With two groups, the three regions of the line containing the group points correspond to different patterns of signs among individuals' weights, as shown in Figure 6:

Figure 6
Patterns of Signs for Individual Weights



Likewise, with three groups, the various regions of the plane containing the group points correspond to different patterns of signs among individuals' weights. There are three patterns, as shown in Figure 7:

Figure 7
Individual Positions in Two-Space
With Possibly Negative Weights



We next consider a simple method for finding weights to apply to group utilities to derive individual utilities that best estimate that individual's choices.

ESTIMATION OF UNRESTRICTED INDIVIDUAL WEIGHTS:

To estimate individual weights we use multinomial logit regression, the same algorithm used to estimate utilities in individual or aggregate analysis.

The first step is to perform a latent class analysis, or to use some other clustering method, to obtain utility estimates for two or more groups.

The second step is to use those group utilities as independent variables in a separate regression for each individual. We find weights that when used to combine the groups' utilities, produce the weighted combination of utilities that best fits that individual's choice data, using a maximum likelihood criterion.

To clarify, the differences between this approach and the more conventional use of multinomial logit regression are as follows:

When used to estimate ***utilities:***

The independent variables are from a design matrix of ***ones and zeros*** indicating the specific attribute levels involved in choice alternatives.

The parameters estimated are utilities for individual attribute levels, of which there are usually ***many***. For example, with six attributes, each with five levels, there would be 6 * (5 - 1) = 24 parameters to be estimated for each individual.

The dependent variables are the observed choices made by respondents.

When used to estimate ***individual weights***:

The independent variables are **utility sums** for choice alternatives, evaluated for each group.

The parameters estimated for each individual are weights for each group, of which there are **few** (say, between 2 and 10).

The dependent variables are the same observed choices made by one respondent.

This method shares with latent class and hierarchical Bayes the characteristic that data from all respondents are involved in the estimation for each respondent. Although each respondent has a unique set of utilities, they are constrained to be linear combinations of the underlying groups' utilities. In the case of two groups, each individual lies somewhere on the line joining those groups. In the case of three groups, each individual lies somewhere in the plane defined by the three groups. In exchange for that restriction, we need to estimate only a few parameters for each respondent, equal to the number of groups - 1, which should increase robustness and decrease data requirements. This method is similar in some respects to an approach suggested by Hagerty (1985), who used Q-type factor analysis to solve for individual utilities in a space of reduced dimensionality for ratings-based conjoint context.

The success of this approach should depend very little on the total number of attributes and levels, but strongly on the number of choice tasks performed by each respondent. Further, one would expect that prediction of holdout choices would be best for a middling number of groups. With too few groups, the space will not be rich enough to capture every individual's preferences adequately. With too many groups, there is likely to be over-fitting.

We now turn to a Monte Carlo test of this method with synthetic data sets for which the correct results are known.

ANALYSIS OF SYNTHETIC DATA:

The first example considers three groups of synthetic respondents and three attributes, each with three levels. "Average" utilities were constructed for each group as follows:

Table 3
Average Utility Values for 3 Groups

	Group 1	Group 2	Group 3
Att 1 Lev 1	0	1	-1
Att 1 Lev 2	-1	0	1
Att 1 Lev 3	1	-1	0
Att 2 Lev 1	0	1	-1
Att 2 Lev 2	-1	0	1
Att 2 Lev 3	1	-1	0
Att 3 Lev 1	0	1	-1
Att 3 Lev 2	-1	0	1
Att 3 Lev 3	1	-1	0

In this example each group has the same average utilities for every attribute just to keep things simple. (Note also that the groups' utilities sum to zero across rows.)

Heterogeneous individual utilities were constructed for 100 synthetic respondents in each group by adding random values to the average utilities for that group. The values added to create heterogeneity were independent and normally distributed, with mean of 0 and standard deviation of either 1, 2, or 3, depending on how much heterogeneity was being modeled. A fourth population was also constructed containing ***no*** within-group heterogeneity. Each individual's "true" utilities were saved for later comparison with estimated values.

A customized computer-administered questionnaire was constructed for each respondent, using Sawtooth Software's CBC System. Individuals had either 10, 20, 30, or 40 choice tasks. Each task presented three alternatives consisting of concepts specified on all attributes, and did not include a "none" option. Respondent choices were modeled by forming the sum of that respondent's utilities for each alternative, adding to each sum a random normal variable with standard deviation of unity, and then choosing that alternative with the highest modified sum.

For each combination of heterogeneity and questionnaire length, latent class analyses were done with from 2 through 6 groups. Each latent class analysis was replicated 5 times from different starting points, and the best-fitting solution was used in each case. We consider 80 combinations of treatments: 4 levels of heterogeneity times 4 levels of questionnaire length, times 5 different numbers of latent classes.

For each combination, individual utilities were estimated using the traditional probability weighting method and the new method with unrestricted individual weights. Most individuals had probabilities in the .90s of belonging to one group or another, so the latent class estimates were similar to what would be obtained just by classifying each individual into his highest-probability group. The quality of estimation was measured by computing the r-square between true utilities and estimates produced by each method. We summarize the 160 r-square values in terms of main effects in Table 4.

Table 4
Average R Square Values For Each Method

	Probability Weights	Unrestricted Weights	Ratio
Heterogeneity = 0	.899	.900	1.00
Heterogeneity = 1	.524	.709	1.35
Heterogeneity = 2	.355	.630	1.77
Heterogeneity = 3	.325	.616	1.90
10 Tasks	.526	.645	1.23
20 Tasks	.523	.714	1.37
30 Tasks	.529	.744	1.41
50 Tasks	.525	.752	1.43
2 Dimensions	.322	.535	1.66
3 Dimensions	.540	.654	1.21
4 Dimensions	.567	.722	1.27
5 Dimensions	.589	.793	1.35
6 Dimensions	.611	.864	1.41
Overall Average	.526	.714	1.36

Average r square values are better for unrestricted weights, and substantially so in most cases:

The effect of heterogeneity: The methods are nearly equal in the case of zero heterogeneity, which latent class assumes. (The unrestricted weights win when fewer than three dimensions are used, having an advantage due to the zero-sum nature of the utilities, but the probability weighting does better when three or more groups are used.) The unrestricted weight method is superior when there is any heterogeneity, and its margin of superiority increases rapidly as heterogeneity increases.

The effect of questionnaire length: The success of probability weighting appears insensitive to questionnaire length, but the success of unrestricted weights is much more so, so its relative superiority increases as the number of tasks increases. Fortunately, Johnson and Orme (1996) found that interviews with many choice tasks per respondent suffered no degradation in data quality as the interviews progressed. They studied only interviews with up to 20 choice tasks, but the trends in their data suggested that even longer interviews, such as those with up to 30 choice tasks, should be feasible without loss of data quality.

The effect of number of dimensions: Recall that these data sets were constructed to contain three groups. Probability weighted utilities are quite badly estimated when ***too few*** groups are used, but there is some benefit from using more than the underlying three groups. For the new method, performance increases more dramatically with the number of dimensions, up to the limit of 6, which is the same as the number of independent utility values estimated for each respondent.

These results indicate that individual utilities are better estimated by permitting weights to have either positive or negative signs, rather than by using probabilities of membership as weights.

The next simulation focuses on hit rates rather than recovering true utilities. It uses six attributes, each with three levels. Populations of synthetic respondents were again generated, with different amounts of heterogeneity. Each population contained three groups of respondents with the average utilities in Table 5.

Table 5

Average Utilities for Three Groups

	Group 1 Levels			Group 2 Levels			Group 3 Levels		
Attribute	1	2	3	1	2	3	1	2	3
1	1	0	-1	0	1	-1	1	-1	0
2	-1	0	1	0	-1	1	-1	1	0
3	0	1	-1	1	-1	0	1	0	-1
4	0	-1	1	-1	1	0	-1	0	1
5	1	-1	0	1	0	-1	0	1	-1
6	-1	1	0	-1	0	1	0	-1	1

The three levels of an attribute always had average utilities of 1, 0, and -1. Within each group, each attribute displayed one of the six possible patterns of those values, and no groups had identical values for any attribute.

Heterogeneous individual utilities were constructed for 100 synthetic respondents in each group as before, by adding random values to the average utilities for that group. The values added to create heterogeneity were independent and normally distributed. In this simulation, heterogeneity levels were 0, 1, and 2.

A unique, computer-administered questionnaire was constructed for each respondent, containing 50 choice tasks. Each task presented three alternatives consisting of concepts specified on all six attributes, and did not include a "none" option. Respondent choices were again modeled by forming the sum of that respondent's utilities for each alternative, adding to each sum a random normal variable with standard deviation of unity, and then choosing that alternative with the highest modified sum.

For each population, a latent class analysis was done using only the ***first 20 tasks*** for each respondent. Solutions were obtained for 2 through 5 groups. Each was replicated five times from different starting points, and the solution with highest likelihood was retained.

Utilities were estimated for each respondent using the traditional method of probability weighting, and also using unrestricted individual weights, with separate estimates based on the first 10, 20, 30, and 40 choice tasks. Tasks not used in the estimation were treated as holdouts, and average hit rates were computed for those choices as predicted by each set of utilities.

There are three levels of within-group heterogeneity (0, 1, 2), 4 numbers of latent classes (2, 3, 4, 5), and 4 numbers of choice tasks (10, 20, 30, 40). For each combination there is a hit rate for the new method and a corresponding hit rate for the traditional method based on the same set of holdout tasks. We again summarize the data in terms of main effects in Table 6.

Table 6
Hit Rates for Each Method

	Probability Weights	Unrestricted Weights	Ratio
Heterogeneity = 0	.773	.748	.97
Heterogeneity = 1	.608	.631	1.04
Heterogeneity = 2	.520	.562	1.08
2 Dimensions	.559	.567	1.01
3 Dimensions	.651	.662	1.02
4 Dimensions	.661	.677	1.02
5 Dimensions	.664	.682	1.03
10 Tasks	.636	.620	.97
20 Tasks	.635	.651	1.03
30 Tasks	.631	.656	1.04
40 Tasks	.633	.661	1.04

The effect of heterogeneity: Hit rates show a pattern similar to that of squared correlations in the previous simulation, although the differences among methods are less dramatic when measured by hit rates. Probability weighting wins when the latent class assumption of no within-group heterogeneity is met. However, unrestricted weights are superior when there is within-group heterogeneity, and more superior as heterogeneity increases.

The effect of number of dimensions: Table 6 also shows how hit rates vary with the number of dimensions used. Unrestricted weighting method shows a slight advantage over probability weighting in each case. However, the more interesting comparison in this table is among rows rather than columns. The data were constructed so as to have three fundamental groups of respondents. For both methods, hit rates are sharply lower when too few groups are considered, but there is no apparent penalty for using too many groups.

The effect of questionnaire length: Since 20 tasks were used in every latent class analysis, probability weighting is insensitive to the number of tasks. Questionnaire length has an effect on relative performance of unrestricted weights. With only 10 choice tasks per respondent, probability weighting wins. There is an increase in relative performance of the new method when going from 10 to 20 tasks per respondent, and its relative performance continues to increase as questionnaire length increases to 30 tasks. There are modest further increases in relative performance as questionnaire length increases from 30 to 40 tasks per respondent, except for the case of greatest heterogeneity.

Summary of Synthetic Data Analysis: Estimating individual utilities by permitting weights with both positive and negative signs seems to work better than the traditional method in the presence of within-group heterogeneity. Its superiority is strongest when there is more within-group heterogeneity, when more dimensions are used in estimation, and when more choice tasks are available for estimation.

We turn now to the study of three data sets from human respondents. We believe human data contain considerable heterogeneity that is not accounted for by multiple segments, so we expect the new method to be successful when enough choice tasks are available per respondent for reliable estimation.

A Consumer Product: These data were furnished by Griggs Anderson Research. The product category was identified as a "computer peripheral," and the data are typical of those obtained in many commercial choice studies. There were 6 attributes with a total of 25 levels. Six hundred consumers responded to a CBC interview in which there were 20 choice tasks. Each task contained three alternatives plus the option of "None." A large proportion (43%) of the choice questions were answered with selection of "None," so that alternative was retained in the analysis.

The first 16 tasks were used to perform a latent class analysis, obtaining solutions for 2 through 9 groups. For solutions involving 5 or fewer groups, 5 replications were conducted from random starting points, and only the best solution was retained for each number of groups. Only one solution was obtained in each case with 6 or more groups.

The CAIC criterion indicated that the 4-group solution was best, while the relative chi square criterion indicated that the 2-group solution was best. Individual utilities were estimated by the traditional method and also by the new method using each latent class solution. Results are shown in Table 7.

Table 7
Hit Rates for Consumer Product Data Set
(16 Choice Tasks for Estimation)

Number Groups	Traditional Method	New Method	Ratio
2	.592	.620	1.05
3	.607	.654	1.08
4	.628	.653	1.04
5	.648	.653	1.01
6	.652	.656	1.01
7	.654	.650	.99
8	.665	.643	.97
9	.665	.653	.98

The new method was slightly superior to the traditional method when 6 or fewer groups were used, but inferior when more groups were used. Our analysis of synthetic data found no penalty for using more than the correct number of groups, but it only considered up to 5 groups. The loss of relative performance here for larger numbers of groups shows that there is danger of overfitting and poorer prediction if too many groups are used with too few choice tasks per respondent.

Although these results are mostly favorable, we should confess that a preliminary analysis of these same data was less so. Initially we had deleted all tasks with answers of "None." That resulted in retaining an average of only nine tasks per respondent. With so few tasks, hit rates for the new method were inferior to those for the traditional method, corroborating the earlier finding that the new method requires a larger number of choice tasks per respondent for success.

An Industrial Product: These data were provided by an end-user company. There were again 6 attributes and a total of 25 levels. A total of 692 individuals responded to a CBC interview containing 16 choice tasks. Each task presented only three alternatives, without the option of "None."

The first 12 tasks were used to perform a latent class analysis, obtaining solutions for 2 through 5 groups. Three replications were conducted from random starting points, and only the best solution was retained for each number of groups. The CAIC criterion again indicated that the 4-group solution was best, while the relative chi square criterion again indicated that the 2-group solution was best. Individual utilities were estimated for each solution by the traditional method and the new method. Results are shown in Table 8.

Table 8
Hit Rates for Industrial Product Data Set
(12 Choice Tasks for Estimation)

Number Groups	Traditional Method	New Method	Ratio
2	.567	.597	1.05
3	.579	.588	1.02
4	.595	.571	.96
5	.609	.572	.94

The new method is better for the two and three group solutions, but inferior when more groups are used. These results confirm that with only a few choice tasks for estimation, the new method has difficulty with over-fitting when latent class solutions contain more than a few groups.

The Zwerina-Huber Data: Zwerina and Huber kindly provided the data from their study cited earlier, in which they were able to estimate individual utilities from choices. Their respondents were 50 MBA students who participated in a two-part computer-administered interview. The subject was laptop computers, and there were 6 attributes, each with 3 levels. One attribute was brand, but the others had a-priori orders which could be used to provide order constraints on utilities within attribute.

In the first session respondents answered six hold-out choice tasks, an attribute rating task, and a full-profile conjoint task. The attribute ratings were used to construct a customized choice questionnaire for each individual containing 30 choice tasks in which the alternatives were approximately balanced in utility. During the second session, respondents answered those 30 choice questions, as well as repeating the initial 6 holdout choices. All choice tasks presented three alternatives, with no “None” option.

Zwerina and Huber estimated individual utilities with multinomial logit analysis, both with and without order constraints conforming to ratings of the desirability of attribute levels. They also estimated utilities from the full profile conjoint exercise and from self explicated ratings. They found that choice-based utilities had better hit rates for predicting holdout choices than utilities from either the full profile or self explicated data.

They found the test-retest reliability for the repeated holdout concepts to be .773. This provides an indication of the amount of error in the holdouts themselves. Their multinomial logit estimates of individual utilities had an average hit rate of .733. When they constrained their estimated utilities to have the same orders within each attribute as respondents’ desirability ratings, their average hit rate rose to .763.

We conducted a latent class analysis of the Zwerina-Huber data, for solutions with from 2 through 8 groups. Because there were so few respondents, 10 replications were conducted from random starting points. Order constraints were imposed for all utilities but brand. The CAIC criterion was optimized for the 5 group solution, but the relative chi square criterion again indicated superiority of the 2 group solution.

Individual utilities were then estimated for each of the latent class solutions, using all 30 choice tasks. The same order constraints were imposed on these estimates as had been imposed on the latent class solutions. We report results for 2 through 7 groups. With larger numbers of groups, at least one group contained a single respondent. Hit rates are given in Table 9.

Table 9
Hit Rates for Zwerina-Huber Data Set
(30 Choice Tasks for Estimation)

Number Groups	Traditional Method	New Method	Ratio
2	.717	.738	1.03
3	.695	.763	1.10
4	.702	.775	1.10
5	.717	.785	1.09
6	.730	.805	1.10
7	.738	.788	1.07

For all solutions beyond the 2-group case, hit rates for the new method tied or exceeded that of Zwerina and Huber (.763), and also the test-retest reliability of the holdout data (.773). They also exceeded hit rates for the traditional method of estimating individual utilities by probability-weighting latent class group utilities. The new method had previously been found to work best when many choices are made by each respondent, and we believe success with these data is due to the fact that 30 choices were available for estimation for each respondent.

Hit rates are of interest to researchers, but managers are often more interested in the accuracy of aggregate share predictions. Zwerina and Huber also examined the accuracy of prediction of ***aggregate*** choice shares for the holdout tasks. They computed the Mean Absolute Error (MAE) between actual choice shares and predictions using a "first choice" or "maximum utility" rule for each respondent. Their reported values were .024 for unconstrained utilities and .041 for constrained utilities. They found that that utilities from choice data were better than full-profile conjoint utilities or self-explicated utilities for predicting aggregate choice shares.

Our constraints were derived from a-priori knowledge of five of the attributes, for which it was obvious that "more is better." Unlike Zwerina and Huber, we imposed no constraints on levels for the brand attribute. Despite the differences in the way our constraints were imposed, our final results were similar. We have computed similar MAE statistics for our constrained estimates, which are reported in Table 10.

Table 10
Mean Absolute Errors in Predicting Choice Shares
for Zwerina-Huber Data Set
(30 Choice Tasks for Estimation)

Number Groups	New Method
2	.118
3	.080
4	.066
5	.050
6	.036
7	.030

Our results generally improve as the number of groups increases. For smaller numbers of groups, our predictions of choice shares are not so good as those of Zwerina and Huber (.041), but our last two cases are better.

DISCUSSION AND CONCLUSIONS:

From a practical point of view, this method of estimating individual utilities from choice data seems to have worked well.

With synthetic data, its predictions were superior to those of latent class analysis when there was within-group heterogeneity and when there were more than about 10 choice tasks per respondent.

With data from human respondents it was generally more successful than latent class analysis, and its superiority was greatest in those cases where there were more choice tasks per respondent.

With 30 tasks per respondent, the new method produced utilities which were slightly superior to individual utilities estimated from individual choice designs, full profile conjoint data, and from self-explicated data.

Like hierarchical Bayes methods, this method employs the general idea of using data from ***all*** individuals to help in the estimation of values for ***each*** individual. However, its method of doing so is simpler and less elegant. Each individual's utilities are estimated as a linear combination of a set of basis vectors from some previous source. We have based the solutions reported here on latent class analyses, but elsewhere we have analyzed several data sets using basis vectors derived with the "latent segment" approach suggested by Moore, Gray-Lee and Louviere (1995) and described in the CBC Latent Class Module manual as "KLogit." Any other clustering method useful with individual choice data would probably work nearly as well.

One attractive aspect of this method is its speed. Compared to the computational effort to obtain a latent class solution, for example, individual utility estimation is trivial, requiring less than one percent as much time as the underlying latent class analysis. If a faster clustering method were used, the entire computation could be quite fast.

One limitation is that the individual utilities are only useful for "first choice" predictions, rather than for traditional logit predictions which depend on their scale. Logit estimates of individual utilities tend to be unstable, and individuals whose choices are fit very well may have utilities that are scaled quite radically. We have handled this problem by scaling each individual's utilities arbitrarily. Perhaps a subsequent likelihood-of-buying task could be used to scale utilities, as is done in ACA.

One of the problems of market researchers, particularly those working with choice data, is that of predicting the market's response to complex combinations of interactions, differential cross effects, and varying similarities among products. It seems likely that all of these problems will be diminished when modeled at the individual level. If so, the payoff of being able to estimate individual-level utilities from choice data will be significant. We hope future research will compare this approach systematically with others, including completely individual estimation and hierarchical Bayes, and will clarify which method or combination of methods is most effective in such complex environments.

REFERENCES:

Allenby, G. M.. and J. L. Ginter (1995) "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 37, Nov, 392-403.

Allenby, G. M., J. L. Ginter, and N. Arora (1997), "On the Identification of Market Segments," Working Paper, Ohio State University.

DeSarbo, W. S., V. Ramaswamy, and S. H. Cohen (1995), "Market Segmentation with Choice-Based Conjoint Analysis," *Marketing Letters*, 6, 137-148.

Hagerty, M.R. (1985) "Improving Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis," *Journal of Marketing Research*, 22, May, 168-184.

Lenk, P. J., W. S. DeSarbo, P. E. Green, and M. R. Young (1966), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2) 173-191.

Johnson, R. M. and B. K. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?," Proceedings of A.R.T. Forum, American Marketing Association.

Moore, W. L., J. Gray-Lee and J. Louviere (1995), "A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation," Working Paper, University of Utah, November.

Zwerina, K. and J. Huber (1996) "Deriving Individual Preference Structures from Practical Choice Experiments," Working Paper, Duke University, August.

ASSESSING THE VALIDITY OF CONJOINT ANALYSIS—CONTINUED

Bryan K. Orme
Sawtooth Software, Inc.
Mark I. Alpert[1]
The University of Texas at Austin
Ethan Christensen
The University of Texas at Arlington

INTRODUCTION

Despite over 20 years of conjoint research and hundreds of methodological papers, very little has been published in the way of formal tests of whether conjoint *really* works in predicting significant *real-world* actions. As ancillary cases, there are interesting questions of whether one method works better than another, and under what circumstances each method should be preferred.

Most of the debate to this point has focused on *calibration of utilities*. Our research focuses on the other side of the equation: *the validity measurement*. In most validity studies, researchers have begged off the measurement of real validity by settling for attempts to predict holdout concepts administered in the same interview. Because the holdout concept is usually so similar (even identical) to the conjoint exercise, most validity studies really only measure internal consistency. When viewed with any perspective at all, calling such exercises validity studies seems a presumptuous stretch. Further, in the typical conjoint validity study, as much as 95% of the effort goes into measuring respondent utilities, and as little as 5% goes into measuring what it is we want to predict. It seems as though validity studies should invest much more in measurement of that which is to be predicted.

There are some widely-recognized shortcomings of conjoint methods. For example, it is thought that respondents sometimes use simplification strategies to answer difficult full-profile tasks. Respondents may consider only the few most important attributes, which would result in exaggerated differences in importance between the most and least important factors. And it is also thought that ACA sometimes errs in forcing individuals to pay attention to every attribute, whether important or not, which would result in ACA's importances being "too flat." Of course, lacking a proper validity study based on real-world purchase observation, both of these claims remain only conjecture.

The title for our research is taken from a paper entitled "Assessing the Validity of Conjoint Analysis" presented by Rich Johnson in the 1989 Sawtooth Software Conference Proceedings. We refer to important points from that paper, and then report an original pilot study which attempts to overcome some of the weaknesses of traditional validation research. This pilot study

[1] Dr. Mark Alpert holds the Foley's Centennial Professorship in Marketing at The University of Texas at Austin. Ethan Christensen is a doctoral student at The University of Texas at Arlington. We thank Rich Johnson and Joel Huber for their insightful comments and direction. The authors accept responsibility for any errors.

featured an intensive holdout exercise which may better reflect real world purchase behavior than traditional holdouts asked during the course of conjoint surveys.

Our main emphasis is on the principles of design for conjoint validity studies. We also compare the results from full-profile, ACA and choice-based conjoint. Our results are not powerful enough to reach strong conclusions about methods, but we think we illustrate a way for strengthening traditional validity studies. Finally, we note the limitations of our pilot study and suggest directions for additional research.

DESIGN CONSIDERATIONS FOR HOLDOUT TASKS

In conjoint validation studies, holdout tasks are not used in the estimation of partworths. They are presumed to represent how the respondent would choose in the real world. Researchers measure validity by comparing how well conjoint utilities predict choices from the holdout tasks.

Ideally, the actual purchase event should be the criterion measurement. Lacking real purchase data, guidelines for constructing experimental holdout tasks include:

- At least one of the holdout tasks should be repeated to assess the reliability of holdout judgements. This allows the researcher to determine the proportion of error in prediction due to errors in the partworths versus errors in response to holdout judgements themselves. It also provides a way to adjust hit rates when comparing results from independent samples of not necessarily the same response reliability.
- The validation measurement should closely mimic the stimulus presentation and depth of processing of the real world purchase event.
- Attribute order effects should be controlled. The holdout task should present the attributes in a different order than full-profile calibration tasks.

HOW REALISTIC ARE TRADITIONAL HOLDOUT TASKS?

The majority of conjoint validity tests have used full-profile evaluations as holdout tasks. Hit rates for correctly predicted choices (from choice sets of pairs, triples, etc.) are a popular measure, as well as correlation or MSE for ratings-based holdouts. It has been argued that full-profile holdouts best represent how products are viewed and evaluated in the real world. We think this is reasonable, but we question whether buyers process and evaluate full-profiles in the real world the same way they do during the context of a survey.

Particularly for high-involvement purchases, respondents exert more effort making real-world decisions than while making judgments in conjoint surveys. Once warmed up to the task, respondents can take as little as 12 seconds on average to make choices in full-profile choice questionnaires (Johnson and Orme, 1996). Huber observes: “Purchases of laptops are generally not made in anything like 30 seconds; people spend significant time discussing a wide range of features” (Huber *et al*. 1992).

Full-profile interviews involving many attributes may encourage respondents to adopt simplification heuristics. By focusing on just a subset of the attributes, respondents can more easily complete long and monotonous conjoint interviews. Simplification strategies can lead to more extreme attribute importances, with relatively little weight given to the factors the respondent

has chosen to ignore. If simplification heuristics are indeed being used in full-profile judgements, it would lead to some critical questions:

- Do respondents focus on just a subset of key attributes in the real world, when many attributes are involved and real dollars are on the line?
- Are hastily-answered full-profile holdout concepts realistic criteria for measuring conjoint validity?
- If not, what criteria should we use for validity comparisons?

These questions form the crux of our research. Again, the ideal validation study would attempt to predict actual purchase behavior. In the absence of real-world judgements, other steps might be taken to improve the quality of the holdout task. For our study, we designed a "Super Holdout Task" (described below) to address in part the shortcomings of tradition holdouts and better simulate real world behavior.

We hypothesize that especially with significant decisions, individuals may broaden their range of attention to product features, perhaps resulting in flatter importances than are captured with traditional full-profile holdout concepts. With this hypothesis in mind, we turn to details of our study design.

STUDY DESIGN

MBAs from The University of Texas at Austin, The University of Texas at Arlington and the University of Washington were employed as respondents. The subject of the study was personal computers for a hypothetical new computer lab at the respondents' respective business schools.

Nine attributes were studied: brand, warranty, microprocessor, number of lab assistants, ergonomic keyboard/mouse, hard drive, RAM, modem/Internet access, and price. All attributes had either two or three levels described in succinct phrases. (See Appendix A for a full listing of attribute levels).

There were two main components of the study, administered in two separate sessions:

1) **Computer-Administered Survey.** Respondents received a packet which included a survey disk programmed using Ci3 along with 22 full-profile conjoint cards printed on cardstock paper. The order of tasks in the survey was:

   a) **Demographic questions**. (Experience with and familiarity with personal computers/ past purchase influence for PCs.)

   b) **Full-profile card-sort/ACA**. (Each respondent received both, in rotated order). Sawtooth Software's CVA system was used to design and analyze the full-profile data. Twenty-two hard-copy cards were sorted into four piles based on preference, and then rated using a 100-pt scale. To control for attribute order bias, two versions of the cards were printed. (See Appendix B for details of the full-profile design). ACA v4.0 was used with default settings.

   c) **Five full-profile holdout choice tasks with three product concepts each**. These tasks were constructed randomly using Ci3. Respondents indicated first and second choices. The second and fifth tasks were identical (with rotated concepts) to measure test-retest re-

liability. Brand was always the first attribute, and Price the last. The interior seven attributes were randomly rotated across respondents.

2. **Super Holdout Task.** After completing the disk-based survey, respondents participated in a 10-minute in-class exercise. Students divided into committees of three to evaluate just one choice task with four PC configurations described in full-profile. These tasks were constructed randomly, varying across committees. Attribute order was randomized for the interior seven attributes.

   Each committee was instructed to reach consensus regarding the best, second, third and worst PC configuration for the new computer lab. After the group evaluations were recorded, respondents were asked to record their personal evaluations—but they did not know beforehand that we would ask for their personal opinions.

We expected that the Super Holdout Task might better reflect real-world behavior than traditional validation tasks, particularly for high-involvement categories. Since more is at stake with high-dollar purchases, buyers spend a great deal of time weighing the pros and cons of available alternatives. Buyers also seek additional information by consulting with others. Also, for business-to-business markets (or even for households), purchase decisions are frequently decided by some sort of committee after some debate.

Another unique aspect of this study is the use of randomized holdout tasks. For many-attribute designs (such as ours) the randomized approach generally achieves a fair degree of utility balance, which is a desirable condition for testing predictive validity. Consistently dominated choice tasks would be less useful, since predicting choices for dominated tasks is trivial. Perhaps most useful is that randomized designs permit group-level utility estimation for the holdout judgments. We are able to compare utilities and importances from four sources: full-profile ratings, ACA, standard holdout choices, and the Super Holdout Task. We expect to assess whether partworths differ between traditional holdouts and the Super Holdout Task.

Based on previous research (Huber 1992, Pinnell 1994), we expected that attribute importances from full-profile and full-profile choice would be more extreme than ACA importances. We also hypothesized that importances derived from the Super Holdout Task would be less extreme than the holdout choice exercise that was part of the computer-administered interview, reflecting greater depth of processing.

TIMING DATA AND RESPONDENT PROFILE

A total of 80 completed surveys were analyzed. Median interview time was 27 minutes for the disk survey, including 13 minutes for the full-profile exercise (time to sort and write scores on the cards), and 8 minutes for ACA. ACA took significantly less time to complete than full-profile, with a t-value for the mean difference in interview time of 6.9.

There were five standard holdout choice tasks during the computer-administered portion of the survey. These took 48, 35, 32, 28 and 26 seconds to complete. Although we don't have timing data for the Super Holdout Task, it took respondents about 10 minutes to complete.

Seventy-four percent of the respondents had been the main decision maker for purchasing a PC before.

CALCULATING HOLDOUT HIT RATES

Conjoint validity is usually assessed by observing how well partworths can predict holdout evaluations. First choice hits are the most common measure. For choice tasks including judgments beyond first choice, we may evaluate hit rates for an expanded set of implied comparisons. For the choice-based holdout tasks in our study, we asked for a full ranking of alternatives. The choice-based tasks in the disk survey involved three product concepts. Assuming the preference order was a, b, c, there are three implied inequalities: a>b, a>c, b>c. The Super Holdout Task presented a choice-based set with four concepts, leading to six implied inequalities (assuming preference order of a, b, c, d): a>b, a>c, a>d, b>c, b>d, c>d.

As mentioned previously, it is desirable to repeat at least one of the holdout concepts to assess test-retest reliability. This would be especially critical if, for instance, one group of respondents had completed ACA and the other group completed full-profile. In order to determine whether the conjoint method that one group received performed better than the other, we would need to adjust hit rates by the test-retest reliability for each group.

For our study, each respondent completed an identical set of calibration tasks (but with rotated task order), so we do not need to be concerned with comparing hit rates across independent samples.

Repeated holdout tasks permit us to calculate a theoretical upper limit for holdout predictability. The second and fifth choice tasks in the disk-based survey were identical (but with concepts rotated). The test-retest reliability for all implied inequalities was 90.0%. Wittink and Johnson (1992) demonstrated that the maximum expected hit rate for predicting a fallible criterion measure is equal to:

$$\pi = \frac{1+\sqrt{(2p-1)}}{2}$$

where π is the maximum expected hit rate and p is the agreement between independent replications of the criterion measure. Given test-retest reliability of 90.0%, the maximum possible hit rate for predicting the standard holdout choices is 94.7%. Since we did not repeat holdout tasks for the Super Holdout Task, we cannot compute its test-retest reliability. Future studies could administer a replication of the Super Holdout Task during the course of the survey to permit test-retest reliability adjustments in the case of independent samples.

TRADITIONAL HOLDOUT HIT RATES

Hit rates for full-profile and ACA are provided in the table below. Due to the small sample size, hit rates for first choices were not very stable. Hit rates for all implied inequalities included more information per respondent, and are used throughout the remainder of this paper.

OLS partworths were calculated for full-profile. *A priori* attributes were constrained to remedy sign reversals using CVA's tieing algorithm. We used a logit transform of the 100-pt purchase likelihood scale response.

Table 1
Traditional Holdout Choice Hit Rates

ACA	76.9%
Full-profile	82.4%

n=80

For the eighty respondents in the pilot study, the full-profile method does a better job predicting the standard holdout choices. The t-value for the mean difference in prediction for full-profile versus ACA for the standard holdouts is 2.85.

Why does full-profile do better than ACA for predicting the standard full-profile holdout choice tasks in our pilot study? Investigating differences between attribute importances and partworths helps answer that question.

ATTRIBUTE IMPORTANCES

We define attribute importances in the standard way, by percentaging the ranges of attribute utilities. Our study design permits us to compute attribute importances from four sources:

1) Full-profile ratings
2) ACA
3) Standard holdout choices (choice-based conjoint)
4) Super Holdout Task (choice-based conjoint)

The two choice-based sources were analyzed in the aggregate using logit, including information from all choices. In general we suggest only using first choices within each task for utility calculation (Johnson and Orme, 1996), but the additional information was valuable for obtaining reasonable estimates given our limited sample. Also, we felt that the bias from second choices reported by Johnson and Orme would have minimal or no impact on the analysis of importances.

Respondents' enthusiasm for (or attention to) non-ordered attributes is not reflected in importances calculated from aggregate utilities when there is disagreement about which levels are preferred. Two of the nine attributes were not *a priori* ordered (brand and ergonomic keyboard/mouse) and were dropped. Relative importances for ACA, full-profile and the traditional (standard) holdout choices are shown in the table below:

Table 2
Attribute Importances

	ACA	Full-Profile	Standard Choices
Internet Access	21.5	24.6	25.7
RAM	19.8	21.6	23.8
Price	15.0	16.5	18.1
Hard Drive	14.0	14.1	12.2
Warranty	11.6	10.2	8.4
Processor	10.7	9.5	9.5
Lab Assistants	7.4	3.5	2.3
TOTAL	**100.0**	**100.0**	**100.0**
STD. DEVIATION	**4.6**	**6.8**	**7.9**

n=80

The standard deviations in the last row of the table reflect the amount of dispersion in the importances for each column. As expected, the ACA importances show the least variation from the most important to least important attributes. This confirms similar findings which have shown ACA importances to be flatter than full-profile and choice-based results (Pinnell 1994, Huber 1992). Although the rank-order of attribute importance is identical for ACA and full-profile, the full-profile importances more closely line up with the importances derived from the traditional holdout choice exercise. This difference largely (if not principally) accounts for full-profile's edge in predicting the traditional holdouts over ACA, as will be shown below in Table 3.

The fact that full-profile closely matches the full-profile choice importances is not surprising given the similarities between the two full-profile tasks. Full-profile conjoint should have an advantage over ACA for predicting standard holdouts also shown in full-profile—especially if respondents adopt simplification heuristics.

When the ACA utilities are re-scaled at the individual level to full-profile importances, the hit rate for ACA approximates the hit rate for full-profile:

Table 3
Traditional Holdout Hit Rates
ACA Utilities Scaled to Full-profile Importances

ACA (before rescaling)	76.9%
ACA (after rescaling)	**82.6%**
Full-profile	82.4%

n=80

Scaling ACA utilities to full-profile importances significantly improves ACA's ability to predict the standard holdout choices (t=3.69).

SUPER HOLDOUT TASK RESULTS

For the Super Holdout Task, respondents were divided into groups of three individuals. Each group received a piece of paper showing four PCs described in full-profile. Over a 10-minute period, each group discussed the options and worked to consensus regarding the most preferred to least preferred PC for the proposed computer lab. After the group had come to consensus, respondents were asked to record their own personal judgements. Interestingly enough, most individuals did not change their answers from the group ranking. This could reflect respondent homogeneity, peer-influenced bias, or perhaps lack of dedication to the task.

To review, we hypothesized that the group exercise might better reflect the depth of processing and consideration that respondents would undertake if they were actually making the decision in the real world. It might also mimic the seeking and sharing of information that typically accompanies high involvement purchases. We personally observed most of the Super Holdout sessions, and in our opinion, respondents deliberated with a good deal of effort. However, we cannot know whether we really accomplished our goal of stimulating respondents to make more life-like judgements.

Table 4 shows predictive results for full-profile and ACA.

Table 4
Super Holdout Task Hit Rates

ACA	73.3%
Full-profile	76.7%

n=80

Full-profile does a better job predicting the Super Holdout Task than ACA, although the margin of victory is slightly less than for predicting the traditional holdout choices. The t-value of the difference in mean prediction is 1.25 (not significant).

Table 5 displays the result which was at the heart of our research: attribute importances for traditional holdout choices versus the Super Holdout Task.

Table 5
Attribute Importances for Holdout Choices

	Standard Choices	Super Holdout Choices
Internet Access	25.7	29.0
RAM	23.8	22.9
Price	18.1	16.2
Hard Drive	12.2	7.2
Warranty	8.4	6.5
Processor	9.5	9.4
Lab Assistants	2.3	8.8
TOTAL	**100.0**	**100.0**
STD. DEVIATION	**7.9**	**8.1**

n=80

The precision of estimates is greater for the Standard Choices than the Super Holdout Choices. Recall that each of the 80 respondents completed five standard choice tasks during the computerized survey, but only one Super Holdout Task. Regrettably, the least stable estimates are the most important to our research.

The ratio of the most to least important factor is 11:1 for Standard Choices, and 4:1 for Super Holdout Choices. This conforms to our hypothesis—but focusing only on the extreme points is quite susceptible to error, given the instability in the estimates for the Super Holdout Choices. The standard deviations suggest there is very little difference in the spread of importances between the standard holdouts and the Super Holdout Task.

There are several competing explanations for why we didn't observe significantly flatter results for the Super Holdout Task relative to the traditional holdouts:

1) Maybe people really don't display flatter importances in more carefully considered decisions.

2) Perhaps the students didn't take the Super Holdout Task as seriously as we had hoped.

3) Maybe they did take it seriously, and they also took the standard choice tasks as seriously.

We are more inclined to believe the second explanation. Until more evidence is shown, however, the main hypothesis of our paper remains unproven.

ANATOMY OF ACA IMPORTANCES

While we weren't able to find significant differences in attribute importances between the traditional holdouts and the Super Holdout Task in our pilot study, we confirmed that ACA importances tend to be "flatter" than full-profile importances.

ACA utilities are derived from two sources: the pairs and priors.

Priors: We used default settings for priors, including a 4-point scale for stated importances.

Pairs: We again used default settings. Respondents judged pairs on a nine-point graded scale. The design included eighteen total pairs. Twelve pairs were shown on two attributes, and six more pairs on three attributes.

Table 6 displays the importances for ACA as given earlier in Table 2, along with importances derived from just the priors. As before, the importances were computed from average utilities on attributes with assumed *a priori* order.

Table 6
ACA Importances by Component

	ACA Final Importances	ACA Priors
Internet Access	21.5	17.4
RAM	19.8	16.6
Price	15.0	15.5
Hard Drive	14.0	14.6
Warranty	11.6	13.0
Processor	10.7	12.9
Lab Assistants	7.4	10.0
TOTAL	**100.0**	**100.0**
STD. DEVIATION	**4.6**	**2.3**

n=80

The importances are less extreme in the priors than in the final utilities. Indeed, with a 4-point importance scale, even if all respondents were in agreement about the most and least important attributes, the maximum ratio between priors importances would be 4:1.

In Version 4 of ACA, optimal weights for the two components (priors and pairs) are fit to best predict purchase likelihood judgments in the calibration concepts, which for our study were customized to include the six most important attributes for each respondent. For this data set, we may infer that the pairs information is more extreme than both the priors and final optimally-weighted importances.

Critics of ACA have suggested that the 4-point stated importance scale is too coarse. In 1991, Bill McLauchlan reported results from an experiment which tested different scales for the stated importances for ACA priors (McLauchlan 1991). For that study, the 4-point implementation performed as well (predicting holdout concepts) as customized ACA versions using a 9-point scale and an analog version which accommodated up to 100 scale points. McLauchlan did not report importances for the attributes involved in his research, however, so we do not know if his study featured such extreme importances as our study.

It seems plausible that designs with greater extremes from the most important to least important attributes might benefit from more than four scale points in the priors section, but until further research is presented on this, it remains speculation. ACA permits the user to customize the scales used in the priors, so one could easily experiment in this area.

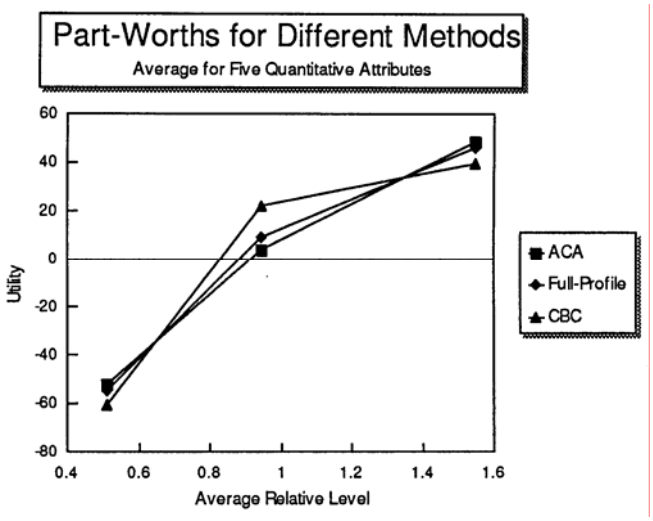
We should again emphasize that this was a pilot study with a small sample and atypical respondents. Lacking evidence about the *true* impact of these attributes on actual purchase decisions for a given product category, one really can't know whether average priors importances are really less valid than importances reflected in the pairs, or the final optimally-weighted result.

ATTRIBUTE PARTWORTHS

Attribute importances reflect utility differences between the best and worst levels for attributes. But importances ignore interior levels. The partworths for intermediate levels may differ between conjoint methods, and may also account for differences in predictability of holdout concepts. Most of the attributes in our study included a middle level, and five of these were quantitative in nature.

Huber has recently noted differences in partworths between Choice and traditional conjoint methods. He has found that CBC partworths tend to show more curvature than full-profile ratings-based conjoint and known utilities (Huber *et al*. 1997). On his advice we investigated this issue with our data set. The results are summarized in Figure 1.

Figure 1



The CBC partworths are based on first choices from the traditional choice tasks administered during the disk-based survey. We don't show partworths for the Super Holdout Task due to instability in the partworth estimates. The partworth utilities are zero-centered and scaled so that the difference between the best and worst levels is 100 points.

Partworths from ACA and full-profile are very similar, with the ACA utilities showing the least amount of curvature on average. This is not surprising given the linearity assumption from the priors. The CBC partworths displayed the most curvature for all five attributes. In some cases the curvature was quite pronounced, and the average across the five attributes reflects this trend.

Why do CBC partworths appear to display more curvature? Huber suggests that respondents may adopt a simplification heuristic that involves scanning choice sets for products with the worst level on key attributes (Huber *et al.* 1997). Respondents focus more on avoiding products with the least preferred levels rather than choosing products with enough good features to surpass some utility threshold, which biases the worst level downward. Under the scaling procedure for the data in Figure 1 which fixes the spread of the worst to best levels, this causes the difference in utility between the best and middle levels to narrow.

It is interesting to note that Johnson and Orme found a similar pattern of curvature when comparing CBC utilities derived only from second choices compared to first choices (Johnson and Orme, 1996). Perhaps the phenomenon which causes second choices to show more curvature than first choices is related to the process which influences CBC partworths to display more curvature than with ACA or full-profile ratings-based conjoint. This remains conjecture, however, and we look forward to more research on these issues. Since we do not know the true shape of the partworths for our study, we can only note the differences between methods without judging which method best predicts real world events.

Attribute Order Effects

It is no great secret that order effects occur in survey research whenever we present lists of items. Researchers have also reported strong attribute order effects for full-profile conjoint and choice-based conjoint (Johnson 1991, Chrzan 1994), but in general we don't see much attention paid to this in practice.

For one full-profile data set, Johnson reported that attribute order effects accounted for roughly 16 percent of the total error variance of conjoint predictions (Johnson 1991). Since it is not reasonable to expect that respondents encounter attributes in the real world in the same order as seen in conjoint tasks, it is natural that we should control for order effects when comparing the validity of ACA and full-profile judgements.

Due to its adaptive nature in the partial-profile pairs section and the ability to randomize attribute presentation in the priors, ACA should be immune to order effects. Based on the past evidence, we expected that order effects could impact the remaining three aspects of our design: full-profile conjoint, computer-administered holdout choices, and the Super Holdout Task.

For our small pilot study we did not find significant attribute order or task order effects. Enough compelling evidence exists from other studies to suggest that we probably would have discovered significant effects given more data points.

Conclusion and Suggestions for Future Research

This was a small pilot study to test an approach for improving holdout data and designing better conjoint validity studies. A number of caveats and limitations come to mind:

1) Sample size was small: only 80 respondents.

2) MBAs are not a very representative sample.

3) The nine attributes tested were described in very succinct statements. The respondents already had a high degree of familiarity with the attributes. In the real world, nine-attribute full-profile studies might not always be so manageable for respondents.

4) Not enough data points were collected to gauge whether importances derived from the Super Holdout Task were significantly different from those of traditional holdout choices.

5) Perhaps the Super Holdout Task we implemented failed to create a significantly different (and more realistic) experience than the traditional holdouts administered during the course of the survey.

We hope to see further research done in this area. **We cannot stress enough that the ideal validity study would include actual purchase as the holdout criterion.** The conjoint community thirsts for this type of research to be published. Indeed we might call this the holy grail of conjoint validation research. We encourage individuals who have the resources to conduct and publish a carefully designed research study with actual purchase choice as the validation criterion. However, just *one* well-done study with real world purchase data would still leave unanswered questions. The ideal conjoint method for predicting high involvement purchases such as computers or cars may not be ideal for predicting purchases for beverages or bubble gum.

In the absence of actual purchase choice, better validation exercises can be designed for comparing conjoint methods. We can imagine that Super Holdout Tasks could take on many forms—many of which could be more realistic and effective than that which we implemented in this pilot study. Regardless of form, the spirit of the task is to put respondents in the same frame of mind as the real world event and to more closely match the consideration and depth of processing as would be expended in the actual purchase decision. Along with the general design principles we've reviewed, we hope aspects of the Super Holdout Task will be used in methodological studies in the future.

Appendix A
Conjoint Attribute Levels

1) Compaq
2) Dell
3) Zeos

4) 90-day warranty
5) 1-yr warranty
6) 2-yr warranty

7) Pentium 100 MHz
8) Pentium 133 MHz
9) Pentium 166 MHz

10) 1 lab assistant
11) 2 lab assistants

14) 600 Mbyte hard drive
15) 1.2 Gbyte hard drive
16) 2 Gbyte hard drive

17) 8 Mb RAM
18) 16 Mb RAM
19) 32 Mb RAM

20) NO modem/Internet access
21) Modem and Internet access

22) $1,000[2]
23) $1,500
24) $2,000

[2] Respondents were told that their university had budgeted $30,000 for purchasing PCs. It was stressed that recommending a $2,000 PC would allow only 15 PCs to be purchased for the lab.

The attribute levels were described exactly the same in the ACA, full-profile card-sort and holdout concepts.

12) Ergonomic keyboard/mouse
13) Standard keyboard/mouse

Appendix B
Full-profile Card Sort Design

One of the challenges of full-profile designs is to keep the total number of stimuli to a reasonable number while still capturing enough information for stable individual-level utility estimation. This was particularly important for this study since respondents would complete ACA, full-profile and additional holdout tasks.

In a previous comparison of full-profile and ACA, Agarwal and Green (1989) used 18 cards to measure six attributes having three levels each. We used 22 cards in our design. With 22 cards, the number of cards to parameters ratio of 1.47 (22/15) is roughly equivalent to Agarwal and Green's ratio of 1.50 (18/12).

Sawtooth Software's CVA version 2 iterative designer was used to generate the design (shown below), which has a D-efficiency of 95.2% (Kuhfeld *et al*. 1994).

Card 1= 2 6 8 11 12 14 18 20 23
Card 2= 1 5 9 11 13 15 18 20 24
Card 3= 1 6 9 11 13 15 17 21 24
Card 4= 2 4 7 11 13 15 18 21 23
Card 5= 3 6 9 10 12 15 19 21 23
Card 6= 3 5 8 10 13 16 18 21 24
Card 7= 2 6 8 11 13 16 19 21 22
Card 8= 1 5 7 10 13 14 17 21 23
Card 9= 2 6 7 10 12 15 19 20 24
Card10= 1 5 8 10 12 15 19 20 22
Card11= 1 6 7 10 12 16 18 21 22
Card12= 3 4 9 11 12 14 18 21 22
Card13= 3 5 7 11 13 14 19 20 22
Card14= 3 6 7 11 12 16 17 20 24
Card15= 2 5 9 11 12 16 17 21 22
Card16= 2 4 9 10 13 16 19 20 24
Card17= 1 4 8 10 13 16 17 20 23
Card18= 2 5 8 10 12 14 17 21 24
Card19= 2 6 9 10 13 14 18 20 22
Card20= 1 5 9 11 12 16 19 20 23
Card21= 1 4 8 11 12 14 19 21 24
Card22= 3 4 8 10 12 15 17 20 22

Each concept was printed in hard-copy on 3 1/2" x 5" cards. Below is an example:

Card# 1
How likely would you be to recommend...

Dell
2-yr warranty
Pentium 133 MHz
2 lab assistants
Ergonomic keyboard/mouse
600 Mbyte hard drive
16 Mbyte RAM
NO modem/Internet access
$1,500

Write a number below between 0 and 100, where 0 = Definitely would NOT; 100 = definitely WOULD.

Your answer _____

Instructions on the survey disk asked respondents to sort the cards into two piles based on preference, and then to divide those two once again. After sorting the cards into four piles, respondents were instructed to write their evaluations on the cards. Then, the cards were shown on the computer screen one at a time, and respondents were asked to type the answers they wrote for the cards. Respondents were encouraged to modify their answers if they desired as they recorded them. Respondents were randomly given a set of either blue or yellow hard-copy cards for the full-profile task. The attribute rotations in the two versions were as follows:

Blue	**Yellow**
A) Brand	A) Brand
B) Warranty	E) Ergonomic Features
C) Processor	F) Hard Drive
D) Lab Assistants	G) RAM
E) Ergonomic Features	H) Internet Access
F) Hard Drive	B) Warranty
G) RAM	C) Processor
H) Internet Access	D) Lab Assistants
I) Price	I) Price

REFERENCES

Agarwal, Manoj K. and Paul E. Green (1989), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results," *International Journal of Research in Marketing*.

Chrzan, Keith (1994), "Three Kinds of Order Effects in Choice-Based Conjoint Analysis," *Marketing Letters,* 5:2, April, 165-72.

Huber, Joel, Dick R. Wittink, Richard Johnson, and Richard Miller (1992), "Learning Effects in Preference Tasks: Choice-Based Versus Standard Conjoint," *Sawtooth Software Conference Proceedings*, 275-82.

Huber, Joel, Dan Ariely, and Gregory Fischer (1997), "The Ability of People to Express Values with Choices, Matching and Ratings," Working Paper, Fuqua School of Business, Duke University.

Johnson, Richard M. (1989), "Assessing the Validity of Conjoint Analysis," *Sawtooth Software Conference Proceedings*, 273-80.

Johnson, Richard M. and Bryan K. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint?" ART Forum, Beaver Creek, Colorado, June.

Kuhfeld, Warren, Randall D. Tobias and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research,* (November), 545-57.

McLauchlan, William G. (1991), "Scaling Prior utilities in Sawtooth Software's Adaptive Conjoint Analysis," *Sawtooth Software Conference Proceedings*, 251-68.

Pinnell, Jonathan (1994), "Multistage Conjoint Methods to Measure Price Sensitivity," ART Forum, Beaver Creek, Colorado, June.

SOLVING THE NUMBER-OF-ATTRIBUTE-LEVELS PROBLEM IN CONJOINT ANALYSIS

Dick R. Wittink, William G. McLauchlan and P.B. Seetharaman[1]

ABSTRACT

Based on an experimental study, we find that the relative importances of attributes in CBC suffer from the same number-of-levels effect as do all ratings- or rank order-based conjoint methods. In a study of PCs this effect is larger for CBC than it is for ACA. We also show how the predicted shares that form the basis for market simulations are sensitive to this effect. To remedy the problem we introduce a customized version and we obtain superior improvement in share predictions for the customized ACA version relative to two versions of the traditional ACA method.

INTRODUCTION

One persistent problem that reduces the validity of conjoint results is the number-of-levels effect, originally identified by Currim et al. (1981). They found that the partworths for the best and worst levels of an attribute tend to be farther apart as the number of intermediate levels increases. This artificial phenomenon affects not only the partworths and the derived attribute importances, but potentially the preference share predictions produced in market simulations as well (Wittink and Krishnamurthi 1981).

In the latest survey of the commercial use of conjoint analysis (Vriens et al. 1997), conjoint users indicate substantial awareness of the number-of-levels effect. In North America, 74 percent (31/42) of ratings/rankings-based conjoint users indicated familiarity with the effect. Among European users, 63 percent (45/71) said they were familiar with it. However, of those familiar, the majority of respondents said they do not make adjustments in the design of a study to counter the problem. We propose, therefore, to show how the problem can be avoided in ACA, which is by far the preferred method among commercial users in North America and in Europe (Vriens et al. 1997).

Many experimental studies have been conducted to determine the magnitude of the number-of-levels effect under a variety of data collection methods, measurement scales, estimation methods, etc. (e.g., Wittink et al. 1982, Wittink et al. 1989, Wittink et al. 1992, Steenkamp and Wittink 1994). Although the effect always obtains, there is disagreement about its origin. Green and Srinivasan (1990) argue for a behavioral response explanation, i.e., respondents may be sensitive in their preference judgments to the number of levels used for each attribute. However, for rank-

[1] Dick R. Wittink is the Henrietta Johnson Louis Professor of Management, and Professor of Marketing and Quantitative Methods, Johnson Graduate School of Management, Cornell University; William G. McLauchlan is Principal, McLauchlan & Associates, Cincinnati; P.B. Seetharaman is a doctoral candidate in marketing, Johnson Graduate School of Management, Cornell University. The authors thank the Marketing Science Institute for financial support, Richard M. Johnson for many intensive discussions and software support, and Christopher King for software support.

order preference judgments, it is possible to derive the effect mathematically (Wittink et al. 1989).

At the 1992 Sawtooth Software Conference, Wittink et al. (1992) reported the results of an experimental study in which alternative explanations for the effect were considered. They found that two manipulations designed to reduce the magnitude of the level effect, if behavioral explanations matter, failed to generate supportive results. The only study we know of in which direct evidence of a behavioral cause is Johnson (1991) who focused on the price respondents would be willing to pay for improved products. In all other studies, it is impossible to rule out alternative explanations. Even if the occurrence of the effect can be described in elaborate detail, as has been done for ACA (Wittink et al. 1991), researchers still can disagree about its interpretation. Nevertheless, this detailed description provides an opportunity to modify ACA so as to eliminate the effect.

THE LEVEL EFFECT IN ACA

It is well known that ACA combines self-explicated data with preference-intensity judgments in the generation of idiosyncratic partworths (Sawtooth Software 1993). In a comparative experimental study, Wittink et al. (1991) found the magnitude of the number-of-levels effect to be approximately twice as large for full profile as for ACA. However, they also found that the initial ACA solution, based on the self-explicated data, did not show a level effect. They did find that in ACA the effect comes about in the preference-intensity judgments. And for respondents for whom paired objects happened to be approximately "utility balanced" (utility balance exists if the paired objects have equal predicted utilities based on the initial ACA solution), the level effect was found to be close to zero.

Specifically, an analysis of the utilities predicted from the initial ACA solution for paired objects showed that the difference is often far from zero. In a study of refrigerator preferences, Wittink et al. (1991) find that this predicted difference is non-zero 73 percent of the time. They also observe that for any non-zero predicted difference, respondents' average preference-intensity judgment is between the predicted difference and zero. This means that on average the attribute that accounts for the largest part of the predicted utility difference between the paired objects will see its levels' partworths move toward zero (and the opposite change occurs for the other attribute, when the objects are defined on two attributes). It turns out that, more often than not, this updating favors attributes with more levels (i.e., attributes defined on a relatively large number of levels tend to see the partworths move farther apart). And the higher the imbalance in predicted utilities, the stronger this effect, on average.

Such an analysis of how the number-of-levels effect comes about in ACA indicates that the problem can be circumvented by having all paired objects equal in predicted utilities. That is, the "mechanical" explanation of the phenomenon in ACA (Wittink et al. 1991) depends on an imbalance in predicted utilities for paired objects. With "utility balanced" paired objects, any remaining number-of-levels effect would have to be due to a behavioral cause (to be explored in the future). It turns out that there are many pairs of objects that satisfy the utility balance constraint if the number of levels for each attribute equals the attribute's self-explicated importance (on a four-point scale) plus one. We use, therefore, a customized ACA version, created by Rich Johnson and Chris King that accomplishes this objective.

The idea of customizing the number of levels to the respondent's self-explicated importances is also appealing for another reason. Once the range of variation for an attribute is defined, it seems intuitively attractive to use the following principle: the greater the importance of the difference between the best and worst levels of an attribute, the larger the number of intermediate levels. Thus, for a respondent who rates this importance = 1, only the extreme levels are used. But for a respondent who rates the importance = 4, we use three intermediate levels in addition to the extreme levels. This customization makes sense because it allows the conjoint user to obtain more refined utility functions of each respondent's more important attributes.

To illustrate the idea of utility balance in paired objects, we present two scenarios. In the first scenario, we consider pairing objects defined on attribute A with four levels but a (self-explicated) importance = 1 and attribute B with two levels but an importance = 3. We show in the matrix below all predicted utilities.

		A			
		+0.5	+0.17	-0.17	-0.5
B	+1.5	2	1.67	1.33	1
	-1.5	-1	-1.33	-1.67	-2

It is clear from this matrix that there is no pair of objects available with utility balance. Indeed, the smallest difference in predicted utilities is 2. Thus, if just these two attributes are used for the construction of paired objects in ACA, a large amount of utility imbalance is unavoidable.

In the second scenario, we employ the customization principle. Since A has an importance = 1, it will have two levels. And B will have four levels, given an importance = 3.

		B			
		+1.5	+0.5	-0.5	-1.5
A	+0.5	2	1	0	-1
	-0.5	1	0	-1	-2

It is clear from this second matrix that there are three pairs of objects with equal predicted utilities. In general, it is easy to show that utility balance exists for at least one pair of objects as long as the number of levels is customized in this manner.

ALTERNATIVE SOLUTIONS TO THE PROBLEM

Before we discuss the details of the experimental study to test the customized ACA version, we briefly discuss alternative solutions to the number-of-levels effect.

1. **Adjust the results** based on maximum and minimum possible partworths or attribute importances. Currim et al. (1981) made such a correction and showed that the substantive conclusions changed dramatically. However, this can only be done if **rank-order** preferences are collected.

2. Use an **equal number of levels** for all attributes. Many recently published academic studies have this property. In commercial applications involving a mixture of continuous (e.g., price) and discrete (e.g., a feature that is present or not) attributes, this is impractical. In addition, it does not satisfy the "utility balance" principle that we employ in the customized ACA.

3. Collect preference data on **other response scales**. Steenkamp and Wittink (1994) used magnitude scaling and obtained a reduction in the magnitude of the level effect under some conditions. However, the level effect still occurs when a theoretically superior measurement scale is used.

4. Collect **choice data**. Conjoint studies that ask respondents to choose one object out of several objects for multiple choice sets, defined according to experimental design principles are gaining in popularity. To date, there is no evidence that choice-based conjoint results are subject to the number-of-levels effect.

5. Use only **self-explicated data**. In ACA, when the self-explicated importances are elicited, the respondent only sees the best and worst levels if the preference order for the attribute levels can be assumed a priori. In that case, the self-explicated importances cannot show a number-of-levels effect. However, if respondents first provide a preference order of an attribute's levels, the self-explicated importance may be sensitive to the number of levels. To date, there is no evidence that the initial ACA solution is sensitive to the number of levels.

EXPERIMENTAL DESIGN

Our discussion of the alternative solutions indicates that alternatives 1-3 have potentially unacceptable characteristics. The use of either of the last two alternatives would imply the rejection of ACA, currently the most popular conjoint method. We prefer, therefore, to test a modified version of ACA (with customized numbers of levels). Due to the absence of evidence regarding the occurrence of a number-of-levels effect in choice-based conjoint, we also include CBC (Sawtooth Software, 1994) in our study.

We have chosen the personal computer as the product category and we employ a maximum of five levels for each of the following five attributes:

Attribute	Levels
- Brand name	Compaq; Dell; Gateway; IBM; Packard Bell
- Speed	200; 175; 166; 150; 133 mhz
- Hard Drive	2.0; 1.8; 1.6; 1.4; 1.2 GB
- RAM	64; 48; 32; 24; 16 MB
- Price	$1,200; $1,400; $1,600; $1,800; $2,000

For all attributes except brand name, we assume that all respondents have preferences for the levels according to the order stated above. Only for brand name is this order elicited. In the customized version, the specific levels used for an attribute are chosen as follows:

Self-Explicated Importance	Number of Levels	Specific Levels
4	5	1, 2, 3, 4, 5
3	4	1, 2, 4, 5
2	3	1, 3, 5
1	2	1, 5

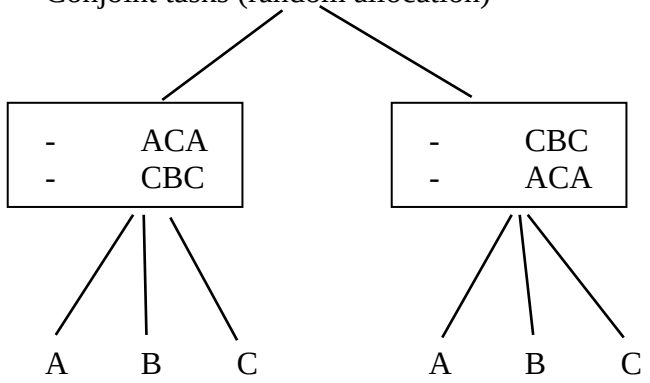
Thus, for each attribute, the best and worst levels are always used. We note that the use of four levels is most likely to produce substantial nonlinearity in the partworths. For example, price would then have $1,200; $1,400; $1,800; and $2,000 as specific options. The linearity imposed on the initial ACA solution is then perhaps especially inappropriate, and the question is whether the paired objects will allow the partworths to be sufficiently updated (to be explored in the future).

Three versions. We use three conjoint design versions. Version A is the customized version. Respondents receiving version A complete both ACA and CBC with customized numbers of levels. Version B employs the 'equal number of levels' principle (alternative 2 described above). Respondents receiving version B do both ACA and CBA with all five attributes defined on three levels: the best, the worst and the median preferred level. Version C uses three levels for Brand Name (the specific levels for Brand Name depend on the respondent's preference order, as in versions A and B), but two levels for Speed and RAM, and four levels for Hard Drive and Price. Respondents receiving version C also complete both ACA and CBC with specific levels chosen based on the number of levels, consistent with the scheme used in the customized version (except for the linkage to self-explicated importances).

Respondents were recruited in shopping malls in five metropolitan areas. They were screened on intention to purchase a PC in the next six months. Each respondent received $5 for the completion of the task. Two hundred respondents completed each of the versions. Details of the procedure and the design are shown below.

Data Collection

- Self-explicated data
  (preference order for brand names, importances for all attributes)
- 5 Holdout choice sets
- Conjoint tasks (random allocation)



- 10 Holdout choice sets (5 replicates)
- Personal data (PC ownership, PC attributes if known)

All holdout choice sets consist of two PCs defined on all five attributes. These PCs do not vary in brand name (only each respondent's preferred brand name is used). For the other four attributes, only the extreme (best, worst) levels are used, because only those levels are necessarily included in all designs.

HYPOTHESES

We expect to observe the number-of-levels effect between versions B and C. For ACA, this is simply a confirmation of earlier results. However, for CBC, this effect has not yet been demonstrated.

H1a. In ACA, Hard Drive and Price attain greater distances between the partworths for the best and worst levels in version C (four levels) than in version B (three levels), while Speed and RAM attain greater distances in version B (three levels) than in version C (two levels).

H1b. Same as H1a for CBC.

If we restricted ourselves to the manipulations involving versions B and C, we would not be able to say to what extent the effect is due to the mechanical explanation provided in Wittink et al. (1991) and/or to a behavioral phenomenon. Since the mechanical explanation does not apply in version A, we can use the version A results to find evidence in favor of a behavioral explanation. Specifically, for a behavioral explanation to matter, we should observe that respondents disproportionately favor the objects in the preference-intensity section of ACA that are superior on the attribute with more levels.

H2. The preference-intensity judgment in ACA for the object favored on an attribute with more levels is a positive function of the (positive) difference in the number of levels between the two attributes (to be explored in the future).

We note that respondents do not have immediate awareness of the numbers of levels as the preference-intensity judgment section in ACA begins. Instead, their sensitivity to the amount of variation indicated by the number of levels may increase throughout this section. Thus, we will allow for an interaction between the predictor defined in H2 and the sequence number of paired objects. We will also consider the relevance of covariates. For example, a behavioral effect may be less likely to occur for respondents with a high degree of PC expertise.

We claim that by definition the mechanical explanation for the number-of-levels effect does not apply to the customized version (A) in ACA. Everything else being equal, the version A results should then have higher validity than either of the other versions. We use the holdout choices to test this. This requires that the predictions are also subject to the number-of-levels effect. Specifically, we expect that the predicted shares for a given object are higher in version C than in version B when the object is favored on an attribute with more levels or disfavored on an attribute with fewer levels.

H3. The ACA-based predictions of a holdout profile differ systematically between versions B and C based on the (positive) difference in the number of levels between the attribute(s) on which the profile is favored and on the (negative) difference in the number of levels between the attribute(s) on which the profile is disfavored.

If H3 verifies the existence of a number-of-levels effect on predicted shares, we can then propose superior predictive validity for version A. Between versions B and C, we expect superior predictions for version B. This expectation is partly based on the fact that having all attributes with the same number of levels is often considered a solution to the number-of-levels problem.

H4. Version A provides the best predictive validity results, followed by version B.

We note that the customization of the numbers of attribute levels has a logical basis in ACA. However, we have not provided any arguments why the same principle should apply to CBC. Indeed, at this point, we cannot provide such arguments. Rather, given that we examine the existence of a number-of-levels effect for choice-based conjoint (version B versus C), it seemed merely prudent and efficient to also use a customized version of CBC (the CBC results will be explored in the future).

RESULTS

We obtained complete data for 182 respondents in version A, 203 in version B and 204 in version C. The individualized ACA partworths were averaged across all respondents, independent of the order of methods, within each version. The absolute distances between the partworths for the extreme levels in versions B and C are reported in Table 1.

Table 1
Distances Between Average Partworths for Extreme Levels
(ACA)

Attribute	Version B	# of levels	Version C	# of levels	Difference
Speed	0.60	3	0.44	2	+0.16
Hard drive	0.53	3	0.62	4	-0.09
RAM	0.68	3	0.57	2	+0.11
Price	0.45	3	0.44	4	(+0.01)

It is clear from Table 1 that the distance is much greater when the number of levels is larger for each attribute, except for Price. The direction of the difference is consistent with H1a, except for Price. Across the four attributes, the average difference (in the expected direction) is 0.09. Note that in version B RAM has the largest distance while in version C it is Hard drive. In addition, Speed has the second largest distance in version B while it is tied for last in version C. Thus, substantive conclusions based on these numbers differ strongly between the two versions.

For CBC we also used individualized partworths (provided by Rich Johnson based on a newly developed program) and averaged those in the same manner as described above for ACA. The absolute distances are shown in Table 2.

Table 2
Distances Between Average Partworths for Extreme Levels

(CBC)

Attribute	Version B	# of levels	Version C	# of levels	Difference
Speed	1.38	3	1.10	2	+0.28
Hard drive	0.90	3	1.68	4	-0.78
RAM	1.76	3	2.05	2	(-0.29)
Price	1.54	3	2.17	4	-0.63

The CBC results also show a larger distance between the extreme levels' partworths when the number of levels is greater, for three of the four attributes. These results are consistent with H1b. For Hard drive and Price the difference (comparing 3 with 4 levels) is on average -0.70, while for Speed and RAM the difference (comparing 3 with 2 levels) is on average zero.

We note that the ACA results show the larger difference for Speed and RAM (average +0.14) while the CBC results show the strongest effect for Hard drive and Price. This comparison of ACA and CBC results both across versions and across methods is complicated by the difference in scale values (across methods). To enhance the comparability we next focus on relative importances. We show in Table 3 these importances based on the average partworths, for ACA and CBC (and also show the average of the relative importances calculated for each respondent separately). For both methods, all observed differences are now consistent with the hypotheses. And for both methods the greatest difference occurs for Speed and Hard drive.

Table 3
Relative Importances Based on Average Partworths by Method, Version, and Attribute
(Average Relative Importances from Individual Partworths in Parentheses)

Method: ACA

Attribute	Version B	# of levels	Version C	# of levels	Difference
Speed	27% (25)	3	21% (19)	2	+6% (+6)
Hard drive	23% (25)	3	30% (31)	4	-7% (-7)
RAM	30% (27)	3	28% (23)	2	+2% (+4)
Price	20% (23)	3	21% (27)	4	-1% (-4)

average absolute difference = 4%
(5)

Method: CBC

Attribute	Version B	# of levels	Version C	# of levels	Difference
Speed	25% (24)	3	16% (10)	2	+ 9% (+ 8)
Hard drive	16% (17)	3	24% (31)	4	- 8% (-14)
RAM	32% (33)	3	29% (27)	2	+ 3% (+ 6)
Price	28% (26)	3	31% (26)	4	- 3% (0)
					average absolute difference = 6% (7)

The last column in Table 3 shows the difference in the relative importances between versions B and C. On average, for ACA the relative importances based on average partworths differ by 4 percentage points (in the expected direction), while the average of the individually computed relative importances differs by 5 percentage points. The larger difference for the second measure is due to the fact that unreliability in the individualized partworths contributes to the number-of-levels effect (and this unreliability is eliminated when the partworths are first averaged).

It is also noteworthy that on both measures CBC shows a higher number-of-levels effect in the difference column. For CBC the relative importances from the average partworths differ by 6 percentage points (in the expected direction), while the average of the individually computed relative importances differs by 7 percentage points. In that sense CBC is more like "full profile" which also was found to generate a larger number-of-levels effect (Wittink et al. 1992b) than ACA did.

Predictions. For each of the ten holdout sets we show the predicted shares, based on utilities predicted from the individual ACA partworths, in Table 4. The data from the replicates of five holdout sets, obtained prior to the ACA and CBC tasks, are ignored. Each object in a holdout set is defined in terms of the best or worst levels of four attributes: Speed, Hard drive, RAM, and Price, in that order (all objects are the same on Brand, which is always defined as the respondent's preferred level). Between versions B and C we expect that if the left object has the best level on an attribute with more levels and/or the worst level on an attribute with fewer levels, this object will have a higher predicted choice share. For example, in the first holdout set the left object differs from the right object on the first two attributes. On the first attribute the left object has the best level (and it has 3 levels in B versus 2 in C) while on the second it has the worst (where it has 3 levels in B but 4 in C). Due to the number-of-levels effect, we expect B to have a higher predicted share than C, as is the case. Interestingly, for all holdout sets the difference in the last column is in the expected direction, consistent with H3. We note that the standard error of this difference is 5%. Thus, almost all the observed differences are statistically significantly different from zero. On average, the **absolute** difference in the last column is 11.3%.

Table 4
Predicted Choice Shares for Left Object
(based on individual ACA partworths)

Holdout Set[2]		Version A	Version B	Version C	Difference (B-C)
1211	2111	51%	51%	37%	14%
1121	1112	33%	42%	45%	-3%
1212	2121	61%	61%	45%	16%
2111	1112	44%	42%	51%	-9%
1211	1121	67%	64%	45%	19%
2121	1112	17%	16%	25%	-9%
1212	2111	26%	25%	13%	12%
2121	1211	15%	15%	25%	-10%
1121	1212	62%	73%	82%	-9%
1222	2121	26%	25%	13%	12%

Thus, the predicted choice shares show strong sensitivity to the number-of-levels effect, consistent with H3. We note that the absolute difference between versions A and B is on average only 3% (while it is 14% between versions A and C).

To quantify the contribution each attribute makes to the observed difference in predicted choice shares between versions B and C we regressed the difference (last column in Table 4) against four predictor variables. Each predictor is defined equal to 1 if the left object has the best level and the right has the worst, 0 if the two objects both have the best or the worst level, and -1 if the left object has the worst and the right object has the best level.

The result is $\hat{Diff} = 2.5 + 4.7\ X_1 - 4.9\ X_2 + 5.5\ X_3 - 1.0\ X_4$. All the coefficients have the expected sign (for X_1 and X_3 version B has more levels while for X_2 and X_4 version C has more levels), consistent with H3. The R^2 is 0.93 which is highly significant ($p < .01$). Based on this equation we predict the difference for the third holdout pair in Table 4 (1212 vs 2121) to be 18.6% (the actual difference is 16%). This is the maximum possible difference that is captured by the equation.

Having demonstrated the influence of the number-of-levels effect on predicted shares, we now consider the final hypothesis. We again use the individualized ACA partworths and compare the predicted shares shown in Table 4 against the actual shares. However, since the initial ACA solution may have differential predictive validities between the three versions we want to consider the improvement in prediction relative to the initial ACA result. In addition, we consider the quality of the improvement in prediction relative to the reliability of the holdout choices.

[2] We use 1 to indicate best and 2 for the worst level of the attributes Speed, Hard drive, RAM, and Price, in that order.

We show in Table 5 the absolute difference between actual and predicted shares based on the individual final ACA partworths, averaged over the ten holdout choice sets. Version A has a slight edge (9.4%) over version B (10.5%) while version C has a relatively poor result (15.7%). The next line shows the error in shares based on the initial ACA partworths, averaged over the same ten holdout choice sets. Here version A is also best, followed by version C. Below these numbers we show the equivalent absolute percentage errors averaged across the five holdout sets on which we have replicates. Here the error is smallest for version B, followed by version C.

Table 5
|Actual minus Predicted Share| for final ACA partworths, initial ACA partworths and for replicate pairs by version

	|Actual Share - Predicted Share|		
	Version A	**Version B**	**Version C**
Final ACA partworths	9.4% (11.0%)[3]	10.5% (11.6%)	15.7% (13.4%)
Initial ACA partworths	21.2% (21.2%)	25.9% (23.4%)	23.3% (22.0%)
Replicate pairs	6.8%	3.4%	5.0%

To obtain a comparable statistic we define the percent improvement (initial minus final ACA solution) relative to the difference between the initial ACA solution and the replicate pair. For version A we obtain (21.2 - 9.4)/(21.2 - 6.8) = 82 percent, while for version B we get 68 percent, and for version C it is 42 percent. These differences are consistent with H4. However, since the error for replicate pairs is based on only five of the holdout choice sets, we repeat this analysis by using the corresponding errors in shares for these five choices (the numbers in parentheses). For version A we now obtain (21.2 - 11.0)/(21.2 - 6.8) = 71 percent. For version B it is 52 percent, and for version C it is 47 percent. The differences are still consistent with H4.

CONCLUSION

We have obtained evidence that the number-of levels effect also exists in choice-based conjoint results. In fact, the magnitude of this effect quantified in terms of relative importances is higher for CBC than it is for ACA. We also show how predicted holdout choice shares based on individualized ACA partworths are affected. Versions B and C differ on average by 11.3 absolute percentage points in predicted shares, and for each holdout pair the difference is in the expected direction. Given the strong reliance on market simulations in commercial applications this linkage of the number-of-levels effect to predicted shares is critical.

[3] We show in parentheses the absolute difference between actual and predicted shares averaged over the five holdout choice sets on which the replicate pairs are defined.

All current conjoint approaches suffer from the number-of-levels effect problem. We claim that the customized version of ACA does not suffer from the "mechanical" explanation for the effect offered by Wittink et al. (1991). We hope to show later that this customized version is also free from an effect that has a behavioral or psychological basis.

We also find that the customized ACA version provides better predictions than either of the traditional versions. The predictive performance calculation takes into account both the reliability of the holdout choices and the performance of the initial ACA solution.

Several questions remain to be addressed. We need to determine the predictive validity of the three CBC versions. We also want to examine the predictive performance for both ACA and CBC at the individual level (e.g. hit rates). By using individual-level predictions we can also take individual differences into account.

Another issue that needs attention is that the traditional ACA version is "level balanced" while the customized version is "utility balanced." Level balance means that the paired comparisons in the preference intensity section are chosen to satisfy as best as possible the objective that all levels of an attribute are used equally frequently. By favoring equal predicted utilities in the customized version it is possible that the level frequencies are quite unbalanced at least for some respondents. Since all holdout pairs are defined only on extreme levels, it is of particular interest to determine how the predictive performance at the individual level depends on the frequency with which the partworths of these extreme levels are updated.

REFERENCES

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

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

Johnson, Richard M. (1991), "Comment on "Attribute Level Effects Revisited . . .," *Advanced Research Techniques Forum, Proceedings of Second Conference*, Rene Mora (ed.). Chicago: AMA, 62-4.

Sawtooth Software (1993), "Adaptive Conjoint Analysis," Sun Valley.

Sawtooth Software (1994), "Choice Based Conjoint," Sun Valley.

Steenkamp, Jan-Benedict E.M. and Dick R. Wittink (1994), "The Metric Quality of Full-Profile Judgments and the Number-of-Attribute-Levels Effect in Conjoint Analysis," *International Journal of Research in Marketing*, 11 (June), 275-86.

Vriens, Marco, Joel Huber, and Dick R. Wittink (1997), "The Commercial Use of Conjoint in North America and Europe: Preferences, Choices, and Self-Explicated Data," working paper in preparation.

Wittink, Dick R., Lakshman Krishnamurthi, and Julia B. Nutter (1982), "Comparing Derived Importance Weights Across Attributes," *Journal of Consumer Research*, 8 (March), 471-4.

Wittink, Dick R., Lakshman Krishnamurthi, and David J. Reibstein (1989), "The Effect of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters*, 1 (Number 2), 113-23.

Wittink, Dick R., Joel Huber, John A. Fiedler, and Richard L. Miller (1991), "Attribute Level Effects in Conjoint Revisited: ACA versus Full Profile," *Advanced Research Techniques Forum, Proceedings of Second Conference*, Rene Mora (ed.) Chicago: AMA, 51-61.

Wittink, Dick R., Joel Huber, Richard M. Johnson, and Peter Zandan (1992), "The Number of Levels Effect in Conjoint: Where Does It Come From, and Can It Be Eliminated?" *Sawtooth Software Conference Proceedings*, Margo Metegrano (ed.) Sun Valley, 355-64.

COMMENT ON WITTINK, MCLAUGHLAN, AND SEETHARAMAN

Rainer Paffrath
Simon, Kucher & Partners

Congratulations to you and your co-author's work!

The Number-of-Attribute-Levels Problem causes according to Mr. Wittink's and his co-author's results and their former research a serious bias in Conjoint analysis. Depending on the number of levels different results (relative importances, predicted shares) can be achieved. If one wanted to attach special importance to an attribute, he could choose 5 levels for this attribute.

Mr. Wittink and his numerous co-authors have dealt with this problem for a long time. If you read carefully the references to his paper, his work in the 80s can be characterized with the *identification* and *explanation* of the phenomenon. In the 90s he has worked towards a *solution*. The title of today's paper lays claim to solve the Number-of-Attribute-Levels Problem and focuses on the "mechanical" part of the problem. Of course, I like Mr. Wittink's work as it is an *active* research result, rather than a passive, criticizing research.

The main results are:

- The Number-of-Attribute-Levels Problem is pervasive (ACA, CBC, full profile)
- Predicted shares in simulation models will also be affected
- The core of Mr. Wittink's work is the customized version of ACA which seems to enhance predictive validity.

A customized ACA-version means a move towards more individually composed interviews. In my opinion there is a clear tendency towards this kind of interviewing. If you listened to my talk, I claimed to take into account the individual importance structure and to compose individual evoked sets. With the help of the microcomputers we have the capacity to do this.

There are two points I would like to say about the customized version of ACA. This experimental version lets the number of levels depend on its self-explicated importance. ACA uses this importance ratings section to initialize a solution which is then refined with the answers to the paired comparisons. If one uses this section to decide on the number of attributes, the importance of this section will increase. In my opinion this is problematic because the contribution of direct questions should be avoided. I know from experience that respondents tend to overvalue attributes and tend to assign high importance ratings to each attribute without differentiating the importance.

The other point is that we do not know how people will react to utility-balanced pairs. At a glance it seems that pairs will not get easier for respondents.

This does not mean I'm against Mr. Wittink's solution. In my opinion Mr. Wittink's work is very valuable because the Number-of-Attribute-Levels effect can be huge and can be abused. The solution shows potential to enhance predictive validity.

I'm sure at the end of the day we will have a version "5" of ACA which will be the most "customizable" version we have ever seen!

WHAT WE HAVE LEARNED FROM 20 YEARS OF CONJOINT RESEARCH: WHEN TO USE SELF-EXPLICATED, GRADED PAIRS, FULL PROFILES OR CHOICE EXPERIMENTS.

Joel Huber
Duke University

Conjoint analysis as a commercial method has been available for more than 20 years (Green, Wind and Jain 1972). Ten years ago, 1987, at a Sawtooth conference, I raised the question of how conjoint could work so well, given obvious differences between the task and market choice. I proposed that conjoint works primarily because the simplification in a conjoint task mirrors the simplification in the marketplace. That is, the complexity of the marketplace encourages people to make choices based on relatively few attributes, effectively selecting the attributes that they will attend to and value in a given choice. In the same way, the complexity of a full-profile conjoint task also encourages respondents to attend to a subset of the attributes. Thus conjoint works by simulating the *attribute selections* process that occurs in actual choices.

Much has changed in the last decade. The biggest change is the expanded arsenal of methods that enable us to measure people's values (Green and Srinivasan 1990). There are sophisticated self-explicated methods that break down choice into values for levels of attributes, weights for different attributes, and perceptions of alternatives on those attributes. ACA in its new version permits different kinds of self-explicated assessments and allows various scales to be used. Full-profile can be built with more general choice designs (Kuhfeld, Tobias and Garratt 1994). Perhaps the biggest change has been the availability of easy-to-use choice-based conjoint systems, which measure tradeoffs with responses to hypothetical choice sets. The question for the next millennium is not whether conjoint works, but defining the contexts in which different methods are appropriate.

One fact is irrefutable: different methods, and even the same method in different contexts, give different results. Consider the following three examples:

1. On high technology products such as computers, ACA's price partworth for knowledgeable respondents needs to be doubled to accurately predict subsequent choice-based conjoint (Pinnell 1994). That is, choices among brands are best predicted if the raw ACA utilities are doubly weighted before being entered into a choice simulator.

2. The relative value of price over brand doubles from the early choice tasks compared to later ones (Johnson and Orme 1996).

3. In contrast to ACA and full-profile conjoint, choices reflect 20%-30% greater emphasis on the most important attributes and put 30%-40% more weight on the least preferred levels of each attribute. (Zwerina and Huber 1997, Orme, Alpert and Christensen 1997, Huber, Ariely and Fischer 1997).

These results have been profoundly disquieting to me, and should be to you. They make clear that the myth of measuring a person's true utility structure is just that, a myth. However, in a positive sense they also suggest that our goal of formulating one way to best measure values needs to be replaced with a goal of matching the characteristics of decisions in the marketplace with those of the task. The three results above are not anomalies. I propose that we can understand the differences among techniques exhibited by examining three characteristics of methods to measure values: the attention that they place on various attributes, the way they alter competitive expectations, and, most importantly on the kinds of values they promote. Depending on the market being simulated, different methods may be appropriate. The purpose of this paper is to provide some guidelines to help you make such a matching.

It is important for me to acknowledge a strong, and somewhat controversial belief that I bring to this discussion. I believe, and hope to convince you, that the purpose of the typical trade-off study ought *not* to be the prediction of short-run market behavior. Short-run behavior is both quite predictable and of minimal strategic value. If we want to know what people will choose, the best predictor is what they chose last time. Part of the reason market behavior exhibits so much inertia is that most market choices take very little time. Even when time is taken, decisions are made using heuristics that permit reasonable choices despite poor comparative information and relatively little effort on the part of consumers. Instead of asking what a heuristic laden and opportunistic customer would buy, I believe that companies need to know:

1. What customers will choose if they *attend* to the attributes.
2. What they will do if customers' competitive *expectations* change.
3. What customers will do if they think about how much they *value* the attribute.

Put differently, if a company needs to know what customers do now, that is best approximated through econometric analysis of current sales data. For the majority of problems, short-term prediction is less important than knowing what a customer would do if and when the competitive environment changes. What happens if a price is lowered and customers gradually notice? What happens if the competitors promote a new feature? What happens if a consumer magazine makes side-by-side comparisons of different competitors? The sections below detail three ways in which various tradeoff tasks alter these three characteristics of the simulated purchase experience: they increase attention to displayed attributes, they alter expectations about the competitive offerings, differentially encouraging the evaluation of various attributes.

THREE PROPERTIES OF EVALUATION TASKS

Increase Attention. Evaluation tasks intentionally force respondents to attend to attributes that they might otherwise not notice. In doing so, attention can elevate the importance of particular attributes to a level that is greater than would occur in the marketplace. For example, featuring the attribute "surge protector" may make this attribute salient even though it may not be salient in actual choices.

Drawing on the ideas given above, this lack of correspondence to the market may be seen as an advantage rather than a disadvantage. It certainly suggests that measured value of an attribute should be viewed as *conditional* on the respondent noticing the attribute. Being conditional on attention is advantageous since in most markets important but unnoticed attributes tend to

become noticed over time. Consumers in markets may be slow, but they are not stupid. If a product has better features or is lower priced, it will be noticed eventually, initially by vigilant consumers, and thereafter through word-of-mouth, rating services and by retailers. Finally, if a product feature needs to be noticed to affect choice, then advertising or promotion of that superior feature is a relatively simple task. The important point is that by forcing attention on specific features or prices, various evaluation tasks provide a prediction of marketplace behavior as customers become aware of those attributes.

Alter Competitive Expectations. An elaborate set of beliefs assist us in our choices. Two kinds are relevant here. The first involve reference levels within attributes. For example, we have price and performance expectations among competitive offerings, enabling us to spot an important new feature or a price that is out-of-bounds. The second kind of expectation involves associations between attributes, as when one uses one attribute to draw inferences about other attributes. These beliefs also help us to simplify choice by using one attribute as a proxy for others.

Research has shown that these expectations are important but fragile. Once successfully challenged in the market, then the new expectations alter the way we process information and make choices. To the extent that measurement tasks also break down beliefs, they simulate what is likely to happen if competitive conditions change. In doing so they assess in a short period of time changes in behavior that the market may take longer to accomplish.

Encourage Evaluation. All trade-off methods, to a greater or lesser extent, encourage respondents to think about the meaning of an attribute in terms of their lives. The primary mechanism generating these thoughts is the trade-off task itself, which encourages respondents to think about the value of one attribute compared with another. Notice that the type of task can encourage or discourage this evaluation. For example, by loading the alternatives with features, a number of attributes may never be noticed; alternatively, by loading an attribute with a large number of levels, it can draw attention to itself. (Wittink, McLauchlan and Seetharaman 1997).

Some studies only gauge reaction to the idea of a feature, while others ask respondents to use the different versions. These latter studies simulate the effect of trial use and evaluation of the feature. The point here is that the degree and depth of the evaluation is part of the study design. In the next section we will explore how particular tasks differentially elicit values.

Differences among Four Tasks: Self-Explicated, Pair Comparisons, Full Profile and Choices

We examine four tasks commonly used to measure values: self-explicated methods, graded pair comparisons, full profile ratings and choices. We initially explore simple and somewhat stylized versions of each method, examining the ways they focus attention, alter competitive expectations and ultimately transform the values generated. After having considered these standard forms, we will then discuss how varying implementations of the tasks further alter the values generated by each method.

Self-Explicated Methods. Self-explicated methods typically involve two data collection stages. First, the respondent provides the relative value of levels within each attribute. For example, ACA's default asks for a ranking of the levels of the attributes, while a different method

might assign 100 points to the most preferred level and corresponding values to other levels. Then the self-explicated model needs to determine a weight for each attribute. ACA's default accomplishes this task with a 4-point scale, while other options permit a continuous scale or point allocation. The value of an alternative is then the weighted combination of the values of each of its levels.

However they are operationalized, self-explicated models are best applied to gauge reaction to a particular offering, in the absence of an explicit competitive offering. They are also useful where there are a great many attributes or outcomes associated with the choice. For these reasons, self-explicated models are commonly used for services, such as how much more positively you would feel about a hotel if its service level improved. They also work well for actions lacking comparable alternatives, such as how likely you would be to trade in a car. This distinction between within- and between-alternative orientation is important. The self-explicated models (e.g. Fishbein, weighted-additive models) were primarily developed to measure attitudes for an alternative, to assess how the characteristics of a product lead you to like it. Unfortunately, positive attitudes often do not translate into purchase decisions. Witness the strongly vocal Macintosh users who nonetheless buy DOS or Windows-based machines. As this example illustrates, attitude toward a brand may be a poor predictor of purchase in a competitive setting.

Because direct evaluation draws attention to attribute levels, it tends to overweight attributes that might otherwise be unimportant in a competitive context. It is easy to think of reasons why, for example, an audiovisual feature (such as Bose speakers) might be important in evaluation of a computer but much less important in choice. In the market, this attribute may become overshadowed by more important attributes, or not perceived to differ sufficiently.

Self-explicated models also become problematic in the face of correlations among attributes, where there are strong expected associations between attribute levels. For example, computers that handle numerical calculations quickly are often faster at handling strings. The customer may assume one attribute is a surrogate for the other, both being measures of 'speed'. Suppose, however, the manufacturer needs to know the relative importance of each. Should each be included as a separate attribute? Does the value respondents provide in a direct rating of numerical speed include an implicit component for string handling? There are no simple answers to these problems within the context of the self-explicated methods.

While self-explicated models can become unstable or ill-defined in the face of correlated attributes, standard forms of conjoint get around this problem by breaking up expected associations among attributes. When exposed to computers with both high reliability and low weight, the heuristic of using one to 'stand for' the other is both less credible and less useful. By contrast, self-explicated tasks do little to break down pre-existing assumptions about the world. Indeed, they tend to reinforce both beliefs about the available levels of attributes and their associations. Accordingly, self-explicated tasks are most useful in simulating those markets whose offerings are stable.

There is some evidence that the self-explicated process works quite well when the alternatives reflect stable and veridical beliefs about the world. For example, in two studies of MBA job choice from offers received, the self-explicated model provides a better prediction than a conjoint model (Srinivasan and Chan Su Park 1997). In my view that makes sense because the job offers are likely to reflect expected levels and correlations. Thus, the expectation-based heuris-

tics
inherent in the self-explicated model work well. By contrast, where the offerings are substantially different from expectations, then a task that breaks from those beliefs should work better.

Summarizing, self-explicated models are best for:

1. Contexts in which many attributes are important for the decision.
2. Markets where expectations about levels and associations among attributes are stable.
3. Decisions where the action depends on the attitude towards an individual alternative or action, rather than in the context of competitive offerings.

Graded pair comparisons. The graded pair comparison task puts two alternatives next to each other and asks how much more one is liked than another, and in doing so immediately draws attention to the *differences* in attribute levels for the pair. For example, in the comparison between a 500MB laptop at $2400 against an 800MB model at $2900, the focus is on whether 300MB additional memory is worth the $500 price difference.

Three evaluative effects follow from the pairwise task. First, the task tends to flatten attribute importances by making otherwise unimportant attributes salient. Since attribute differences are easy to assess in a pairwise task, the importances are spread out across different attributes. This flattening of attribute importances tends to be particularly strong when there are only a few attributes differentiating the pair, as occurs in ACA's default. In tasks where the dominant attributes are missing, respondents come to value attributes that might otherwise be overshadowed.

Second, the pairwise orientation tends to reduce the importance of external reference levels. In particular, in valuing the differences between levels, there is less attention placed on their absolute levels. For example, suppose there is a real resistance to paying more than $3000 for a computer. However, in a pair task, there will typically be relatively little differences in respondents' reaction to computers with $2400-$2900 prices, compared to those with $2600-$3100 prices. The problem arises because thinking about the $3000 resistance level takes an extra processing step after the difference has been evaluated. This focus on differences becomes stronger and stronger as respondents repeat the pairwise task.

Finally, the focus on differences increases the relative value of attributes about which such differences are easy to value. It is easy to assess the implications of a $500 difference in price, but how should one value the difference between, say, Compaq and Dell? More generally, numerical attributes about which it is easier to characterize the difference, such as price, size, rating, will have greater weight in pair tasks than categorical attributes such as brand name, product family or country of manufacture. Nowlis and Simonson (1997) have shown that when one's focus is on individual alternatives then the categorical attributes have more weight, while a pairwise task emphasizes continuous attributes.

In summary, pair comparisons are most appropriate when:

1. Modeling a market in which alternatives are explicitly compared with one another.
2. Approximating a deeper search where the consumers draw information from a broad range of attributes.
3. Contexts in which within-attribute value steps are smooth and approximately linear.

Full Profile Ratings. A full profile rating is a very common form of conjoint in which respondents evaluate a group of, say, 16 alternatives, each defined as a bundle of characteristics. A typical task requires that each alternative be evaluated on a simple scale, say, a 1-7 attractiveness rating, or a 0-10 purchase likelihood. Regardless of the scale used, the critical defining aspect of this task is it encourages respondents to evaluate each profile *individually*. Put differently, attention is focused on the acceptability of an alternative's attributes, rather than differences between alternatives as we found for pairs. This seemingly innocuous attentional difference produces strong effects on shifts in expectations and on values that emerge from the task.

To understand the impact of a within-alternative focus on expectations, think about rating the attractiveness of a laptop. You might have feelings about it, for example, that the price is too high or the processor untrustworthy. However, to give it a rating it is important to know how it compares to others in the set. You need to learn to identify of the kind of laptop that gets "2" rating compared to one that gets a "5". Respondents achieve this mapping quite easily by getting a sense of the range and the average value of the profiles. Evidence for the speed and strength of this adaptation process can be seen by noting how the average rating hovers for most respondents within one point of the center of the scale, *regardless of the respondent's general attitude.* Indeed, in most analyses of ratings the mean is treated as a nuisance variable to be discarded, rather than a measure of attitude toward the category.

This adaptation of respondents to the average profile has important implications for the implementation of ratings-based conjoint. Since the adaptation does not occur instantly, respondents generate more reliable ratings if they understand the range of the profiles. In our work, we have found that full profile ratings predict better when preceded by an ACA task than when followed by it (Huber, Wittink, Fiedler and Miller 1988). The self-explicated and pair sections in ACA permit people taking the subsequent full profile task to have a good sense of the attribute ranges and how they combine to make more or less attractive alternatives. More generally, warm-up tasks are very important in full profile conjoint. Louviere (1985) recommends two warm-up tasks, first rating an alternative that is poor on most attributes, followed by one that is good. These two tasks familiarize the respondent to the scales and stabilize subsequent ratings.

In addition to efficiently shifting respondents from their external reference levels to the average within the set, the full profile task also is quite effective in breaking up associations between attributes. Respondents' associations are weakened by profiles that go against their expectations; for example, when they find quality processors with low power, or light-weight laptops with long battery life. Indeed, the orthogonal designs common in most full profile exercises require that any pair of attribute levels have an equal likelihood of being paired within a profile, thus assuring that respondents will experience profiles that violate their prior expectations. The net effect of the full profile conjoint task is to generate decisions that are relatively free from simplifications that come from reference and associational effects. Even more so than graded-pair comparisons, full profile ratings *decontextualize* respondent values.

In addition to producing values that are relatively context free, full profile's focus on the individual alternative changes the resulting values compared with graded pairs in three ways: fewer attributes are featured, those featured tend to be qualitative, and greater weight is attached to the lowest levels of each attribute. Each of these value shifts is considered below.

The first value shift is a focus on a few attributes. There is no logical reason why ratings-based conjoint should limit attention to a small number of attributes, but that is what happens, study after study. At the individual level the pattern is clear; out of say seven attributes, two or three attributes will be significant, while the rest are virtually zero.

In addition to a small number of attributes becoming prominent in the full-profile task, those featured are more likely to be qualitative (Simonson and Nowlis 1997). This qualitative focus follows from the full profile's orientation to the individual alternative, and contrasts with the focus on numerical attributes in pair comparisons discussed earlier. Qualitative attributes tend to have attitudes attached to them regardless of context. Consider your immediate evaluation of the brand name, Packard Bell, or the feature "multi-media." By contrast, how you feel about a 100MhZ processor depends critically on whether it is compared with an 80MhZ or a 130MhZ. In a conjoint rating task the standard of comparison is implicit, while in pair comparisons the contrast is with the other pair. The implication is that if the market action you wish to simulate by the task involves such explicit comparisons, then a pairwise method is preferred. To the extent that each alternative is evaluated alone (like homes, cars, recordings) then a full profile task is more appropriate.

Finally, there is evidence that full profile puts greater weight on the negative levels of attributes than pair comparisons. For positively-coded attributes, this orientation is reflected in diminishing returns, so that the gain from a one- to a two-year warranty is greater than the gain from a two- to a three-year warranty. For negative attributes, it is expressed as increasing aversion to the negative attribute, so that the loss of moving from 4 to 6 pounds for a laptop is less than the move from 6 to 8 pounds (Orme, Alpert and Christensen 1997, Huber, Ariely and Fischer 1997). The process driving this curvature is a combination of task simplification and loss aversion. The task is simplified by downgrading alternatives containing the least-preferred attribute levels, while avoiding these low levels protects against losses associated with making a bad choice.

To summarize, full profile conjoint is most appropriate when:

1. It is desirable to abstract from short run level and associational beliefs.
2. The market choices demonstrate substantial simplification both in a limited number of attributes being processed and greater weight on the most negative levels.
3. The focus of the decision is within alternative so that the explicit comparisons between pairs of option are rare.

Choices. A choice task can be viewed as a group of full profile concepts, where, instead of requiring that each be individually rated, the respondent is asked to indicate which is best. However, this formal similarity to the conjoint rating task belies strong processing effects that derive from the act of choosing. Choosing shifts attention away from assessing how much better one alternative is compared to another and towards processes that lead one to be reasonably confident that the one chosen is best. This goal encourages even greater simplification than a rating task. Some of this simplification is evident in the time taken. A 9-attribute, 3-alternative choice task took about 30 seconds per choice, while a rating task took about 30 seconds for each alternative (Orme, Alpert and Christensen 1997). Clearly, respondents are not evaluating each of the alternatives and choosing one with the highest score.

What are they doing? First, respondents are looking for dominating, easy choices. If they find none, they look to see if they can exclude any of the alternatives, typically those with low scores on important attributes. Once the choice is down to two, then a quick scan of important attribute differences completes a satisfactory selection process. In part because of such strong simplification, the process is not very reliable. In choice experiments where one choice set has been repeated, the same alternative (5 attributes, 4 alternative) is chosen only 70%-80% of the time (Huber, Wittink, Fiedler and Miller 1993).

What is the impact of choices on expectations? As with pair comparisons and full profile tasks, respondents quickly leave their old reference levels as they adapt to the alternatives provided. Of course, if the alternatives are too bad then people react negatively to the entire process, and if they are too good, they may make very little effort to choose, since all are satisfactory. However, within bounds, respondents in choice tasks quickly learn to evaluate each alternative compared with the local competition within the choice set.

Associations between attributes are more difficult for respondents to see in choice sets, but learning does eventually take place. The Johnson and Orme (1996) study shows that the relative importance of brand name relative to price drops by 30% in the course of 3-4 initial choice sets and to 50% after 10 tasks. Initially, brand name is important. Soon, however, respondents realize that brand name is not predictive of price or features, and evaluate its contribution, *holding other aspects constant.* In the same way, they also learn to evaluate each attribute independently of other attributes commonly associated with it.

Choice tasks shape values in three ways. First, greater simplification leads to even fewer attributes being featured (Orme, Alpert and Christensen 1997, Zwerina and Huber 1997). Second, the attributes featured are different. Since choices combine both within-alternative processes (like full profile) and between alternative judgments (like pair comparisons), the focus is not with respect to quantitative or qualitative, but follows from the fact that choices are more *immediate* and *real.* Choices lack the abstract and hypothetical quality of ratings--respondents are being asked if they would actually choose the alternative. Attributes whose impacts are immediate and concrete come to the fore compared to those that are distant or abstract. Consider the following two examples. First, IntelliQuest (Pinnell 1994) has found that the utility values for price have to be doubled to make their ACA values match the subsequent choice task. Second, in our study of refrigerators, we found that long-term cost of annual energy use was more important in ratings than in choice, whereas the immediately due sales price was more important in choice over ratings (Huber et al. 1993).

The third way in which choices shape values is by putting even greater weight on the poorest attribute levels. This tendency is manifest in large utility differences between poor-middle levels, and relatively small differences between middle-best levels of the attributes (Orme, Alpert and Christensen 1997, Zwerina and Huber 1997). The mechanism here is the same combination of loss aversion and simplification found in the full profile task, but the effect is stronger. Rather than getting lower ratings, alternatives with low levels on important attributes are more likely to be simply dropped from consideration.

To summarize, choice is most appropriate when

1. Simulating immediate response to competitive offerings, especially brand and price studies.
2. Decisions are made on the basis of relatively few, well-known attributes with substantial aversion to the worst levels of each attribute.
3. Consumers make these decisions on the basis of competitive differences among attributes given.

WHAT ABOUT DIFFERENT VERSIONS OF THE METHODS?

To simplify the exposition, the preceding sections have deliberately focused on relatively pure types of the self-explicated, graded pair, full profile and choice tasks. Of course, most implementations of these tasks differ importantly from these pure forms in terms of the ways they affect attention, competitive expectations and values. However, the logic used to understand the task effects of the pure forms can be used to predict the impact of these modifications. A few examples are given below.

ACA: ACA combines a self-explicated and a pair comparison task. The self-explicated task permits a good introduction to attribute levels and tends to bring more attributes into consideration. The pairwise task further increases attention to less important attributes, since attribute differences are so easy to process. Finally, the focus on differences and the linear priors tends to result in quite linear steps in utility between adjacent levels.

If desired, ACA can be made more like choice by encouraging simplification and loss-aversion. Curvature in the ranking of levels can be encouraged by changing the task to having respondents assign the best level of each attribute 100 points and proportional values to lower levels. Furthermore, simplification emulating choice can also be encouraged by having pairs differ on more attributes, say 4-5 attributes differing rather than the default two or three. One may not want to make this modification, however, as there is evidence that ACA works best with 2-3 attributes differing (Huber and Hansen 1988).

Sort Board for Full Profiles: A common way to make full profile ratings more like choice is to give respondents a deck of cards and ask them to sort them on, say, a board with 10 categories from groups from worst to best. This task brings in attentional properties that mirror some aspects of pair comparisons and choice. Respondents typically group the cards into rough categories followed by a more detailed evaluation of alternatives in the same pile. This latter pairwise focus tends to bring attention to less important attributes, since the alternatives sorted together often have the same values on the most important ones. Further the two-stage process of an initial screen followed by more detailed pairwise assessment of final alternatives mirrors what happens in a number of choice contexts (Payne 1976).

Simplified choices: Not only are there ways ratings can be made more like choice, but choices also can be made more like ratings. Since the major property of the choice task is that it encourages simplification, a common way to limit this tendency is to reduce the processing required. Two ways are possible, reducing the number of alternatives per choice (Pinnell and Englert 1997), or reducing the number of attributes differing (Chrzan, Fellerman, 1997). Both

these methods lessen the statistical power of the design, but increase the ability of respondents to respond consistently to the task. Generally, simplifying choice can be expected to increase the number of attributes that are processed and decrease the weight put on the least preferred attribute levels.

SUMMARY: A FRAMEWORK FOR EVALUATING METHODS

Table 1 provides a summary of the different methods and their impact on attention, competitive beliefs, and resultant values, permitting a link between the market decisions and the appropriate task. Below, I reiterate the important ways the measurement task shifts attention, competitive beliefs and resulting values, and suggest ways that this knowledge can be used to guide the development of useful commercial studies.

Attentional shifts are an integral part of any value measure. Simply mentioning an attribute increases its importance, raising the specter of attributes appearing important that otherwise would be ignored in the market choices. One way to limit this problem is to load the task with enough attributes so the unimportant ones are ignored in the task process. This task simplification screening is particularly strong in full-profile ratings and choices. The risk here is that the task may encourage respondents to ignore too many attributes, in which case pair comparisons or self-explicated tasks may be more appropriate. In any event, it is important to think about ways attributes can become important in the market context, for example through a promotional campaign, shelf talkers, or simply gradual understanding of the market over time. It is those attributes that should be featured in the task.

Competitive beliefs are changed by value measurement tasks. Except for self-explicated methods, all the tasks discussed decontextualize judgment by shifting reference levels and changing associations. Reference levels refer to expected levels and ranges of each attribute. These levels assist our market decisions by gauging whether a particular offering is appropriate or not, and enable us to make reasonable decisions in very little time. However, these reference levels are also quite sensitive to the competitive context. Consider the following two examples. What seems like an outrageous price can quickly become acceptable in the face of higher-priced competitive offerings. What seems like an appropriate modem becomes outmoded when compared with the faster models.

To my mind, decontextualizing values from particular or reference levels is an advantage of the various trade-off methods. Just as conjoint easily shakes people from their reference points, so market forces also shift these reference levels. Sticker shock may make people put off buying a car, but eventually they adapt to the new competitive level. Thus, a well-designed tradeoff study can anticipate the effect of adapting to new market offerings. The caveat is that the conjoint context needs to match the future market.

The second change in competitive beliefs is with respect to associations. Like reference levels, associations allow people in markets to make reasonable choices quickly, by selecting a trusted name, store or price tier. Breaking down these associations requires that people really assess the value of, for example, the brand name in itself. Thus, the process of breaking down associations can be seen as a way of approximating what a person would do if expectations are not used to simplify the decision. While it may result in somewhat worse predictions of short-term decisions, it can better approximate the effect of extended thought or discussion.

If measurement tasks focus attention and shift competitive expectations, they also tap different values. We have reviewed three ways in which the task can alter derived values: through an orientation to individual items versus a comparison with others, through the immediacy of the situation it evokes, and through a need to simplify the task. Below is a summary of each of these distortions along with suggestions as to how they might be better handled.

Tasks can evoke either a *comparative or individual-alternative* orientation. As argued earlier, pair comparisons result in greater weight to those attributes whose differences are easy to calculate, whereas full profile ratings put greater weight on categorical attributes such as brand name. The choice of which to use depends on the degree to which the market decision is a comparative one. Thus, the choice of laptop is typically a comparative process, whereas the choice of a job or a house is generally more focused on the fit to one's own values. Further, decisions where the alternatives are not comparable, such as when to sell or whether to buy in a category, focus attention within alternatives and thus are best modeled by ratings-based conjoint or even a self-explicated model.

Tasks can be *immediate or reflective*. Immediate tasks, such a choice experiments, ask respondents which they would choose today. As discussed earlier, the more immediate tasks increase the importance of attributes with short-term implications, such as price, and attributes with visible performance characteristics. The more reflective tasks (say, asking for a tradeoff between two pounds of weight and an hour of battery life) are both hypothetical and relatively timeless. One does not consider one's next business trip, but instead the general pattern of such trips. The implication should be clear. To the extent that the market decision being simulated is based on immediate and short-term considerations, then choice experiments are appropriate. Long-term and repeat purchasing contexts, by contrast, are better modeled by procedures that encourage respondent to abstract from current considerations.

Finally, tasks can evoke varying degrees of *simplification*. Respondents simplify tasks both across and within attributes. Across attributes, respondents simplify by attending to the most important attributes at the expense of less important ones. Within attributes, they discard alternatives that have low levels on important attributes, typically producing the appearance of strong diminishing returns in the partworths. Choices result in the most simplification, followed by full-profile conjoint, pair comparisons and the self-explicated task. Thus the various tasks provide a way of simulating more or less simplification.

As we examine the ways these tasks focus attention, alter competitive beliefs and change revealed values, the dilemma posed at the beginning of this paper emerges. If we wish to predict short-term, heuristic-bound behavior, then none of the methods reviewed are very good, although some, like choice, may be better than others. However, I believe marketing research and marketing firms will be better off if they err on the side of encouraging more elaborate over less elaborate processing: drawing attention to more attributes rather than less, encouraging a long term rather than an immediate focus on the problem, and by breaking apart the problem for the respondents so that they do not too grossly oversimplify it for themselves. This leads to a recommendation to use pair comparisons and self-explicated methods, and to be particularly cautious with choice-based methods.

There are two reasons for taking this posture. First, individuals certainly use all kinds of shortcuts in making market decisions. However, while individual decisions may make them

vulnerable to opportunistic marketers, customers do learn, both individually and collectively. Thus we want a technique that enables people to be most happy with outcomes of the decisions that they make, not to make the decision that they will see later as foolish or short-sighted. The more aspects customers consider, the more long term their orientation, the more they are able to cope with the complexity of the problem, the more satisfied they will be with the offerings they choose.

Konosuki Matsushita said it best:

> "Don't sell customers goods that they are attracted to. Sell them goods that will benefit them." (Fortune, 1997)

I would like to close with a thought experiment. Suppose as part of this conference you win a laptop; but you do not get to choose the laptop. Instead, you get to choose the method that will select the laptop for you. You choose from among the four methods discussed here: self-explicated, pair comparisons, full profile or an individual choice experiment. You will then get the laptop that optimizes your values as expressed by your chosen technique.

Most knowledgeable people do not like this question, seeking control over the choice rather than the method to choose. However, when pushed, most knowledgeable people prefer the biases and distortions from the more thoughtful exercises such as self explicated or pair comparisons compared with full profile ratings or the choice-based task. Choices are seen as evoking in too much simplification, with too much focus on near-term consequences, and too much emphasis on avoiding the 'worst' levels of an attribute. Pair comparisons and self-explicated tasks, by contrast, may err by putting too much weight on unimportant attributes, and perhaps not enough weight on the worst levels, but they are robust and reliable.

The question then is, knowing what you know about the strengths and weakness of the different tasks, which method would you choose to select your own laptop? Perhaps more relevant, which method would you select for your own customers?

SUMMARY OF DIFFERENCES AMONG VALUE MEASUREMENT TASKS

	Self-Explicated	Graded pair comparison	Full profile ratings	Choice
Attentional Focus	Individual alternatives	Pair differences	Alternatives in a general context	Alternatives in a competitive context
Impact on Competitive Expectations	Reinforces prior expectations	Shifts expected trade-offs (e.g. price-quality) and levels.	Orthogonal arrays break down associations.	Initial reference levels are dominated by attribute differences.
Valuation Focus	Feelings towards attributes	Tradeoffs between levels	Selection of attributes and levels	Short term and concrete attributes, loss avoidance
Emphasis	Less important attributes	Quantitative attributes	Qualitative attributes	Near-term, concrete attributes
Ideal use	Non-competitive contexts, many attributes	Stable trade-offs	Gauge simplification strategies	Immediate competitive effects, simple choices.

REFERENCES

Chrzan, Keith and Ritha Fellerman (1997), "A Comparison of Full- and Partial-Profile with Best-Worst Conjoint Analysis," *Sawtooth Software Conference Proceedings.*

Green, Paul, E, Yoram Wind and Arun K. Jain (1972), "Preference Measurement of Item Collections," *Journal of Marketing Research, 9,* (November) 371-377.

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

Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research, 23,* (August) 307-317.

Huber, Joel (1987) "Conjoint Analysis: How We Got Here and Where We Are," *Sawtooth Software Conference Proceedings,* Ketchum ID: Sawtooth Software, 237-253.

Huber, Joel, Dan Ariely and Gregory Fischer (1997), "The Ability of People to Express Values with Choices, Matching and Ratings," Working Paper, Fuqua School of Business, Duke University.

Huber, Joel, Dick R. Wittink, John A. Fiedler and Richard Miller (1993) "The Effectiveness of Alternative Elicitation Processes in Predicting Choice," *Journal of Marketing Research* (February), 105-114.

Johnson, Richard M. and Bryan K. Orme (1996), "How Many Questions Should You Ask in Choice Based Conjoint?" Presented at the ART Forum, Beaver Creek CO.

Kuhfeld, Warren, Randall D. Tobias, and Mark Garratt (1994) "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research,* (November), 545-57.

Louviere, Jordan (1988), "*Analyzing Decision Making: Metric Conjoint Analysis,"* Newbery Park, CA: Sage Publications.

Orme, Bryan K., Mark I Alpert and Ethan Christensen (1997) "Assessing the Validity of Conjoint Analysis—Continued," *Sawtooth Software Conference Proceedings.*

Pinnell, Jonathan (1994) "Multistage Conjoint Methods to Measure Price Sensitivity," Presented at the ART Forum, Beaver Creek, CO.

Pinnell, Jonathan and Sherry Englert (1997) "Design Considerations in Choice and Ratings-Based Conjoint," *Sawtooth Software Conference Proceedings.*

Payne, John W. (1976) "Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis," *Organizational Behavior and Human Performance, 16,* 366-387.

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

Wright, Peter and Mary Ann Kriewall (1980) “State of Mind Effects on the Accuracy with Which Utility Function Predict Marketplace Choice,” *Journal of Marketing Research, 17,* (August) 277-293.

Zwerina, Klaus and Joel Huber, “Deriving Individual Preference Structures for Practical Choice Experiments,” Working Paper, Fuqua School of Business, Duke University.

COMMENT ON HUBER

Carl T. Finkbeiner
National Analysts, Inc.

Put simply, I believe that Joel's fundamental message is that context matters, an opinion with which I strongly concur. In conjoint research, the context within which judgments are made affects the utility structure that results from the analysis of those judgments, regardless of the conjoint method being used.

To me, the implications of Joel's basic message are as follows:

- We ought to consider extending our models to cover multiple methods. "True" utility structure isn't necessarily a "myth," as Joel states, if we can successfully take context into account in the modeling.
- We *should* select methods that match the context being modeled. The "real world" also provides a context within which judgments are made. Try to match that context with the method used to estimate outcomes for the "real world."

Using holdout choice tasks to measure the success of a model is inherently biased in favor of choice models. The holdouts choices are obtained within a context that is most like that of discrete choice methods, creating, to the extent that context affects judgments, the aforementioned bias. Thus, for example, a holdout task favoring full-profile conjoint would be ratings of holdout full-profiles.

In a good effort at clarity, Joel offers decision criteria for selecting methods. As much for my own benefit as for the reader, I summarize them briefly here. Joel does an excellent job of amplifying on these points.

Number of attributes allowed (or required).

Is the goal estimation for an individual alternative or of competitive shares?

Is the goal to reflect stable marketplace expectations or preferences, or is it to investigate hypothetical "what-if" scenarios?

Does the market being modeled reflect over-simplification in decision-making?

Is the focus of the modeling a short-term estimate or is it longer-term?

I do have one significant area of concern with Joel's comparisons of methods. He makes statements which I paraphrase as "Full-Profile Conjoint and Choice place a relatively greater weight on the negative level(s) of a quantitative attribute as compared to Self-Explicated models." By this, I take him to mean that the Self-Explicated utility function is relatively flatter on the negative end, at least as compared to Full-Profile Conjoint or Choice utility functions.

To test my own experience regarding this assertion, I examined results from nine studies that I had readily at hand. I cannot provide results from most of these studies since they are proprietary, however, I can say that several counter-examples to Joel's hypothesis occurred. Two of the studies I examined are in the public record: one is the Finkbeiner-Platz study I presented at the

ACR conference in Toronto in 1986; the other is the 1996 Zwerina-Huber study which is still "in-publication," but was reported by Rich Johnson and myself in our papers at this conference. In the first study, a Full-Profile Conjoint (using sort boards) was compared to the Self-Explicated portion of ACA and the full ACA including Paired Comparisons. In the second study, Self-Explicated, Full-Profile Conjoint (using one-at-a-time ratings), and Choice data were collected and analyzed by several methods. I can provide the interested reader with all of the partworths from both of these studies. In both sets of data, there is evidence consistent with and evidence counter to Joel's hypothesis.

I conclude from my examination that, though there exist cases where Joel's assertion is true, there are also a number of cases where Conjoint and Choice do *not* more heavily weight negative levels. It might be useful to determine the cases where Joel's hypothesis should hold, and where it won't.

Beyond this disagreement, I found Joel's paper to be a very brave and helpful attempt at bringing organization and understanding to a confusing topic. For whatever they are worth, my own conclusions about this topic are as follows.

- I agree with Joel about Self-Explicated Ratings being most useful when there are many attributes, when we are modeling stable expectations and preferences, and when we wish to estimate for individual product alternatives. These circumstances are most clearly met in customer satisfaction modeling studies, and so that is where I tend to use the Self-Explicated method the most.
- I believe that the Paired Comparison and Choice methods *both* encourage over-simplification of the judgment task , and so may not predict well when the context being modeled is one in which over-simplification cannot occur.
- One-at-a-time Full-Profile Conjoint is too focused *within* product and benefits from forcing some *between* product comparison into the task. Consequently, whenever the setting permits it, I prefer using sort boards to one-at-a-time ratings to obtain Full-Profile ratings data.

Models which estimate many parameters with few degrees of freedom (as is the case with individual-level models such as Full-Profile Conjoint) tend to produce counter-intuitive partworths which, though non-significant statistically, may create confusion in interpretation of choice simulations. For example, if a respondent is not attending to price, the partworths for price may apparently indicate that the respondent prefers high prices over low prices, all other things being equal. This implausible outcome usually arises because of sampling error. Such models often benefit from being constrained to be consistent with Self-Explicated ratings, which almost never show this counter-intuitive outcome at the individual level.

CURRENT PRACTICES IN PERCEPTUAL MAPPING

Thomas A. Wittenschlaeger
Hughes Aircraft Company
John A. Fiedler
POPULUS, Inc.

The paper uses data from a proprietary Hughes survey to demonstrate the principles which underlie current practices in perceptual mapping using discriminant analysis-based maps. The paper discusses the advantages of using discriminant analysis in creating perceptual maps, criteria for selecting brands and products for respondents to rate, and principles for optimizing a perceptual space.

INTRODUCTION

Perceptual mapping has been used as a strategic management tool for about thirty years. It offers a unique ability to communicate the complex relationships between marketplace competitors and the criteria used by buyers in making purchase decisions and recommendations. Its powerful graphic simplicity appeals to senior management and can stimulate discussion and strategic thinking at all levels of all types of organizations.

Despite their high value as a decision-making tool, perceptual maps are easy to produce. Most currently popular mapping procedures utilize readily available ratings data which satisfy management's need for a competitive score card. Despite their having been around for thirty years, perceptual maps are still viewed as an innovative technique.

WHY USE DISCRIMINANT ANALYSIS TO PRODUCE PERCEPTUAL MAPS?

From the authors' experience, two approaches are most commonly used today to produce perceptual maps: Correspondence Analysis (CA) and Discriminant Analysis (DA). CA is generally easier to use than DA; it can be used with aggregated data such as cross-tabulations while DA cannot. However, DA offers several advantages over CA which make it the authors' usual first choice:

(1) DA has a close linkage between product points and attribute locations. When high proportions of information are accounted for by the map, the products' projections on each vector are perfectly correlated with their means on that attribute. There is a hard-to-understand relationship between products and attributes in CA, but even Michael Greenacre has argued that nobody should try to infer anything about those relationships from CA maps.

(2) Unlike CA and factor-analysis-based mapping, DA maps do not change if attributes are added that are linear combinations of those already present in the space.

(3) DA is alone in paying attention to "between product" information, after scaling it so that "within product" differences are equal for each dimension and uncorrelated. That means

that DA uses a "yardstick" to give every dimension common metric (in terms of equal unexplained variance). Neither CA nor factor-analysis-based mapping techniques distinguish between-products differences from within-products differences at all.

(4) DA is the most efficient method, in terms of cramming into a space of low dimensionality the most information about how products differ. After implicitly rescaling the data to have "spherical error," DA provides in its map the least squares approximation to the entire data matrix for that number of dimensions. Since managers have severe problems understanding higher-dimensional structures, and DA gives you the most information in the fewest dimensions, DA permits superior communications.

(5) Unlike mapping based on distances or similarities, DA makes use of attribute ratings, which are easy and natural for respondents, and useful for their content even if mapping is not done with them.

(6) Fiedler (ART) showed that DA was more successful than CA at reproducing a known map when the data were distorted in various ways.

DEMONSTRATION DATA SET

The authors use data from a proprietary Hughes project to demonstrate the principles of current best practices in DA-based mapping. The study dealt with air traffic management systems; it was a world-wide project with 301 decision makers from both the public and private sector. The interview was programmed in three languages, conducted at three different international conference and via D-B-M, and programmed in Ci3, ACA, and APM. The questionnaire made extensive use of visual aids.

The focus of the study was system design and development of future ATM systems which cannot be discussed due to the confidentiality requirements. Data relating to vendors' perceptions of ATM systems was tangential to the objectives of the research and was made available to the authors for this conference. Limited masking of data has protected Hughes' proprietary interests.

ADAPTIVE PERCEPTUAL MAPPING

Sawtooth Software's Adaptive Perceptual Mapping (APM) was utilized in the questionnaire. APM employs discriminant analysis.

Advantages of APM. It is exceptionally easy to use. It permits the use of an incomplete design in which respondents only use rating criteria which they believe to be important, and they only rate products which they know best. This tends to result in meaningful tasks. The software offers an interactive rotation option which greatly simplifies the process of producing effective maps.

Disadvantages of APM. APM's weaknesses reflect its age. The respondent interface does not offer the flexibility of SSI's newer products. The programming interface reflects the product's Ci2 heritage.

CHOOSING WHAT TO RATE

All mapping techniques attempt to show the comparative differences in how products are rated on attributes. The validity of a map depends on both the overall set of attributes and brands in the study as well as the subset of attributes and brands evaluated by each respondent.

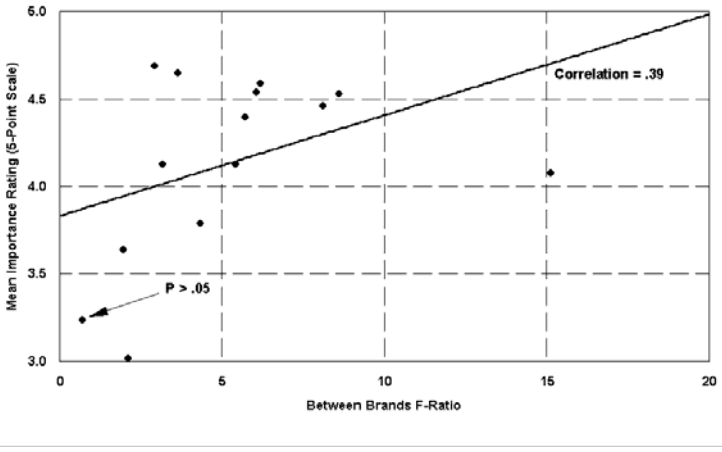
Most studies suffer from too many attributes. Manufacturers and service providers see hundreds of ways in which their products and services differ—or might differ—from those of their competitors. Often the research analyst is unable to impose the discipline necessary to develop a reasonably short list of attributes. In most studies it is usually desirable (or necessary) to select a subset of attributes for respondents to rate. This can be accomplished by using one of two approaches:

(1) *Select a subset of most important attributes.* Each respondent rates all attributes on importance. The questionnaire is programmed to select a subset of the important attributes for rating. This may assure more meaningful questionnaire tasks for respondents.

(2) *Randomly select a subset of attributes.* The questionnaire randomly selects a subset of attributes for each respondent. This has the advantage that there will be roughly equal sample sizes for each of the evaluative criteria. The obvious disadvantage is that the respondent task may be less interesting.

Which approach is better? Figure 1 compares average importance scores and F-ratios from the Hughes ATM study.

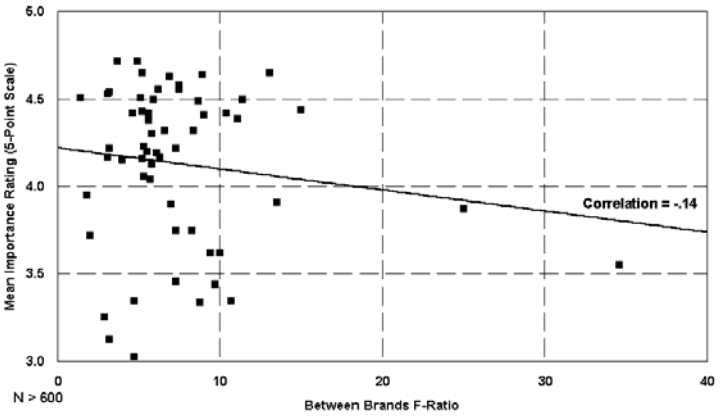
Figure 1
Discrimination Versus Stated Importance: ATM Systems



Respondents rated products on five most important attributes (of 14 altogether). Discrimination and importance are correlated; allowing respondents to use those attributes which they judged to be important was the correct decision.

Figure 2 compares discrimination and importance from another study. In this study a random subset of attributes (20 out of 57) was chosen for each respondent.

Figure 2
Discrimination Versus Stated Importance: Consumer Non-durable



Discrimination and importance are uncorrelated. Two of the most discriminating attributes are among those judged to be least important; several of the "most important" attributes are among the least discriminating.

This is not an uncommon result. Very often category-defining attributes (such as "fluoride" in toothpaste or "good taste" in food) are included in a study. These define the price of entry into the category, are generally rated very important, and usually fail to discriminate. It is often difficult to successfully argue for their exclusion from a study. Conversely, brands are often differentiated by attributes which consumers judge to be irrelevant.

Conclusion and Recommendations. Restricting ratings to "most important" attributes may overlook attributes critical to marketplace differentiation; such a restriction may limit ratings to attributes which define the category rather than describe brands. The design objective should be to maximize discrimination. We believe this can be accomplished in two ways:

(1) *Rate more products at the expense of attributes.* We are interested in how individuals compare products. As the *APM System Manual* states: "For each attribute, we assume that the *differences* [emphasis in original] among a respondent's ratings provide useful information, but that his *average* ratings for each attribute do not." The entire object of perceptual mapping is to display perceptions in a reduced space. It is a waste to have a respondent rate only one or two brands on dozens of attributes when he or she could rate five or six brands on seven or eight attributes.

(2) *Rate products within attributes.* This is the approach taken by APM and is most likely to maximize discrimination particularly with rational or practical benefit-oriented attributes. Occasionally the reverse is better. If a brand is viewed and evaluated holistically—such as soft drinks, cigarettes, or beer—and is being evaluated using brand personality scales and user imagery checklists, then it is typically better for the respondent to evaluate one brand in terms of all attributes before moving onto another brand.

ACHIEVING THE BENEFITS OF APM WITHOUT ITS LIMITATIONS

Ten years ago, APM made producing DA-based perceptual maps easy and economical. While the system has not evolved with other Sawtooth products, it remains an easy-to-use method for producing superior DA-based maps. It is also possible to achieve all the benefits of APM—and overcome its few limitations—using Ci3, standard statistical software, and spreadsheets.

Getting the Data in Ci3. Sawtooth Software's Ci3, with its LISTS and ROSTERing, can generate a data set with any conceivable brand and attribute selection logic. Given typical client requirements, many studies will have too many attributes with only a subset used by any one respondent and there will be many brands with only a subset used by any one respondent; thus the resulting Ci3 data file will be very large but mostly empty.

Generating a Perceptual Space. A DA-based perceptual space can be generated using SPSS or almost any other statistical package. There are nine steps in the process; SPSS code for each of these steps is included as an appendix to this paper.

(1) Convert data from Ci3.

(2) Build a system file with one record for each set of ratings for each brand. There will be as many records as the product of the total number of brands times the total sample size. Create a ".sav" file for product ratings for each brand. This produces as many files as there are brands.

(3) Concatenate all files and sort records by brand number within respondent number.

(4) Eliminate non-rated brands.

Steps 5 through 7 implement APM's approach of centering each person's attribute ratings across brands rated. This permits each person to use a unique subset of attributes and maximizes between-brand discrimination. From the APM manual: "We convert all ratings to 'deviation scores' so that each respondent's average for each attribute is zero. Experience has shown that this results in a reduction of random variation and increases precision of the measurement of the differences between products."

(5) Aggregate mean attribute ratings by respondent. Create a file with each respondent's average ratings for all attributes rated across all brands rated (excepting an ideal brand).

(6) Match the mean ratings file to each respondent's individual brand ratings record. Subtract the means from each rating.

(7) Recode all non-rated data to zero.

(8) Run discriminant analysis. The brand rated is the dependent variable; restrict the solution to two (or rarely three) dimensions. Aggregate mean discriminant scores by brand, segment, etc. and save to a worksheet.

(9) Correlate attribute ratings with discriminant scores and copy to spreadsheet.

Creating Perceptual Maps. Maps can be readily created using Excel, 1-2-3 or any other spreadsheet package. There are five steps to the process:

(1) Scale both the attribute and brand centroid matrices so that the largest absolute value in each is unity.

(2) Insert pairs of zeros between each row of attribute correlations.

(3) Create X-Y chart for attributes by specifying a line connecting all data points (these will be your vectors)

(4) Add series point labels for each attribute at the end of each vector. These can be rotated manually if the chart is subsequently pasted into a graphics program.

(5) Create a second X-Y chart for brand centroids and other points such as ideal point segments.

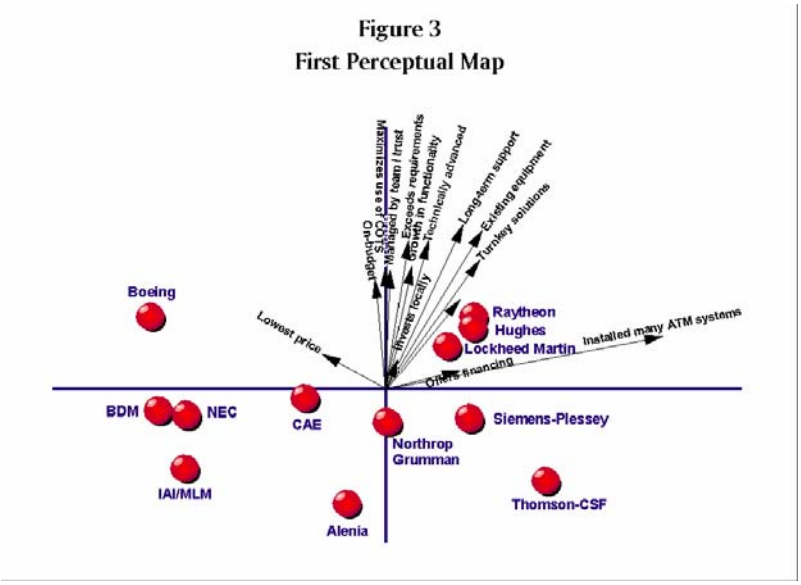
(6) Overlay the two charts.

OPTIMIZING A PERCEPTUAL SPACE: FIVE PRINCIPLES FOR POWERFUL MAPS

The procedures necessary to create a perceptual space using discriminant analysis are relatively straight forward. However the results may be initially disappointing or difficult to interpret and communicate.

The *first map* in the Hughes ATM study is shown below:

Figure 3
First Perceptual Map



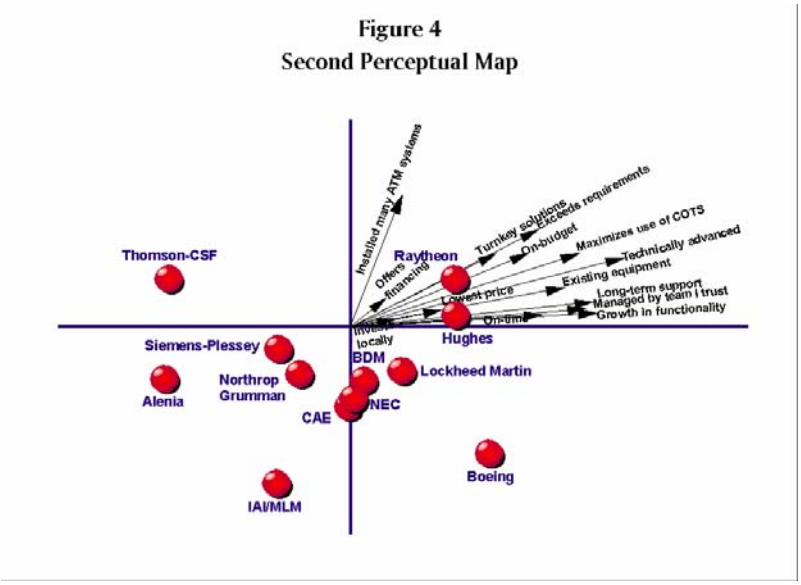
The client observed that the map failed to discriminate between the major competitors who are mostly in the upper right quadrant. Most of the space is determined by the way smaller fringe companies are perceived. The client sought to capture the differences between the major competitors and then fit in the other companies.

The analytic solution was to re-run the discriminant analysis using only a sub-set of brands as the dependent variable and subsequently calculating and plotting discriminant scores for non-specified companies.

The First Principle: A perceptual map should discriminate between *major* brands. If you're trying to show the relationships between cities in Washington State, don't include Washington, D.C., in your map.

The *second map* based on based major ATM competitors is shown below:

Figure 4
Second Perceptual Map



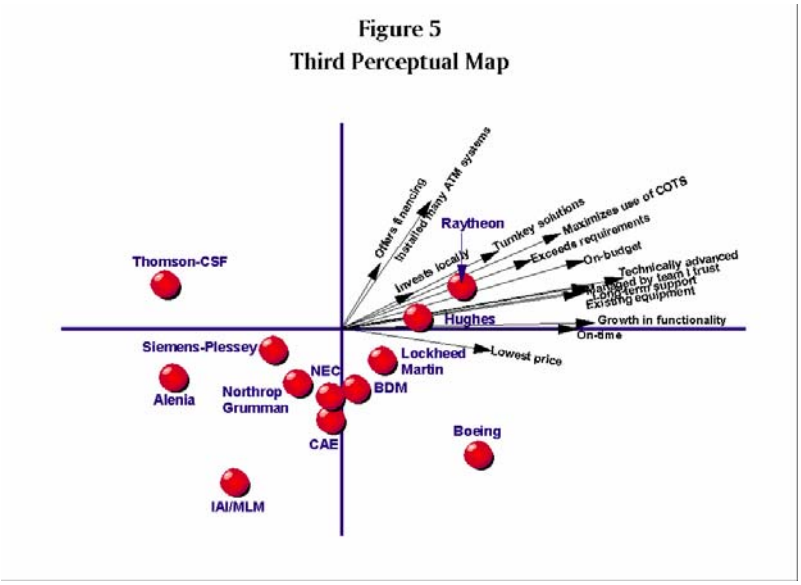
The client observed that this second map was better, but still not very useful. All the attributes appeared highly intercorrelated. Was there some way to "fan" them out?

The analytic solution hypothesized that different people with different needs have different perceptual frameworks. Segmentation may reveal that companies are perceived differently by different groups. This hypothesis was tested by replacing—as sample size permitted—each brand with five "brands," one for each of the ACA-based clusters.

The Second Principle: Attributes should occupy as much of the perceptual space as possible.

The *third map* based on major ATM competitors broken out by segment is shown below. The segment differences are not shown to protect proprietary findings.

Figure 5
Third Perceptual Map

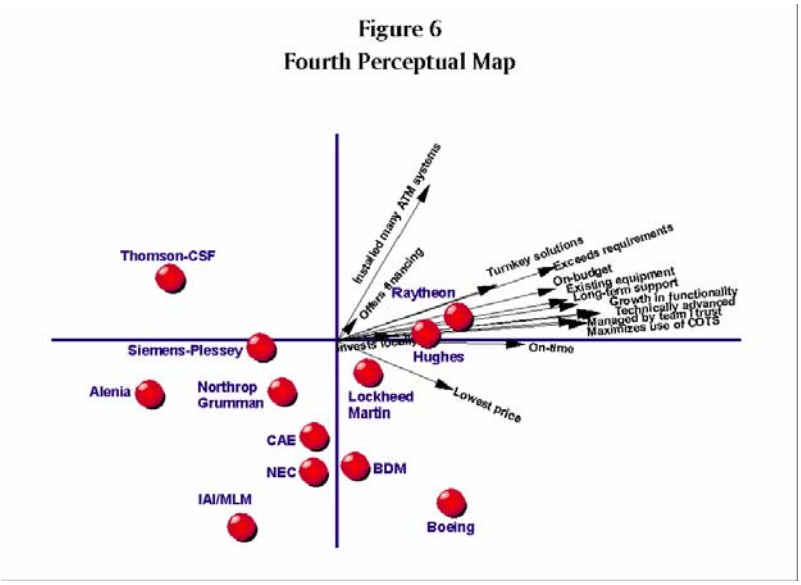


The client observed that, while this space accounted for need-based segment differences, variances in perceptions might be more related to regional differences rather than differences in product needs.

The resulting analysis was straight forward: as sample size permitted, replace each brand with six "brands," one for each of the geographic regions.

The Third Principle: The map should capture the *most important* sources of variance between brands.

The *fourth map* based on major ATM competitors broken out by geographic region is shown below. Again, the segment differences are not shown to protect proprietary findings.



The client observed that while the space was right, it would not be intuitively clear to management. Could the space be rotated so the axes are more closely aligned with some of the attributes?

The analytic solution was to rotate the space 22° counterclockwise. To rotate a perceptual space, multiply x-y coordinates of each point (or pairs of individual-level discriminant scores) by the following transformation matrix:

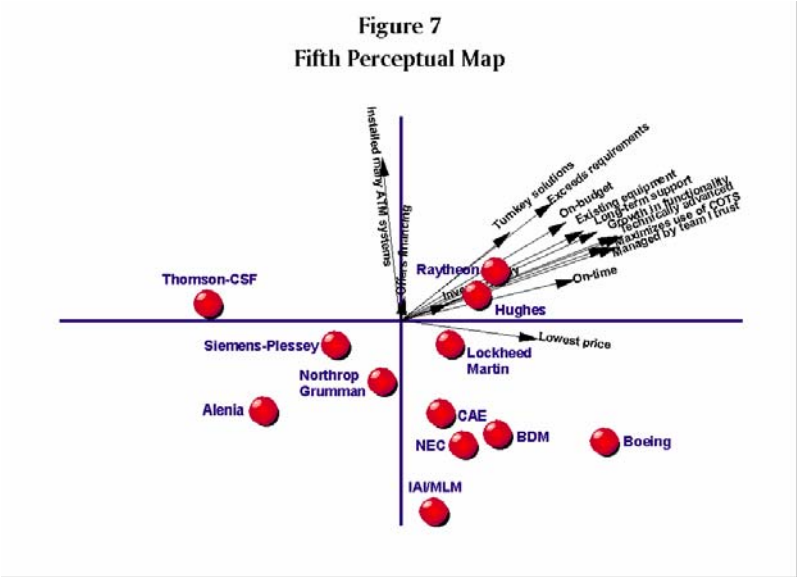
$$\begin{vmatrix} (1-\alpha^2)^{\frac{1}{2}} & -\alpha \\ \alpha & (1-\alpha^2)^{\frac{1}{2}} \end{vmatrix}$$

where α is the cosine of the desired angle of rotation.

The Fourth Principle. Always rotate a map so that the axes are aligned with understandable attributes and so that desirable movement is typically "up" and "to the right." An ideal, measured or hypothesized, should be in upper right quadrant.

The *fifth map*, rotated, is shown below:

Figure 7
Fifth Perceptual Map



At this point, analysis begins. The types of questions that can be readily answered include:

(1) How each company is perceived by each segment and in each region?

(2) What is the difference in perceptions of Hughes between current customers and prospective ones?

(3) What are the differences in perceptions between those who are very familiar with Hughes and those who are less familiar?

(4) Is the pattern of differences in brand familiarity the same for all competitive vendors?

The Fifth Principle: Show all major study findings in the context of a single perceptual space.

REFERENCES

APM System, Version 1, Sawtooth Software, Inc. Sequim, WA, 1987-95.

Fiedler, John A. "A Comparison of Correspondence Analysis and Discriminant Analysis-Based Maps," *American Marketing Association: Advanced Research Techniques Forum.* Beaver Creek, CO, June 9-12, 1996.

Greenacre, Michael J. "The Carroll-Green-Shaffer Scaling in Correspondence Analysis: A Theoretical and Empirical Appraisal," *Journal of Marketing*, 26 (August, 1989), 358-365.

Johnson, Richard M. "Multiple Discriminant Analysis," unpublished paper, "Workshop on Multivariate Methods in Marketing," University of Chicago, 1970.

SPSS Professional Statistics, Release 6.1. SPSS, Inc. Chicago, 1993.

Appendix

SPSS Code for Generating DA-based Perceptual Maps

```
Step 1.
SAVE OUTFILE = "f:\proj\study\temp\tmp_rtgs.sav".

Step 2.
COMPUTE  brand   = 1.
COMPUTE  recnum  = (respnum$ * 100) + brand.
SAVE
  OUTFILE = "f:\proj\study\temp\tmpr01.sav"
 /KEEP    = respnum$ recnum brand r.1.1 to r.1.14
 /RENAME    (r.1.1 to r.1.14 = rate01 TO rate14).

Step 3.
ADD FILES
  FILE = "f:\proj\study\temp\tmpr01.sav"
 /FILE = "f:\proj\study\temp\tmpr02.sav"
                          ..
                          ..
 /FILE = "f:\proj\study\temp\tmprNN.sav".
SORT CASES BY recnum.

After Steps 1-3, the resulting data file looks like this:
1001 01 100101 - - - - - - - - - - - - - -    Respondent 1001
1001 02 100101 2 4 - 5 1 4 - - - 1 2 - 1 2   rates brands 2, 5, and 8
1001 03 100101 - - - - - - - - - - - - - -    on attributes
1001 04 100101 - - - - - - - - - - - - - -    1-2, 4-6, 10-11, 13-14.
1001 05 100101 3 4 - 3 4 1 - - - 4 5 - 3 2
1001 06 100101 - - - - - - - - - - - - - -
1001 07 100101 - - - - - - - - - - - - - -
1001 08 100101 5 5 - 4 5 3 - - - 3 4 - 4 5

1002 01 100101 2 - 4 5 4 - - - 5 3 - 2 - -    Respondent 1002
1002 02 100101 - - - - - - - - - - - - - -   rates brands 1, 5, and 6
1002 03 100101 - - - - - - - - - - - - - -    on attributes
1002 04 100101 - - - - - - - - - - - - - -    1, 3-5, 9-10, 12
1002 05 100101 5 - 3 2 4 - - - 4 3 - 1 - -
1002 06 100101 3 - 5 4 1 - - - 2 3 - 1 - -
1002 07 100101 - - - - - - - - - - - - - -
1002 08 100101 - - - - - - - - - - - - - -

Step 4.
COUNT emptyrec = rate01 to rate14 (1 THRU 5).
SELECT IF (emptyrec > 0).

Step 5.
AGGREGATE
  /OUTFILE  = "f:\proj\study\temp\aggr_rat.sav"
  /PRESORTED
  /BREAK    = respnum$
  /avgrat01 "Avg Rtng Att01" = MEAN(rat01)
        ..            ..          ..
```

```
        ..          ..          ..
  /avgratNN "Avg Rtng AttNN" = MEAN(ratNN).
```

Step 6.

```
MATCH FILES
   FILE     = "f:\proj\study\temp\tmp_rtgs.sav"
   TABLE    = "f:\proj\study\temp\aggr_rat.sav"
  /BY       respnum$
  /MAP.
DO REPEAT
   tmpa = rate01 TO rateNN
  /tmpb = avgrat01 TO avgratNN.
   COMPUTE tmpa = tmpa - tmpb.
END REPEAT.
```

Step 7.

```
RECODE rate01 to ratNN (sysmis = 0).
```

After Steps 5-7, the resulting data file looks like this:

```
1001 02 100101 -1.3 -0.3  0.0  1.0 -2.3  1.3  0.0  0.0  0.0 -1.6 -1.6  0.0 -1.6 -1.0
1001 05 100101 -0.3 -0.3  0.0 -1.0  0.6 -1.6  0.0  0.0  0.0  1.3  1.3  0.0  0.3 -1.0
1001 08 100101  1.6  0.6  0.0  0.0  1.7  0.3  0.0  0.0  0.0  0.3  0.3  0.0  1.3  2.0
1002 01 100101 -1.3  0.0  0.0  1.3  1.0  0.0  0.0  0.0  1.3  0.0  0.0  0.6  0.0  0.0
1002 05 100101  1.6  0.0 -1.0 -1.6  1.0  0.0  0.0  0.0  0.3  0.0  0.0 -0.3  0.0  0.0
1002 06 100101 -0.3  0.0  1.0 -0.3 -2.0  0.0  0.0  0.0 -1.6  0.0  0.0 -0.3  0.0  0.0
```

Step 8.

```
DISCRIMINANT
   GROUPS     = brand (1 YY)
  /VARIABLES  = rate01 to rateNN
  /ANALYSIS   = ALL
  /METHOD     = DIRECT
  /FUNCTIONS  = 2
  /SAVE       = SCORES (dscr)
  /PRIORS     = EQUAL
  /STATISTICS = MEAN STDDEV UNIVF TABLE
  /CLASSIFY   = NONMISSING POOLED .
```

Step 9.

```
CORRELATIONS
   VARIABLES  = rate01 to ratNN
   WITH         dscr1 dscr2
  /MISSING    = PAIRWISE
  /PRINT      = NOSIG.
```

Rotation.

```
* Enter desired angle of rotation (-90 thru +90) in place of zero.
COMPUTE ang_rot = 0.
* Convert angle to radians.
COMPUTE rad_rot = ang_rot * 3.141593 / 180.
* Compute cosine of angle.
COMPUTE alpha   = COS (rad_rot).

COMPUTE d1a     = SQRT (1 - (alpha**2)).
```

```
COMPUTE d1b     = alpha.
COMPUTE d2a     = alpha * -1.
COMPUTE d2b     = SQRT (1 - (alpha**2)).

COMPUTE r_dscr1 = (dscr1 * d1a) + (dscr2 * d1b).
COMPUTE r_dscr2 = (dscr1 * d2a) + (dscr2 * d2b).
```

COMMENT ON WITTENSCHLAEGER AND FIEDLER

Thomas L. Pilon, Ph.D.
TRAC, Inc.

I think John's paper is excellent. I believe the techniques that he has applied and the manner in which he has applied them represent the best current practices in perceptual mapping as his title claims. The purpose of the comments that follow is to supplement his presentation.

WHY USE DISCRIMINANT ANALYSIS TO PRODUCE PERCEPTUAL MAPS?

I agree that discriminant analysis is the best choice for producing perceptual maps. For in-depth discussions of the various approaches to perceptual mapping see Green, Carmone, and Smith (1989), Hair, Anderson, Tatham, and Black (1995), and Pilon (1989, 1992).

CHOOSING WHAT TO RATE

I would like to underline the conclusion and recommendations that were described in this section:

1. Rate more products at the expense of attributes
2. Rate products within attributes.

In my opinion, these recommendations are critical to the success of a perceptual mapping study.

ACHIEVING THE BENEFITS OF APM WITHOUT ITS LIMITATIONS

The nine step algorithm that John outlines will effectively reproduce the results that one can obtain from APM.

One thing that is critical to point out is that the version of SPSS that John was using employs a **canonical** discriminant analysis routine. If one is using a discriminant analysis routine that is not canonical, it is highly recommended that the discriminant analysis be conducted on the principal components of the variables, rather than on the raw variables themselves.

Secondly, while I agree that the APM data collection interface is dated when compared to Ci3, the data analysis portion of APM is still quite easy and useful to employ. As an alternative to nine steps in SPSS, consider converting your data to be compatible with APM and using APM to analyze your data. While the APM file format is not as straightforward as you would hope and APM's respondent weighting capabilities are limited, I expect that many researchers will find that utilizing the APM two-step algorithm will be easier and quicker than using John's nine-step SPSS algorithm.

CONJOINT ANALYSIS & PERCEPTUAL MAPS

Since the majority of papers at this conference have been about conjoint analysis, it seems appropriate to comment on the relationship between conjoint analysis and perceptual mapping.

I like to think of Perceptual Mapping as telling us how we are perceived at the moment and Conjoint Analysis as telling us where we need to be tomorrow. It is important to note that when most researchers build the base case in conjoint simulators, they set up the base case according to how the researcher perceives the products (let's call this "actual reality"). I believe that it is also very useful to set up the base case according to how the market perceives the products (let's call this "perceived reality"). The "actual reality" approach provides insights with respect to required product changes, while the "perceived reality" approach provides insights with respect to required changes in perceptions.

REFERENCES

Green, Paul E., Frank J. Carmone, and Scott M. Smith (1989). *Multidimensional Scaling: Concept and Applications.* Boston: Allyn & Bacon.

Hair, Joseph F. Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1995). *Multivariate Data Analysis.* 4th ed. New Jersey: Prentice-Hall.

Pilon, Thomas L. (1989). "Discriminant versus Factor Based Perceptual Maps: Practical Considerations." *Sawtooth Software Conference Proceedings,* 166-182.

Pilon, Thomas L. (1992). "A Comparison of Results Obtained from Alternative Perceptual Mapping Techniques." *Sawtooth Software Conference Proceedings,* 163-178.

OBTAINING PRODUCT-MARKET MAPS FROM PREFERENCE DATA

Terry Elrod
University of Alberta

ABSTRACT

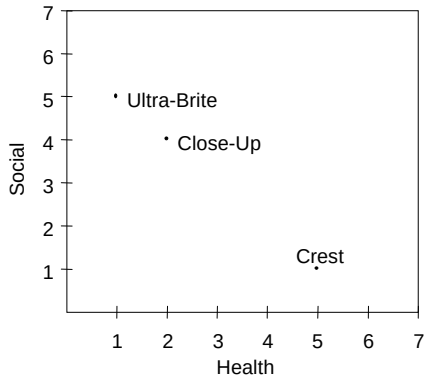
This paper introduces the reader to product-market maps and shows how they can be used to explain and predict brand preference and, ultimately, brand choice. It then considers why it might be desirable to estimate product-market maps from consumer preferences for existing brands. A model for accomplishing this is described and its utility explored using the data analyzed by Wittenschlaeger and Fiedler, whose paper also appears in this volume. A product-market map is fit to pairwise preferences for existing brands obtained from users of air traffic management systems. An additional analysis of brand perceptions assists in interpretation and verification of the map.

AN INTRODUCTION TO PRODUCT-MARKET MAPS

A product-market map uses a picture to characterize both products (i.e. brands) and market (i.e. customers) in terms of the benefits that drive consumer brand preference and choice. Product-market maps are best explained using a simplified example.

Suppose that choice of toothpaste is driven by how the brands are perceived in terms of two fundamental benefits: health and social. And suppose that there are only three brands perceived by consumers as shown in Figure 1. Because this picture portrays only products, it is a product map. You may think of this picture as simply a plot of the three brands in terms of their average ratings on these two benefits using a seven-point ratings scale. In this hypothetical example, Crest enjoys a strong perception in terms of the health benefit, but is weak on the social benefit. Ultra-Brite and Close-Up are more similar to each other than either is to Crest—both are relatively strong in terms of the social benefit but weaker in terms of the health benefit. The position of a brand in a product map signifies how much of each benefit it delivers as perceived by consumers.

Figure 1. A Hypothetical Product Map



A product-market map adds consumers to a product map. Two such segments are added to the product map of Figure 1 to yield the hypothetical product-market map of Figure 2. The location of the two segments in the map reflects how each segment uses each benefit to determine its preferences for the brands.

Determining Brand Preferences from a Product-Market Map

There are two common models for relating product-market maps to brand preferences: the ideal-point model and the vector model. The ideal-point model assumes that the location of a consumer segment in the map represents the consumer's "ideal brand" in terms of the benefits underlying the product category. Brand preference is inversely related to each brand's distance from the consumer's ideal point. The distances underlying brand preference for the teenager segment of our hypothetical example are shown in Figure 3.

Figure 2. A Hypothetical Product-Market Map

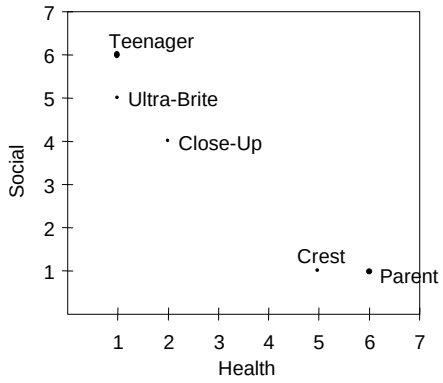
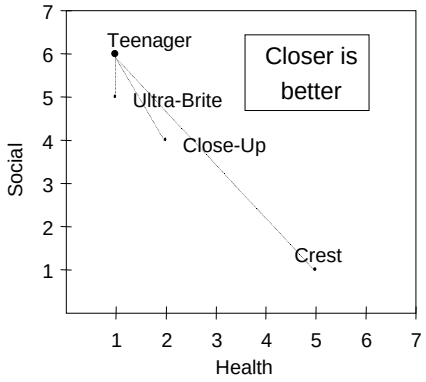
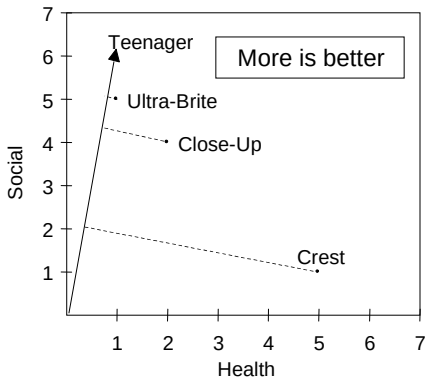


Figure 3. Ideal-Point Model Illustrated for the Teenager Segment



The vector model assumes that more of a product benefit is always better, although consumer segments still differ in terms of how much importance they attach to each benefit. The vector model for the teenager segment is illustrated in Figure 4. Because the vector model underlies the product-market map described later in this paper, we will examine its properties more fully.

Figure 4. Vector Model Illustrated for the Teenager Segment

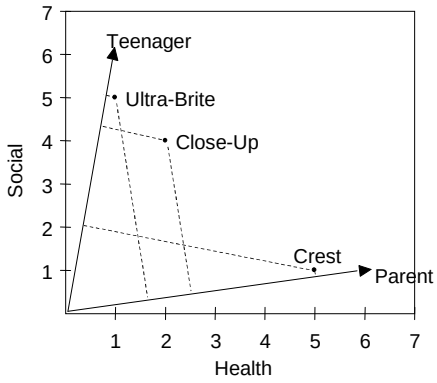


In a vector model, each consumer segment can be represented by an arrow originating at the origin of the map and ending at the location of the segment in the map. The arrow emphasizes the relative importance of the two benefits to the segment. In a vector model, brands that lie farthest in the direction indicated are most preferred, while distance of a brand *from* the arrow is irrelevant. The brand preferences are proportional to the projection (at a right angle) of the brands onto the arrow. These projections for the teenager segment are also indicated in Figure 4. They indicate that Ultra-Brite is most preferred, but Close-Up is a close second. Crest, with its poor perceived performance in terms of the social benefit, is a distant last preference for this segment.

Using the Product-Market Map

Figure 5 portrays the hypothetical product-market map for both segments, as well as the implied preferences for all brands and segments using the vector model. This product-market map, simple as it is, can be used to illustrate much about the importance of market segmentation, product differentiation, and the intimate connection between them.

Figure 5. The Brand Preferences for All Brands and Segments Implied by the Vector Model



Crest is much preferred by the parent segment, which means that Crest enjoys a near-monopolistic position vis-à-vis this group. This is due to Crest being the only brand that is tailored to the greater importance the parent segment places upon the health benefit. Its poor perceived social benefit is not important to this segment. However it is also apparent that Crest has no prospect of attracting significant sales from the teenager segment. Its marketing should therefore be directed to the parent segment. It may be priced at a premium and still be preferred by this segment.

While Ultra-Brite is the preferred brand for the teenager segment, Close-Up is a close second. Aggressive pricing and advertising by Close-Up may suffice to attract significant sales from the teenager segment, and this prospect prevents Ultra-Brite from enjoying large profit margins. Neither brand can hope to attract appreciable sales from the parent segment.

A Tabular Representation of the Toothpaste Example

All of the information in the product-market map of Figure 5 can also be shown in tabular form as in Table 1. The first two columns of numbers in the table show the locations of the brands and segments in the map. The last two columns show the brand preferences for both segments as implied by this vector map. The preference value of the parent segment for Crest is obtained as the sum of the benefits of Crest weighted by the importances the parent segment attaches to these benefits: i.e. $5 \times 6 + 1 \times 1 = 31$. This is a straightforward application of the multi-attribute utility model familiar to marketers.

Table 1. Tabular Representation of the Toothpaste Product-Market Map

	Product benefits		Brand preferences	
	Health	Social	Parent	Teenager
Crest	5	1	31	11
Close-Up	2	4	16	26
Ultra-Brite	1	5	11	31
Parent	6	1	®	
Teenager	1	6		®

ESTIMATING PRODUCT-MARKET MAPS FROM PREFERENCE DATA

So far we have considered product-market maps and how they may be related to brand preferences without considering how such maps may be obtained. Obtaining meaningful product-market maps is a nontrivial exercise because they represent brand *perceptions* in terms of the product *benefits* that underlie brand preference and choice. Neither consumer perceptions nor the product benefits underlying preference are directly observable.

There are three primary methodologies for obtaining product-market maps. The oldest is simply to have consumers provide a product-market map directly by rating the brands and themselves in terms of benefits specified by the researcher in advance. Because fundamental benefits are intangible and rating scales somewhat artificial, this approach tends not to yield product-market maps that predict choices well.

Two other methods exploit the redundancy among the three types of data shown in Table 1: brand perceptions, consumer importances for benefits, and brand preferences. Knowing any two of these types of data allows calculation of the third by application of the multi-attribute utility model. For example, suppose that you have obtained the product part of the product-market map and that you have also collected information from consumers about their preferences for the brands. What is known and is not known to you under this scenario is shown in Table 2.

Table 2. Unknown Consumer Importances for Toothpaste Benefits

	Product benefits		Brand preferences	
	Health	Social	Parent	Teenager
Crest	5	1	31	11
Close-Up	2	4	16	26
Ultra-Brite	1	5	11	31
Parent	?	?	⑧	
Teenager	?	?		⑧

Because of the redundancy of information in Table 2 it is possible to estimate the importance each segment attaches to each benefit. This may be estimated separately for each segment using regression analysis. The dependent variable would be, for example, parent preferences for the three brands, and the two independent variables would be the values each of the three brands have on the two benefits.

With only three brands there are only three observations, so the regression estimates will not be very stable. I will discuss a method for stabilizing these estimates later, but clearly having more than three brands in a market would help to obtain more reliable estimates of the consumer locations in product-market map.

The article by Wittenschlaeger and Fiedler that also appears in this volume provides an excellent demonstration of how a product map might be obtained as the first step towards developing a product-market map. Their product map could serve to determine both the benefits that

underlie brand preference as well as the locations of the brands in terms of these benefits. The regression analysis I just described could then be performed as a second procedure. Obtaining a product-market map by analyzing brand preferences using a given product map is known as an "external" analysis of preferences. The word external refers to the fact that the product map was obtained in advance using other information.

Table 3. Obtaining Both Brand and Consumer Locations from Brand Preferences

	Product benefits		Brand preferences	
	Health	Social	Parent	Teenager
Crest	?	?	31	11
Close-Up	?	?	16	26
Ultra-Brite	?	?	11	31
Parent	?	?	⑧	
Teenager	?	?		⑧

In this paper I illustrate what is known as an "internal" analysis of preference data. The goal of the analysis is to obtain both product and consumer segment locations in a product-market map simultaneously, and only by using brand preference information. This is a more ambitious task. A glance at Table 3 shows how much we seek to estimate from the brand preference data alone.

A Model for Obtaining a Product-Market Map from APM Preferences

This paper will employ the vector model to estimate a product-market map from brand preferences. Ideal-point product-market maps are often very difficult to estimate from brand preferences. Ideal-point models are more general than vector models. However if a vector model accounts adequately for brand preferences, as is often the case, then the data contain little information to allow estimation of the additional generality of the ideal-point model.

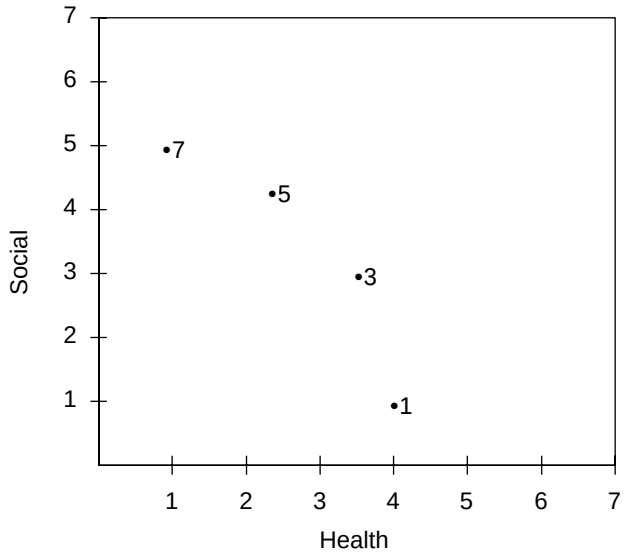
Flexibility of the Vector Model

Note that the vector model assumes more is better when it comes to *benefits* revealed by the model, but not necessarily for the brand *attributes* that characterize the brands. This distinction is important and can be illustrated using our toothpaste example.

I recall an issue of *Consumers Report* some years ago that contained a review article for toothpastes which stated that the primary determinant of the tooth-whitening ability of a toothpaste is its abrasive content. Abrasives help to remove stains from teeth (as well as plaque), but too much abrasive content can accelerate the wearing away of tooth enamel.

If consumers believe that a toothpaste can have too much tooth-whitening power, then this attribute would be nonlinearly related to the health and social benefits. Figure 6 illustrates this by showing four points of a 7-point scale that relates a brand's perceived tooth-whitening ability to its location in the map. A toothpaste with a rating of 1 has virtually no tooth-whitening ability. I postulate that such a brand would have a low rating on the social benefit of toothpaste (assuming that some tooth-whitening ability is deemed essential to this benefit), while I place the brand (arbitrarily) at slightly above the midpoint of the scale for the health benefit.

Figure 6. Hypothesized Relationship Between Tooth-Whitening Ability and the Benefits of Toothpastes



Increasing this brand's rating to 3 on tooth-whitening ability helps its performance on the social benefit with little harm to its perceive health benefit. Increasing its perceived tooth-whitening rating further, however, yields diminishing improvement on the social benefit while harming the brand's perceived health benefit.

Consumer segments which attach approximately equal importance to the health and social benefits of toothpastes (and who would be represented in Figure 6 by a vector pointing to the upper-right corner) would prefer toothpastes with tooth-whitening ratings in the range of 3 to 5 on the 7-point scale. This prediction is consistent with an ideal-point model relating brand preference to brand attributes even though the model used to estimate the product-market map from preferences is vector-based. Extreme segments, such as the parent and teenager segments of the example, might still prefer extreme ends of the tooth-whitening scale.

This discussion of tooth-whitening ability and how it may be nonlinearly related to the benefits underlying product-market maps serves to indicate the flexibility possessed by such maps based on the vector model when the maps are estimated from brand preferences. Given the flexibility of such models, it is not surprising that the additional generality of ideal-point models of product-market maps often makes them hard to estimate reliably from preference data.

Particulars of the Model Employed

The method developed for this paper to derive a product-market map from preference data is a variant of factor analysis. Factor analysis is familiar to many marketers. The output of a factor analysis invariably shows how the variables included in the analysis are related to the two or so factors estimated by the analysis. When analyzing a data matrix of brand preferences, with as many rows as consumers and one column for each brand, the "variables" are the brands and the "factors" are the benefits. Hence the result of a factor analysis applied to such data is a product map. However factor analysis also estimates "factor scores," one for each respondent. Factor scores characterize the respondents in terms of the same factors. They are the coefficients for the respondents that, together with the product map, best reproduce the pattern in the preference data. Thus the factor scores are also the importance weights for the individual respondents.

For the purpose of analyzing consumer preference or choice data to obtain a product-market map, I have adapted factor analysis in four respects to better accommodate this particular marketing application. The remainder of this section provides a brief description of these differences.

(1) A factor analysis usually begins by standardizing the data. The raw data are rescaled so that the mean for each column is zero and the standard deviation is one. It makes sense to do this when some of the variables included in the factor analysis differ from other variables in their units of measurement. However in this setting every variable is a measure of preference and the different columns simply refer to different brands. Differences in average preference across brands is vital information that is retained and accounted for by the analysis.

(2) Factor analysis assumes that the different variables may be measured with different amounts of error. This also makes sense when the variables are measured on different scales or ask fundamentally different questions. However here each variable is an expression of brand preference on the same scale, and the only difference between questions is the brand being rated. Therefore I have not allowed the error variances to differ arbitrarily from one brand to the next.

(3) Factor analysis is often estimated by maximum likelihood assuming that the factor scores have a multivariate normal distribution across respondents. Using a distribution such as the multivariate normal to characterize consumer heterogeneity is a good idea for two reasons. First, the respondents invariably represent a sample from the population of consumers. When a *sample* has been taken in order to learn about the *population* from which the sample has been taken, then it is appropriate that the analysis explicitly recognize this fact. This principle is widely overlooked in marketing research. Estimating factor scores at the individual-level for each respondent (or conjoint part worths, for that matter) is simply incorrect. The analogous error in an analysis of variance would be to model random effects as if they were fixed effects.

   A second reason to use a statistical distribution such as the multivariate normal to characterize customer heterogeneity is that it ameliorates the problem of trying to estimate too many coefficients from too few data. To illustrate the economy that results, assume for the moment that we have a two-dimensional map and 300 respondents. Estimating importance vectors separately for each respondent requires the estimation of 600 parameters.

Estimating the mean and variance-covariance matrix for a bivariate normal distribution, on the other hand, requires estimation of only 5 parameters.

Nevertheless, the assumption of multivariate normality for consumer importances is a strong assumption. I have relaxed this assumption in two ways. First, the multivariate normal distribution has been replaced by the more robust and general multivariate t distribution. The t distribution has "longer tails" than the normal, so it is more robust to outlying respondents. The degrees of freedom is estimated along with the other unknowns of the product-market map. Second, rather than assume that the distribution applies to all respondents, I have only assumed that it applies *within* each segment. Thus consumer heterogeneity within segments is explicitly accounted for. Because of the high degree of indeterminacy in estimating product-market maps from brand preferences, it is possible without loss of generality to scale the maps so that consumers within each segment have independent standard t-distributions about the segment mean. This is convenient because it allows us to represent each segment by its mean alone, without also having to portray consumer heterogeneity within each of the segments.

(4) A final extension to factor analysis is particular to the type of preference data obtained by Sawtooth Software's APM. APM does not provide data on brand preferences as I have described them. Rather, APM obtains *pairwise* preferences using a 100-point "probability of purchase" scale. Thus we do not observe brand preferences directly, but a measure of the difference in preference for pairs of brands. Obtaining product-market maps from pairwise preferences involves additional programming but is not conceptually difficult.

A natural method for analyzing probabilities is to transform them into logits. However stated probabilities can include the endpoints of the scale, and a literal logit transformation of these values is not possible. I have implemented a capability of estimating the best increment to add to the endpoints of the 0-100 interval before applying the logit transformation.

DESCRIPTION OF THE ATM DATA

Thomas A. Wittenschlaeger and John A. Fiedler were kind enough to share with me the data used in their paper. The data pertain to suppliers of air traffic management (ATM) systems.

The study included 12 suppliers, but one company was unfamiliar to all but a few respondents. Because APM only asks for preference judgments involving brands familiar to each respondent, there was little preference information available for this brand—too little information to allow reliable estimation of its location in a product-market map based on preference data. This company was therefore not included in my analysis. Complete data for 11 companies and 14 attributes was available for 292 respondents, all of whom were included in the analyses described below.

Customer Region

The distribution of the 292 respondents among four regions of the world is shown in Table 4. I created the Other region by combining the small numbers of respondents from three regions: the rest of the Americas, Russia, CIS and Eastern Europe and Asia. Regions with few respondents were combined so that every segment's position in the product-market map would be reliably estimated.

Table 4. Region Segment Definitions

Label	Explanation
U	Canada, U.S.
W	Western Europe
M	Middle East and Africa
O	Other

Company Familiarity

Every respondent also provided information about company familiarity using a 5-point scale. The 11 companies included in the analysis, and their average familiarity ratings, are shown in Table 5. (The excluded company had an average familiarity rating of only 1.34.)

Table 5. The Twelve Companies and Their Average Familiarity Ratings

Label	Average Rating
HG	3.61
RY	3.55
BO	3.44
LK	3.23
NR	2.95
SM	2.93
TM	2.89
NE	2.43
AL	2.37
CA	2.00
BD	1.99

Attribute Importance

We have complete ratings of attribute importance for all 14 attributes and all 292 respondents. The attribute definitions are provided in Table 6 along with the average importance rating for each attribute. These data will not be used to derive the product-market map, but will be referred to later when assessing the face validity of the map.

Company Perceptions

Table 6. Definitions and Average Importance Ratings for 14 Attributes

Label	Explanation	Rating
ONT	Delivers on-time	4.69
ONB	Delivers on-budget	4.64
TRS	Is managed by a team I trust	4.60
LON	Provides long-term life cycle system support	4.53
GRW	Provides growth in functionality and capacity	4.53
ADV	Provides technically advanced solutions	4.46
ADP	Provides solutions that can adapt /accommodate to existing equipment	4.40
TUR	Offers turnkey solutions	4.13
COT	Maximizes use of commercial off-the-shelf products	4.12
MAN	Has installed many ATM systems	4.08
EXC	Offers products which exceed requirements	3.79
LWS	Offers the lowest price	3.64
LOC	Invests in local industry / economy	3.23
FIN	Is able to offer financing packages	3.01

In addition, respondents provided partial information about their perceptions of the companies in terms of the attributes. Complete data for all companies and attributes would require $12 \times 14 = 168$ ratings from every respondent, which is too onerous a task and in any case ratings for unfamiliar brands or on unimportant attributes are likely to have little reliability. Therefore Sawtooth Software's APM collects perceptions only for those companies familiar to the respondent and only on those attributes of importance to him or her. Wittenschlaeger and Fiedler used these data to develop a product map. These data will be used here only to interpret the product-market map and assess its face validity.

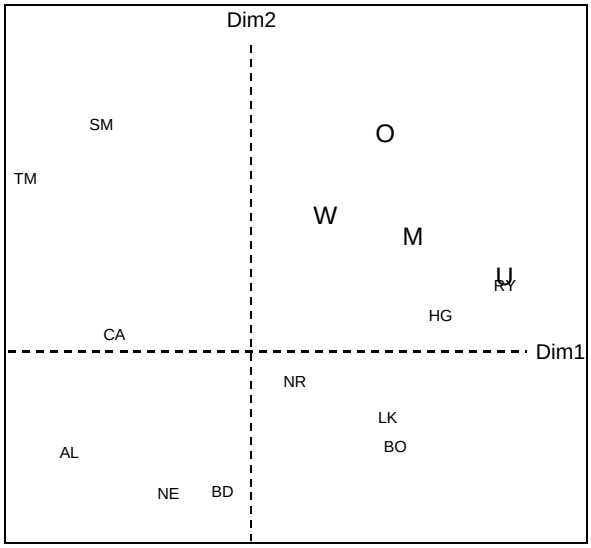
ATM Preference Data

Finally, respondents indicate their relative preferences for each of several pairs of companies. The companies are simply identified by name in this task.

I fit the vector model to a logit transformation of the pairwise probabilities after adding an estimated increment of at least 5 to both ends of the 0-100 interval. (That is, the interval used to transform the probabilities was forced to be at least –5 to 105). This in effect forces the odds of choosing the less preferred alternative to be greater than or equal to 1:21.

The choice of 5 as a minimum size for the increment was somewhat arbitrary. An increment of 10 was found to be optimal when using the multivariate normal distribution to represent within-segment heterogeneity. However, replacing the multivariate normal distribution with the multivariate t improves model fit dramatically while requiring estimation of only one additional parameter. Estimating an unconstrained increment for the logit transformation and the degrees of freedom for the t distribution simultaneously for these data led to no increment for the logit but degrees of freedom for the t distribution so small that the estimation procedure became unstable. Constraining the increment to be at least 5 led to a larger estimate for degrees of freedom and stable estimates of the map. The resultant map is shown in Figure 7, where both axes are to the same scale.

Figure 7. A Two-Dimensional Map of Preferences



Interpreting the Map Using Brand Perceptions

While the map of Figure 7 shows both the companies and customer segments, it contains no information that allows us to interpret the map in terms of the 14 attributes included in the study (Table 6). Such an interpretation can be added to the map using a secondary analysis known as "property fitting."

First, a table is created with as many rows as brands and as many columns as attributes. Each cell of this table shows the average rating received by the brand of that row on the attribute of that column. This newly created table is then related to the map one attribute at a time. Just how this is done is illustrated for the ADV attribute in Table 7. There we show the average ratings on ADV for the 11 brands together with the locations of these brands in the map. The relationship

between ADV and the two other columns was then determined using regression analysis where the columns Dim1 and Dim2 are the independent variables in the regression.

Table 7. Relating ADV to the Map

	ADV	Dim1	Dim2
AL	3.28	-0.35	-0.18
BD	3.49	-0.05	-0.26
BO	3.69	0.29	-0.17
CA	3.43	-0.26	0.05
HG	3.81	0.38	0.09
LK	3.78	0.27	-0.11
NE	3.59	-0.16	-0.26
NR	3.66	0.09	-0.04
RY	3.91	0.50	0.15
SM	3.48	-0.29	0.46
TM	3.47	-0.43	0.36

The degree of success when performing this regression analysis separately for each of the 14 attributes in the study is shown in Table 8. Although each regression is based on only 11 observations, all but two of the regressions were statistically significant. The statistically insignificant regressions were for FIN and LOC. Table 6 shows that these two attributes were rated by respondents as being least important of all, so these attributes should in fact be nearly irrelevant to brand preferences and unrelated to a map that explains these preferences.

Table 8. Relating Attributes to the Map

	R^2	P-value
ONT	0.82	0.00
ONB	0.62	0.02
TRS	0.95	0.00
LON	0.92	0.00
GRW	0.86	0.00
ADV	0.91	0.00
ADP	0.76	0.00
TUR	0.74	0.00
COT	0.60	0.02
MAN	0.58	0.03
EXC	0.88	0.00
LWS	0.65	0.01
LOC	0.39	0.14
FIN	0.15	0.53

The method of property fitting just described assumed that the attributes are linearly related to the benefits of the preference map. We have seen that attributes need not be linearly related to product benefits, but map interpretation is simplified when they are. I report here two analyses which indicate that the linear assumption is appropriate for relating attributes to benefits for these data.

The first check is to fit a quadratic regression of average brand perceptions to the benefit dimensions of the map. That is, three additional independent variables can be added to the regression shown in, corresponding to Dim1^2, Dim2^2 and Dim1*Dim2. These added terms failed to improve upon the vector model to a statistically significant extent for any of the 14 attributes. This is not surprising given that only 11 observations are available for each regression.

A second check makes use of additional information which is often collected by APM. Respondents provided perceptions of their ideal brand along with actual brands. A basis for deciding the appropriateness of a linear relationship between attributes and benefits is to compare the ratings given the ideal brand to the ratings given actual brands. If ratings for actual brands rarely straddle the ideal brand rating on an attribute, then this is a further indication that a linear relationship between the attribute and product benefits is appropriate.

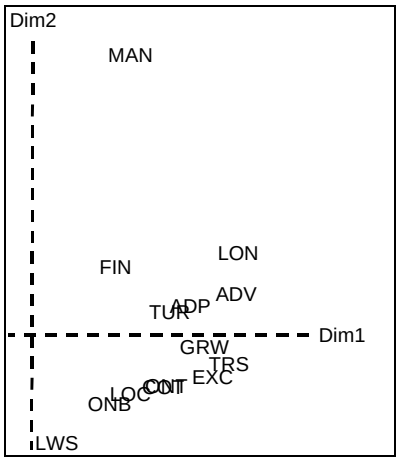
I calculated for each respondent and attribute the ratings given the actual brands minus the rating given the ideal brand. Since the actual and ideal brands are rated on the same 5-point scale, the difference must always be an integer between 4 and –4. Over all respondents, brands and attributes only 5.7% of the differences were positive (4.3% by only one unit). This is not a lot of "straddling" of the ideal point.

Displaying the Relation of Attributes to Map Benefits

Portrayal of attributes in the map is the same as for segments and companies: the coefficients from the regression for each attribute provide its location in the map. Often attributes are shown as vectors radiating from the origin, but because the arrows can obscure other information, I simply plot the attributes as points. A display of the attributes as they relate to the dimensions of the map is shown in Figure 8. (The two attributes that coincide in Figure 8 are ONT and COT.)

The acronyms used for attributes are as shown in Table 6. Dim2 in the map distinguishes companies that "Have installed many ATM systems" (MAN) from "Offers the lowest price" (LWS). All other attributes are more closely associated with Dim1 in the positive direction. Attributes lying to the upper right seem to pertain to companies that best provide a customizable offering with substantial support. Examples are "Provides long-term life cycle support" (LON), "Provides technically advanced solutions" (ADV), "Provides solutions that can adapt / accommodate to existing equipment" (ADP), and "Offers turnkey solutions" (TUR). In contrast, attributes lying towards the lower right pertain to companies that offer a more standardized product, with the predictability that this allows: "Delivers on-budget" (ONB), "Delivers on-time" (ONT), and "Maximizes use of commercial off the shelf products" (COT). (Recall that LOC is not statistically significant.)

Figure 8. Relation of the Attributes to the Map



Putting it All Together: The Final Product-Market Map

Figure 7 and Figure 8 are combined into the single (busy!) product-market map shown as Figure 9. The distance of the attributes from the origin was reduced by the same fraction for all attributes and dimensions so that they would fit better into the display. The display is made somewhat easier to read by using the convention that the customer segments are denoted by single letters (cf. Table 4), the companies by two letters (cf. Table 5), and the attributes by three letters (cf. Table 6).

CONCLUDING COMMENTS

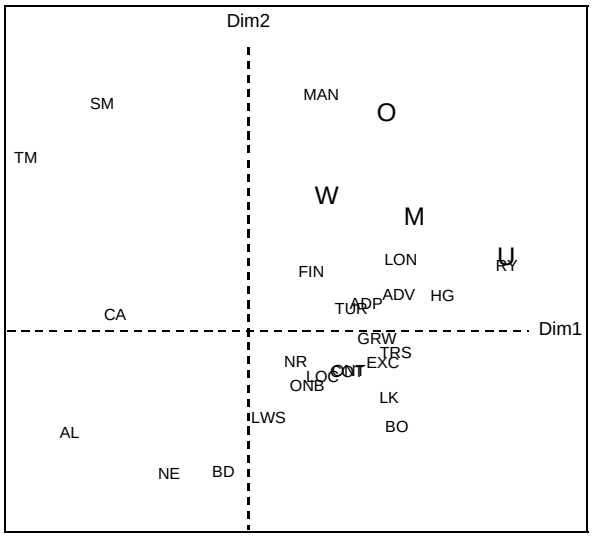
This paper has illustrated a method for estimating a product-market map from pairwise preferences for existing brands, such as is obtained using APM. Map interpretation was aided by regressing average brand perceptions onto the map.

The final map displays 4 customer segments, 14 attributes and 11 companies. In practice the map might be simplified for some purposes, perhaps by replacing the 14 attributes with descriptive labels of the map's dimensions that are based on these attributes.

Because the product-market map is based upon an analysis of customer preferences, it remains closely tied to these preferences in a quantitative sense and this property should be exploited. Simulators can be built using a spreadsheet software package to predict shares for all brands to help assess contemplated new or repositioned brands. An adequate discussion of the details on how to do this must await a separate paper.

Given the strategic value of the information provided by product-market maps, companies that make good use of this technology can expect to enjoy an important advantage over their competitors.

Figure 9. The Final Product-Market Map for the ATM Data



COMMENT ON ELROD

Richard M. Johnson
Sawtooth Software, Inc.

I want to thank Terry Elrod for his interesting work on maps based on preference data. I think his presentation was remarkable clear as a summary of a very complex technique, and I also think it was quite an impressive achievement. To explain that last statement, I need to provide some perspective on the difference between Terry's approach and that of John Fiedler. I'll also make a suggestion for a future development in mapping.

In the early '60s, mathematical psychologists developed theories about how perceptions and preferences might be related. They considered objects to be arranged in some kind of perceptual space, determined by how people perceived them on descriptive attributes. Each individual was also thought to have an ideal point in the space, or an ideal direction, and to prefer objects that were closer to that point, or farther out in that ideal direction. Several kinds of relationships were implied by those spaces, including:

Attribute ratings should be reproduced by the space, in the sense that an object perceived to be higher than others on an attribute should have a corresponding position with respect to that attribute vector in the space.

Preferences should be predictable from relationships between object positions and representations of individual desires.

Attribute ratings have been used to develop product spaces using factor analysis, discriminant analysis, and correspondence analysis. Attribute-based spaces have been extremely popular in marketing research, perhaps because attribute ratings have many uses beyond making perceptual maps.

Preference judgements have also been used to develop product spaces in marketing research, though less often. The first techniques for making maps based on preferences, in the late '50s and early '60s, were Coombs' "Unfolding" Method (which assumed each individual had an "ideal point") and Tucker's "Points of View" approach (which assumed each individual had a preferred direction). I've always thought that these simple theories relating product attributes and preferences were quite beautiful, and they have had an important role in shaping my professional life.

But, there's been a problem: On one hand, spaces like John showed us, based on attribute ratings, are easy to interpret and good at conveying insights but they are not very good at predicting individual preferences. If you estimate an individual's ideal point in an aggregate perceptual map based on attribute ratings, usually that individual ***won't*** prefer the closest product. It may be that individuals' perceptual spaces are so different from

one another that their average does not capture any of them well enough. Our APM software uses a separate simulation module to predict preferences based on each respondent's own product perceptions; but the aggregate maps displayed by APM do not take advantage of preference data, and APM's simulator doesn't use the aggregate map in any way.

On the other hand, spaces like Terry showed us, based on preference data, are better at accounting for preferences, but because they contain less information about product attributes they can be difficult to interpret and may be less useful for producing insights. This may be a serious disadvantage, since one important use of product maps is to help managers visualize strategic opportunities.

I had considered the possibility that maps based on attribute data might never account for preferences very well, and the maps based on preferences might never look very much like those based on attributes. But, to my surprise and delight, the final map that John Fiedler produced from attribute data and the map that Terry produced using only preference data from the same study are strikingly similar. Whether they are similar enough to be considered equivalent is more than we can say from the information available to the reader, but this is certainly a promising development and suggests that the underlying theory relating attributes and preferences may be correct.

However, I think we need a new kind of mapping approach that combines the features of what John and Terry have presented. I'd like to see a technique that maximizes fit to **both** ratings of products on attributes and individual preferences, using a rigorous likelihood criterion such as that which Terry employed. Terry has told me that he considers that a difficult and perhaps impossible thing to accomplish, but I have hope that Terry will produce such a thing for us one day soon.

INTEGRATED CHOICE LIKELIHOOD (ICL) MODEL

Carl T. Finkbeiner
National Analysts, Inc.

ABSTRACT

A unified model (the ICL model) is presented which subsumes many of the popular preference decomposition models in current use, including ordinary full-profile conjoint analysis, discrete choice analysis, adaptive conjoint analysis, and hybrid conjoint analysis. The ICL model integrates self-explicated ratings of attributes, full- (or partial-) profile conjoint choice likelihood ratings, and choice or constant sum ratings. It is an internally consistent model, allowing great flexibility in the mix of the three different data types and in the number of attributes, attribute levels, or products to which an individual respondent is exposed. The model is estimated for individual respondents, with an individual-level choice model as an integral part, so that choice simulation is logically consistent—obtained using a component of the full model. This choice model component is non IIA, a special case of multinomial probit. The incorporation of points-of-view or latent classes in the model stabilizes the individual-level models. Maximum likelihood estimation is used to estimate parameters. The flexible specification of the ICL model makes it ideally suited to computerized interviewing with respondent-tailored data collection.

Note: the notation used in this paper is complex and, for ease of reference, the notation has been collected into two sections at the end of the paper: **Data Notation** and **Parameter Notation**.

INTRODUCTION

Among the most useful tools in market research are the preference decomposition methods which allow:

- The prediction of the likelihood of respondents choosing products in various configurations of attributes (often called *choice shares*).
- The assessment of the separate impacts of those product attributes.

These methods include ordinary conjoint analysis (Green & Wind, 1975), discrete choice modeling (Louviere & Woodworth, 1983), trade-off analysis (Johnson, 1974), and adaptive conjoint analysis (Johnson, 1987b), among others.

It is characteristic of all methods that they represent a product as a combination of separate attributes, each attribute having varying levels. In essence, the concept of these methods is that a controlled experiment is carried out with likelihood of choice being the dependent variable, so that a model can be developed for predicting choice of *any* product configuration and so that attribute effects can be measured.

In some models, the attributes may be treated as either continuous or categorical in their levels. In each approach, a series of attribute combinations is evaluated:

- Sometimes all attributes are included, as in the full-profile approach (common in ordinary conjoint analysis and discrete choice modeling), and sometimes only a subset are included, as in the partial profile approach (common in trade-off analysis and adaptive conjoint analysis).
- Sometimes a respondent evaluates all attribute combinations being included in the study (common in ordinary conjoint analysis and trade-off analysis) and sometimes each respondent only sees a subset of the combinations (common in discrete choice modeling and adaptive conjoint analysis).
- Sometimes large numbers (if not all) of the respondents see the same set of attribute combinations (common in ordinary conjoint analysis, discrete choice modeling, and tradeoff analysis) and sometimes every respondent sees a different set (common in discrete choice modeling and adaptive conjoint analysis).

The particular approach used is a function of the model underlying the method and of practical considerations of cost and respondent burden.

What follows is a brief description of some of the main variations of decompositional preference models and an experience-based evaluation of them, followed by a recommendation as to the most useful. I will discuss five dimensions on which the methods tend to differ: (i) kind of task, (ii) integration of choice model, (iii) type of choice model, (iv) estimation method, and (v) level of aggregation.

Kind of Task

The different kinds of tasks depend to some extent on the model underlying the method. However, they can be generally classified as follows.

- Scale ratings of choice likelihoods for products (either for each product separately or for preferences between products)
- Rank orders of attractiveness of products
- Scale ratings of attribute levels, especially choice likelihoods for products with each level

- Scale ratings of the importance of attributes
- Choice or constant sum ratings among a set of products

In general, rank orders have not been found necessary: scale ratings provide equivalent or superior information about choice likelihood and are easier for the respondent to generate when there are many things being rated (Leigh, *et al*, 1984). Trade-off analysis, which traditionally relied heavily on rank orders, has probably lost some popularity simply on this account.

Attribute level ratings carry some information about the attractiveness of levels relative to other levels within the same attribute. However, it is not clear how to integrate attribute level ratings to produce reliable predictions, as evidenced by the fact that studies usually show that conjoint analysis is more valid than attribute level ratings (e.g., Leigh, *et al*, 1984; Green, *et al*, 1991).

Importance ratings are strongly correlated with attribute level choice likelihoods. We have found these ratings to be useful adjuncts when choice likelihoods of attribute levels are not present, e.g., in customer satisfaction studies with ratings of satisfaction on each of a product's attributes (Cooper & Finkbeiner, 1984; Finkbeiner, 1992). Adaptive conjoint analysis makes use of rough importance ratings as a starting point for its partworth calculations (Johnson, 1987a) because the attribute level data captured by that method are only rankings and do not convey adequate information about the scale of each attribute. However, given the strong correlation noted above, we find that importance ratings are redundant and can usually be avoided in the present modeling contexts.

Simple choice tasks generally produce weaker data than scale ratings: a single choice between two products conveys nothing about the respondent's *degree* of preference. Of course, if enough choices are obtained either from the individual or across many respondents, simple choice data can produce quite reliable models (Oliphant, *et al*, 1992).

A somewhat more general type of choice task is the constant sum rating among product alternatives in which the respondent, by some device, assigns a relative frequency of choosing each product in a set. For example, we may ask a respondent to divide 100 points among each of three products to reflect their likelihood of choosing each in the future. Simple choice may be thought of as a special case of constant sums in which the relative frequency for the chosen product is 1.0 and is 0.0 for all other products. However, unlike simple choice, constant sum ratings do convey something about degree of preference and so provide stronger data than simple choice. Simple choice and constant sum tasks carry a lot of face credibility since they seem to more closely mimic product choice behavior in the marketplace.

There is some evidence that combining different types of data in a single model may be beneficial. Huber, *et al* (1993) demonstrate an improvement in validity for choice prediction when choice likelihood ratings are combined with attribute level and impor-

tance ratings. They also showed that the attribute level and importance ratings act as a beneficial warm-up for the choice likelihood ratings.

Integration of Choice Model

All of these types of models estimate coefficients for each attribute or attribute level. These coefficients are variously called "regression coefficients," or "partworths," or "utilities."

There are two general categories of methods with respect to these coefficients: those which estimate attribute coefficients directly from choice as the dependent variable (e.g., discrete choice modeling), and those which first estimate the coefficients using choice likelihood (or utility) of the product as the dependent variable and then add choice prediction as a transformation of the product utilities (e.g., ordinary conjoint with a first-choice model attached as a choice share simulator).

Parenthetically, we note that these models are generally specified as linear and additive in the attributes, but that some of the models may be modified to include non-linear interaction terms (ordinary conjoint and discrete choice modeling).

Type of Choice Model

The types of choice models used (either as an integral part of the model or as an addendum) include:

- First-choice—product with highest utility has probability of choice equal to 1.0 (Huber & Moore, 1979; Thurstone, 1945)
- Bradley-Terry-Luce—probability of choice for each product is the product's utility divided by the sum of utilities (Luce, 1959; Suppes & Zinnes, 1963)
- Multinomial logit—product utilities have an independent extreme-value distribution so that the probability of choice for each product is the exponential of the product's utility divided by the sum of the exponentials of utilities (McFadden, 1973; Louviere, 1988)
- Multinomial probit—product utilities have a multinormal distribution with product dependencies explicitly part of the model so that probability of choice for a product is the integral over that region of the multinormal distribution in which that product's utility is greatest (Daganzo, 1979; Finkbeiner, 1986)

The distressing aspect of the Bradley-Terry-Luce and multinomial logit models is that they are subject to the so-called Independence of Irrelevant Alternatives (IIA) property in which a new product's share (or choice probability) is drawn from existing products in proportion to the existing products' previous shares. Generally, in the marketplace, a new product will draw its share disproportionately from those existing products which it is most like, a property shared by the first-choice and multinomial probit models. The

Bradley-Terry-Luce and multinomial logit models when developed and applied separately to each individual respondent (as is done in the simulator provided with the adaptive conjoint analysis software), produce share estimates in aggregate that do not generally have the IIA property.

However, this use of those models at the individual level still does not go far enough. If product C, identical to product A, is added to a two product set containing A and B, we expect that the sum of shares for A and C would equal the share of A in the original two product set. Bradley-Terry-Luce and multinomial logit do not predict such an outcome (even when developed and applied at the individual level), but first-choice and multinomial probit do.

Estimation Method

The estimation methods that are used depend upon the data and the model being used. Ordinary conjoint tends to use either scale ratings or rank orders: OLS ANOVA/ regression models are most commonly used with these types of data, although ordinal models, such as ordered logit or MONANOVA, are sometimes used with rank orders. (It is generally felt that the less common ordinal models do not produce superior models (Green & Srinivasan, 1978).) Discrete choice modeling with choice data tends to use maximum likelihood estimation, although variations on least squares estimation methods (weighted or generalized least squares or minimum chi-squared) are also used. Adaptive conjoint analysis uses least squares regression starting from ad hoc prior coefficient estimates.

Generally speaking there is no strong advantage of one estimation method over another and, in fact, most methods that are applicable to a model provide equivalent results in statistical properties for common practical applications in market research. Statisticians generally prefer those estimation methods (such as maximum likelihood) for which they know certain nice theoretical sampling properties. However, I must admit that knowing sampling properties is not a very compelling reason for choosing an estimator, given the availability of resampling techniques for deriving sampling properties of virtually any estimator (Efron, 1982). In addition, in complex models, those “nice” sampling properties of, say, maximum likelihood estimates are generally asymptotic (i.e., for very large samples). This is not much of a relative advantage because, as the sample size gets so large that the sample converges on the population, most reasonable estimates, obtained with the same degrees of freedom, usually converge on the population parameter values and so have similar asymptotic properties to maximum likelihood.

Level of Aggregation

The model may specify its parameters either for each individual respondent or for the aggregated group. Those models which have only aggregate-level parameters are less flexible as analytic tools, since they assume that all respondents in the group have the same parameter values, thereby not allowing for individual differences.

On the other hand, individual-level models require a great deal of data from the individual respondent and so, to burden the respondent as little as possible, usually have few degrees of freedom. Consequently, the model for an individual respondent may not be very stable. (Aggregation of individual models to a large group, however, are quite stable because of the very large composite degrees of freedom.) Improvement in the stability of the individual respondent's model should reduce the variance in aggregate parameter estimates (such as estimates of choice shares), and so is a goal.

Recommended Approach

My personal assessment after many years of experience with these models is that, if there is a single most useful approach, it is ordinary conjoint using scale ratings (structured so the respondent is encouraged to compare the products being rated) with a choice model (preferably first-choice or multinomial probit) added on after estimation. This seems to provide the efficiency of data collection, accuracy of prediction, and flexibility of application that is most nearly optimal. However, it suffers in a few areas.

- Rating a series of product profiles on choice likelihood is an arduous task for respondents, although we have usually been able to obtain acceptably reliable and accurate data in this way. Nonetheless, improving the "user-friendliness" of this task would benefit respondent cooperation, if not model accuracy.
- Individual partworth estimates are quite unstable and, as a result, often show violations of suspected relationships between attribute levels. For example, while I don't necessarily expect that the partworths on price should always be monotonic decreasing (after all, some people feel that "you get what you pay for" and so prefer higher prices), I believe that it is unreasonable for partworths to decrease, increase, decrease, and then increase again as price increases. Individual partworths sometimes show this relationship because of the instability of the individual level model.
- Choice tasks are often viewed by users of the research as having greater face credibility than do scale ratings. I acknowledge this point as a reasonable one, though it has only modest validity: I have never seen choice tasks by themselves produce consistently superior predictions on hold-out data when the full-profile conjoint task was reasonably well constructed. Furthermore, a single choice provides inherently weaker information about a respondent's utilities than does a scale rating and so the cost of obtaining adequate choice data is often greater. Finally, any problem for which discrete choice modeling is applicable may be equally appropriately addressed by ordinary conjoint, whereas ordinary conjoint is also applicable to non-discrete choice (or even non-choice) modeling problems.
- The choice model is an addendum to the conjoint model and often is logically inconsistent with it. For example, a first choice model assumes no error in the utility estimate, an assumption that is almost certainly not true or consistent with the

model used to estimate partworths. Models usually benefit from being fully integrated and logically consistent.

Given the above discussion, I propose the Integrated Choice Likelihood (ICL) model, a model which uses scale ratings of full- or partial-profile products, scale ratings of attribute levels to augment the other data with data that is easier on respondents, and choice probability ratings to add face credibility with a relatively small number of additional tasks. The model directly incorporates choice prediction in a logically consistent fashion (even if the optional choice data are omitted) and uses a restricted version of the multinomial probit choice model. It makes more interesting and less burdensome the data collection task. While models are developed at the individual level, their robustness is improved by incorporating a points-of-view model (Tucker & Messick, 1963; also known as a latent class model, Lazarsfeld & Henry, 1968) in which an individual is taken as a composite or mixture of general respondent types (Finkbeiner, 1985; Hagerty, 1985). Finally, the model offers a great deal of flexibility and allows customization in data collection, making it well adapted to computerized interviewing.

I discuss the data, the model, the loss function, parameter estimation, special cases of the model, choice simulation, design issues, and attribute interactions, and provide an example in the ensuing sections. For the most part in these sections, without loss of generality, I discuss the ICL model as including only main effects (in the experimental design sense), and not interaction effects. The section on "Attribute Interactions" describes the modifications needed for an interaction effects version of the model.

DATA

Data from up to three different kinds of respondent tasks are used by the ICL model.

1. Scale ratings of the likelihood of choosing each of a set of full- or partial-profile products ($\mathbf{y}_i$). These ratings will be referred to hereafter as *profiled product ratings*. This task is the same as that for an ordinary full-profile conjoint analysis and should be structured so as to provide the respondent with opportunities to compare products when rating them.

2. Scale ratings of the likelihood of choosing a product with each level of each attribute ($\mathbf{r}_{ik}$). There are as many of these ratings as there are attribute levels.

3. Choice probability ratings in one or more scenarios ($\mathbf{x}_{is}$). A scenario is a set of products and within a scenario, the respondent rates the probability of choosing each product. The task is a constant sum rating where the probabilities must sum to 1.0 across the products in the scenario. As noted previously, simple choice is a special case of this task. Furthermore, one of the scenarios can be the current market using either pre-specified product profiles or using the respondent's perceived profiles, with $\mathbf{x}_{is}$ representing the actual purchase (i.e., choice) or distribution of purchases (i.e., choice probabilities) of existing products. This latter is a useful

device for producing a model which is consistent with current product purchase behavior.

Any of these three data tasks may be excluded altogether, however, either or both of the profiled product ratings and the choice probability ratings must always be present. The attribute level ratings by themselves do not carry enough information about the scale of each attribute to be sufficient for estimating choices. Profiled product ratings do carry that information, as do choice probability ratings if the profiled product ratings are present or if a parameter of the model is constrained.

MODEL

A respondent's profiled product rating is modeled as a recentering of the product *total utility*. Total utility is defined as the sum of those *respondent* partworths corresponding to the product's attribute levels (as specified in the relevant row of $\mathbf{D}_i$). The respondent partworths are a linear combination, $\mathbf{v}_i$, of the general respondent types' partworths, $\mathbf{U}$, for those attributes and levels included for the respondent: i.e., $\mathbf{G}_i\mathbf{U}\mathbf{v}_i$. Thus, the profiled product ratings are modeled by:

$$\mathbf{y}_i = c_i \mathbf{1}_{J_i} + \mathbf{D}_i \mathbf{G}_i \mathbf{U} \mathbf{v}_i + \mathbf{e}_{yi}$$

A respondent's direct ratings of an attribute's levels, $\mathbf{r}_{ik}$, are modeled as a simple linear rescaling of the corresponding respondent partworths for that attribute. The multiplicative constant, b_{ik}^2, must be non-negative so that $\mathbf{r}_{ik}$ will tend to be consistent in direction with the respondent partworths. Thus, the direct attribute level ratings are modeled by:

$$\mathbf{r}_{ik} = a_{ik} \mathbf{1}_{L_{ik}} + b_{ik}^2 \mathbf{G}_{ik} \mathbf{U} \mathbf{v}_i + \mathbf{e}_{rik}$$

The choice probability rating, $\mathbf{x}_{is}$, by the respondent in a given scenario is modeled by assuming that underlying the choice is $\mathbf{z}_{is}$, the product total utilities (defined as for $\mathbf{y}_i$ with the design matrix $\mathbf{H}_{is}$ taking the place of $\mathbf{D}_i$) plus error:

$$\mathbf{z}_{is} = \mathbf{H}_{is} \mathbf{G}_i \mathbf{U} \mathbf{v}_i + \mathbf{e}_{zis}$$

The choice probability rating of product m in scenario s reflects $p_{ism} = \Pr(z_{ism} > z_{isn} \; \forall \; n \neq m)$. Since any linear rescaling of $\mathbf{z}_{is}$ will not affect p_{ism} (Finkbeiner, 1988), then the only required difference between $\mathbf{y}_i$ and $\mathbf{z}_{is}$ (assuming identical design matrices) is the additive constant c_i in $\mathbf{y}_i$.

The function of the matrix $\mathbf{G}_i$ in the preceding equations is to indicate those levels and attributes which are dropped from individual respondent i's model because they do not contribute anything to respondent i's total utility even when present in a product profile. In effect, the partworths for those dropped levels and attributes are taken to be

zero. Their exclusion is justified on the grounds that, if correct, the remaining parameters are more accurately estimated. An example of when the option of excluding an attribute may be useful is when we know that the attribute is not relevant to a subpopulation of respondents. Dropping a level of an attribute may be useful when we know that the level won't be made available to certain subpopulations.

The dropping of levels or attributes using $\mathbf{G}_i$ just discussed does *not* include dropping them for administrative expediency during data collection. For example, adaptive conjoint analysis (Johnson, 1987b) drops some attributes from each of the profile ratings in order to simplify the respondent's task. In this case, a dropped attribute may well contribute to the respondent's total utility and is not taken to have partworths of zero. This kind of dropping of levels or attributes is controlled by appropriately located zeroes in the design matrices $\mathbf{D}_i$ and $\mathbf{H}_{is}$.

Assume that, for a given respondent, the errors of the three types of ratings are distributed in the multivariate normal distribution with a mean of zero and the appropriate covariance matrix as given below:

$$\boldsymbol{\Sigma}_{yi} = \sigma_i^2 \mathbf{Q}_{J_i}$$

$$\boldsymbol{\Sigma}_{rik} = \sigma_{ri}^2 \mathbf{I}_{L_{ik}}$$

$$\boldsymbol{\Sigma}_{zis} = \sigma_i^2 \mathbf{Q}_{M_{is}}$$

where, in the above, the $\mathbf{Q}$ matrices are defined to have ones in the diagonal and equal values (ρ_i) off-diagonal as follows:

$$\mathbf{Q}_X = (1-\rho_i)\mathbf{I}_X + \rho_i \mathbf{1}_X \mathbf{1}'_X \qquad (-1 \le \rho_i \le 1)$$

There are $1 + K_i + S_i$ error vectors in the above model: one for the profiled product ratings, K_i for the attributes in the attribute level ratings, and S_i for each choice scenario. Assume that these different error terms are uncorrelated with each other and that one of the error terms for one respondent is independent of the error terms for all other respondents.

For the attribute level ratings error vector, the assumption that the error terms are uncorrelated implies that on different occasions, the variation in one respondent's ratings of an attribute level are independent of all other attribute levels. The error variances for these ratings are constant: σ_{ri}^2. Note that, for $\boldsymbol{\Sigma}_{rik}$ to be positive semi-definite, as it must be, σ_{ri}^2 must be ≥ 0.

The error terms for $\mathbf{y}_i$ are assumed to have correlations that are equal across products, where products are defined by rows of the design matrix. This equicorrelation is the equivalent of a "halo" effect and its assumption is justified by the fact that the same

respondent is making all ratings across the products. The same equicorrelation assumption about the error terms for $\mathbf{z}_i$ also holds.

In addition to the above equicorrelation of errors assumption, we also recognize that products that are (nearly) identical should have (near-)identical utilities whenever they are rated together and so, should have error correlation near 1.0. Repeating the same profile product rating is almost never done in practice, except when the duplicates are separated in such a way as to render them nearly independent so that the equicorrelation assumption still applies. However, in the choice probability portion of the task, we assume that duplicate products in the same scenario will be noticed by the respondent. The model takes this into account by treating (near-)identical products as a single product in the choice ratings, splitting equally whatever choice probability would go to the single product. The concept of similarity of products affecting choice has been adopted elsewhere (Lakshmi-Ratan, 1984; Johnson, 1987a; Kamakura & Srivastava, 1984), though the present treatment is a fairly simple one.

Since $\mathbf{\Sigma}_{yi}$ and $\mathbf{\Sigma}_{zis}$ are defined as proportional to the relevant $\mathbf{Q}$ matrices using the same parameter σ_i^2 as the proportionality constant, the variances of the error terms are assumed constant. For these matrices to be positive semi-definite, σ_i^2 must be ≥ 0.

This model structure implies that:

$$\mathbf{y}_i \sim \mathcal{N}(\mu_{yi}, \Sigma_{yi})$$
$$\mathbf{r}_{ik} \sim \mathcal{N}(\mu_{rik}, \Sigma_{rik})$$
$$\mathbf{z}_{is} \sim \mathcal{N}(\mu_{zis}, \Sigma_{zis})$$

Consider an element of $\mathbf{x}_{is}$ (say, x_{ism}) to represent a proportion out of some hypothetical number N_{is}, where N_{is} is the number of hypothetical occasions on which respondent i chooses among the M_{is} products in i's scenario s. Thus, given N_{is} hypothetical choice occasions with this scenario, product m is chosen $N_{is}x_{ism}$ times by respondent i.

In fact, this may be very much like the way in which $\mathbf{x}_{is}$ is actually obtained from the respondent—by asking, for example, "Think about your next N_{is} choice occasions. Assume the products available are as described in this scenario. Please indicate how many times you would choose each of those products out of the next N_{is} choice occasions." Of course, a special case of this is simple choice, for example, "On your next choice occasion, if the products available are those described in this scenario, which product would you choose?"

If there actually were N_{is} choice occasions, then it would follow that

$$N_{is}\mathbf{x}_{is} \sim \mathcal{M}(\mathbf{p}_{is}, N_{is})$$

where $\mathbf{p}_{is}$ is evaluated by derivation from the assumed normal distribution of $\mathbf{z}_{is}$. It turns out that the value of N_{is} is immaterial because it is fixed and so does not affect the parameter estimates.

It remains to define $\mathbf{p}_{is}$. The above multinomial assumption and the multivariate normality assumption on $\mathbf{z}_{is}$ define a multinomial probit model in which the probability for a product is defined as the proportion of the multinormal distribution in which that product's z_{ism} is greater than the z_{ism} for all other products in the scenario. Mathematically, this definition is:

$$\Pr\left[z_m > \max_{m' \neq m}(z_{m'}) \right] = \int_{z_1 < z_m} \int_{z_2 < z_m} \cdots \int_{z_m=-\infty}^{z_m=+\infty} \cdots \int_{z_M < z_m} \Phi\left(\mathbf{z} | \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\right) dz_1 \ldots dz_M$$

where, for simplicity, the subscript "is" has been dropped from z_{ism}, M_{is}, $\mathbf{z}_{is}$, $\boldsymbol{\mu}_{zis}$, and $\boldsymbol{\Sigma}_{zis}$.

One could use the Clark (1961) approximation to estimate the above multivariate integral, but a more accurate approximation is available for the special case of equicorrelation. This method (called the equicorrelated probit approximation) is primarily due to Gumbel (1961) and is well described in Bock (1975, pp. 520 – 522). It yields accurate approximations to the true multinormal integral and so we use it to define $\mathbf{p}_{is}$:

$$p_{ism} = \frac{\exp(\theta_i \mu_{zism})}{\sum_{m'=1}^{M_{is}} \exp(\theta_i \mu_{zism'})}$$

where

$$\theta_i = \frac{\pi}{\sqrt{6\sigma_i^2(1-\rho_i)}}$$

This definition is modified when, as described earlier, two or more products in the scenario are (near-)identical. In that case, one of the identical products is chosen as a surrogate for the others and is the only one included in the calculation for $\mathbf{p}_{is}$. The choice probability obtained by the one surrogate product is split equally among the identical products.

Difficulties in using the equicorrelated probit approximation for the choice probabilities occur when $\sigma_i^2 = 0$, where evaluation of $\mathbf{p}_{is}$ requires dividing by σ_i^2. This condition implies that the respondent's profiled product and choice probability ratings are predicted perfectly. As $\sigma_i^2 \to 0$, the choice probabilities from the above model approach the choice probabilities from the first-choice model, in which the product with the largest μ_{zism} is

chosen with probability 1. Therefore, when a $\sigma_i^2 = 0$, $\mathbf{p}_{is}$ is appropriately estimated from the first-choice model, thereby avoiding division by zero.

LOSS FUNCTION

I use maximum likelihood as the method of estimating the parameters of the above model: σ_i^2, ρ_i, σ_{ri}^2, a_{ik} ($k = 1...K_i$), b_{ik}^2 ($k = 1...K_i$), c_i, $\mathbf{v}_i$, and $\mathbf{U}$. The subscript i indicates parameters unique to a respondent. Note that the parameters $\mathbf{p}_{is}$, $\boldsymbol{\mu}_{yi}$, $\boldsymbol{\mu}_{rik}$, $\boldsymbol{\mu}_{zis}$, $\boldsymbol{\Sigma}_{yi}$, $\boldsymbol{\Sigma}_{rik}$, and $\boldsymbol{\Sigma}_{zis}$ are functions of the previous parameters and so their maximum likelihood estimates are those functions of the other parameters' maximum likelihood estimates.

The loss function is -2 times the logarithm of the likelihood function of the parameters, given the data. Minimizing this function with respect to the parameters yields maximum likelihood parameter estimates. (Throughout this development, restrictions on the values of σ_i^2, ρ_i, and σ_{ri}^2 are noted where relevant. These restrictions will be dealt with in the **Parameter Estimation** section.)

In specifying the loss function to be used for estimation, begin with the probability density function, $\Im$, of the entire sample, including all three data types:

$$\Im = \prod_{i=1}^{N}\left[\Phi\left(\mathbf{y}_i\left|\boldsymbol{\mu}_{yi},\boldsymbol{\Sigma}_{yi}\right.\right)^{\delta_y} \times \prod_{k=1}^{K_i}\Phi\left(\mathbf{r}_{ik}\left|\boldsymbol{\mu}_{rik},\boldsymbol{\Sigma}_{rik}\right.\right)^{\delta_r} \times \prod_{s=1}^{S_i}\Psi\left(\mathbf{x}_{is}\left|\mathbf{p}_{is},N_{is}\right.\right)^{\delta_x}\right]^{w_i}$$

Substituting for the probability density functions Φ and Ψ, convert the probability density function $\Im$ to the likelihood function ℓ by reversing the roles of the variable data and fixed parameters to be fixed data and variable parameters. Then, take -2 times the logarithm of ℓ to yield:

$$\begin{aligned}-2\ln(\ell) = \sum_{i=1}^{N} w_i\Bigg[&\left(\delta_y J_i + \delta_r L_i\right)\ln(2\pi) + \delta_y\left(\ln\left|\boldsymbol{\Sigma}_{yi}\right| + \left(\mathbf{y}_i - \boldsymbol{\mu}_{yi}\right)'\boldsymbol{\Sigma}_{yi}^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_{yi}\right)\right)\\ &+ \delta_r\sum_{k=1}^{K_i}\left(\ln\left|\boldsymbol{\Sigma}_{rik}\right| + \left(\mathbf{r}_{ik} - \boldsymbol{\mu}_{rik}\right)'\boldsymbol{\Sigma}_{rik}^{-1}\left(\mathbf{r}_{ik} - \boldsymbol{\mu}_{rik}\right)\right)\\ &-2\delta_x\sum_{s=1}^{S_i}\left(\ln\left(N_{is}!\right) + \sum_{m=1}^{M_{is}}\left[x_{ism}\ln\left(p_{ism}\right) - \ln\left((N_{is}x_{ism})!\right)\right]\right)\Bigg]\end{aligned}$$

Drop any term in $-2\ln(\ell)$ which does not include parameters, since data are fixed during parameter estimation and since data terms cancel out in the one other use of $-2\ln(\ell)$: the statistical test proposed near the end of the **Parameter Estimation** section.

From the definition of $\boldsymbol{\Sigma}_{yi}$, it can be shown that:

$$\ln\left|\boldsymbol{\Sigma}_{yi}\right| = J_i \ln(\sigma_i^2) + (J_i - 1)\ln(1-\rho_i) + \ln(1+(J_i-1)\rho_i)$$

$$\boldsymbol{\Sigma}_{yi}^{-1} = \frac{1}{\sigma_i^2}\mathbf{Q}_{J_i}^{-1} \qquad \text{where } \mathbf{Q}_{J_i}^{-1} = \frac{1}{1-\rho_i}\left[\mathbf{I}_{J_i} - \frac{\rho_i}{1+\rho_i(J_i-1)}\mathbf{1}_{J_i}\mathbf{1}_{J_i}'\right]$$

Note that this logarithm and this inverse do not exist if $\rho_i = 1$, $\rho_i \le -(J_i\text{-}1)\text{-}1$, or $\sigma_i^2 \le 0$.

The logarithm of the determinant and the inverse of the covariance matrix for $\mathbf{r}_{ik}$ is:

$$\ln\left|\boldsymbol{\Sigma}_{rik}\right| = L_{ik}\ln(\sigma_{ri}^2)$$

$$\boldsymbol{\Sigma}_{rik}^{-1} = \frac{1}{\sigma_{ri}^2}\mathbf{I}_{L_{ik}}$$

Note that this logarithm and this inverse do not exist if $\sigma_{ri}^2 \le 0$.

Substituting the preceding results into the expression for $-2\ln(\ell)$, dropping terms which do not include parameters, and noting that $\mathbf{e}_{yi} = \mathbf{y}_i - \boldsymbol{\mu}_{yi}$ and that $\mathbf{e}_{rik} = \mathbf{r}_{ik} - \boldsymbol{\mu}_{rik}$, the loss function simplifies to:

$$f = \sum_{i=1}^{N} w_i \left[\delta_y \left(J_i \ln(\sigma_i^2) + (J_i - 1)\ln(1-\rho_i) + \ln(1+(J_i-1)\rho_i) + \frac{\mathbf{e}_{yi}'\mathbf{Q}_{J_i}^{-1}\mathbf{e}_{yi}}{\sigma_i^2} \right) + \delta_r \left(L_i \ln(\sigma_{ri}^2) + \frac{1}{\sigma_{ri}^2}\sum_{k=1}^{K_i}\mathbf{e}_{rik}'\mathbf{e}_{rik} \right) - 2\delta_x \mathbf{x}_i' \ln(\mathbf{p}_i) \right]$$

This loss function is minimized in parameter estimation.

PARAMETER ESTIMATION

The procedure used to estimate the parameters σ_i^2, ρ_i, σ_{ri}^2, a_{ik} & $b_{ik}{}^2$ ($k = 1...K_i$), c_i, $\mathbf{v}_i$, and $\mathbf{U}$ is an iterative one which minimizes f, except in the case when only the profiled product ratings ($\mathbf{y}_i$) are present (i.e., $\delta_y = 1$) where a closed-form solution is possible. Broadly speaking, the procedure cycles between estimating the individual-level parameters (σ_i^2, ρ_i, σ_{ri}^2, a_{ik} & $b_{ik}{}^2$ ($k = 1...K_i$), c_i, and $\mathbf{v}_i$) and estimating the aggregate-level parameter, $\mathbf{U}$. In fact, estimation of the individual-level parameters is further broken into two parts: those parameters which have a closed form for their estimates, and those which do not.

When all three data types are present (i.e., $\delta_r = \delta_x = \delta_y = 1$), closed form estimates are possible for σ_{ri}^2, a_{ik} & $b_{ik}{}^2$ ($k = 1...K_i$), and c_i, given the other parameters, but σ_i^2 and $\mathbf{v}_i$

require numerical optimization. As we shall see later, with maximum likelihood estimates of the other parameters, the estimate of ρ_i which minimizes f cannot be determined from the data; in this case, whatever value ρ_i is given appears to be compensated for by the maximum likelihood estimates of the other parameters so that the terms of most interest to the researcher—choice probabilities and choice share simulation, as well as the part-worths $\mathbf{Uv}_i$—are invariant. At the aggregate-level, with individual-level parameters held constant, numerical optimization is required again to estimate **U**. Simplifications are possible when one or two of the three data types are dropped, as described in the next section.

The steps in the iterative estimation process begin with closed-form estimation of σ_{ri}^2, a_{ik} & b_{ik}^2 ($k = 1...K_i$), and c_i, given the other parameters. The following derivatives are necessary for this process:

$$\frac{\partial f}{\partial c_i} = \left(\frac{-2\delta_y w_i}{\sigma_i^2 (1 + (J_i - 1)\rho_i)} \right) \mathbf{1}'_{J_i} \mathbf{e}_{yi}$$

$$\frac{\partial f}{\partial a_{ik}} = \left(\frac{-2\delta_r w_i}{\sigma_{ri}^2} \right) \mathbf{1}'_{L_{ik}} \mathbf{e}_{rik}$$

$$\frac{\partial f}{\partial b_{ik}^2} = \left(\frac{-2\delta_r w_i}{\sigma_{ri}^2} \right) \mathbf{v}'_i \mathbf{U}' \mathbf{G}'_{ik} \mathbf{e}_{rik} \quad \forall\ k = 1 \ldots K_i$$

$$\frac{\partial f}{\partial \sigma_{ri}^2} = \frac{-\delta_r w_i}{\sigma_{ri}^2} \left[L_i - \frac{\sum_{k=1}^{K_i} \mathbf{e}'_{rik} \mathbf{e}_{rik}}{\sigma_{ri}^2} \right]$$

Setting these derivatives to 0, solving, and enforcing previously noted restrictions, yields the following parameter estimates:

when $\delta_y = 1$,

(1a) $$\hat{c}_i = \frac{\mathbf{1}'_{J_i} \left(\mathbf{y}_i - \mathbf{D}_i \mathbf{G}_i \mathbf{U} \mathbf{v}_i \right)}{J_i}$$

when $\delta_r = 1$,

$$
\begin{aligned}
\hat{b}_{ik}^2 &= \max\left(0, \frac{\mathbf{v}_i' \mathbf{U}' \mathbf{G}_{ik}' \left(\mathbf{I}_{L_{ik}} - \frac{1}{L_{ik}} \mathbf{1}_{L_{ik}} \mathbf{1}_{L_{ik}}'\right) \mathbf{r}_{ik}}{\mathbf{v}_i' \mathbf{U}' \mathbf{G}_{ik}' \left(\mathbf{I}_{L_{ik}} - \frac{1}{L_{ik}} \mathbf{1}_{L_{ik}} \mathbf{1}_{L_{ik}}'\right) \mathbf{G}_{ik} \mathbf{U} \mathbf{v}_i}\right) \\
\hat{a}_{ik} &= \frac{1}{L_{ik}} \mathbf{1}_{L_{ik}}' \left(\mathbf{r}_{ik} - b_{ik}^2 \mathbf{G}_{ik} \mathbf{U} \mathbf{v}_i\right) \\
\hat{\sigma}_{ri}^2 &= \max\left(\varepsilon, \frac{\mathbf{e}_{ri}' \mathbf{e}_{ri}}{L_i}\right)
\end{aligned}
\tag{1b}
$$

Note that $\hat{\sigma}_{ri}^2$ is not allowed to fall below ε and $\hat{b}_{ik}^2$ is not allowed to fall below 0, in order to prevent improper conditions.

The other individual-level parameters (σ_i^2, and $\mathbf{v}_i$) must be obtained by numerical optimization conditional on the above equations for $\hat{\sigma}_{ri}^2, \hat{a}_{ik}, \hat{b}_{ik}^2$ ($k = 1 \ldots K_i$) and $\hat{c}_i$. In this optimization for individual-level parameters, solve for σ_i, rather than σ_i^2, in order to force the estimate of the squared parameter to be non-negative. If, at any point in the function evaluation, $\sigma_i < \varepsilon$, reset it to ε.

The optimization algorithm I recommend is the modification of Powell's conjugate directions method described in Press, *et al* (1992, pp. 412-420). This method is relatively simple to implement because it does not require derivatives. However, in case a derivative-based optimization algorithm may be desired, the derivatives with respect to σ_i, $\mathbf{v}_i$, and **U** are given below. In the following derivatives, we make use of the fact that $\mathbf{Q}_{J_i}^{-1} \mathbf{e}_{yi} = \frac{\mathbf{e}_{yi}}{1 - \rho_i}$ when $\mathbf{e}_{yi}$ is evaluated at the above estimate of c_i.

$$
\frac{\partial f}{\partial \sigma_i} = 2 w_i \left[\frac{\delta_y}{\sigma_i}\left(J_i - \frac{\tilde{\mathbf{e}}_{yi}' \tilde{\mathbf{e}}_{yi}}{\sigma_i^2 (1 - \rho_i)}\right) + \delta_x \theta_i (\mathbf{x}_i - \mathbf{p}_i)' \boldsymbol{\mu}_{zi}\right]
$$

$$
\frac{\partial f}{\partial \mathbf{v}_i} = -2 w_i \mathbf{U}' \mathbf{G}_i' \left[\frac{\delta_y}{\sigma_i^2 (1 - \rho_i)} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \frac{\delta_r}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} + \delta_x \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i\right) * \mathbf{x}_i\right]\right]
$$

$$
\frac{\partial f}{\partial \mathbf{U}} = -2 \sum_{i=1}^{N} w_i \mathbf{G}_i' \left[\frac{\delta_y}{\sigma_i^2 (1 - \rho_i)} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \frac{\delta_r}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} + \delta_x \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i\right) * \mathbf{x}_i\right]\right] \mathbf{v}_i'
$$

where the notations of "*" and "~" in the above (and the following equations) are defined as:

- "*" refers to element-by-element multiplication of vectors
- "~" over $\mathbf{e}_{yi}$ and $\mathbf{e}_{rik}$ means those terms are evaluated using whatever maximum likelihood parameter estimates are available for them—in this case $\hat{\sigma}_{ri}^2, \hat{a}_{ik}, \hat{b}_{ik}^2$ $(k = 1 \ldots K_i)$ and $\hat{c}_i$—but that some parameters are involved—in this case σ_i, **U**, and $\mathbf{v}_i$—for which maximum likelihood estimates are not yet available. If all parameters were evaluated at their maximum likelihood estimates, the "~" would have been replaced by a "^".

It turns out that the parameter ρ_i is indeterminate. This can be seen from its derivative:

$$\frac{\partial f}{\partial \rho_i} = w_i \left[\delta_y \left(\frac{1-J_i}{1-\rho_i} + \frac{J_i - 1}{1+(J_i-1)\rho_i} + \frac{\tilde{\mathbf{e}}'_{yi}\tilde{\mathbf{e}}_{yi}}{\sigma_i^2(1-\rho_i)^2} \right) - \delta_x \frac{\theta_i}{1-\rho_i}(\mathbf{x}_i - \mathbf{p}_i)'\boldsymbol{\mu}_{zi} \right]$$

If σ_i is estimated at its maximum likelihood value for which its derivative is zero, then it can be shown that the above derivative for ρ_i reduces to:

$$\frac{\partial f}{\partial \rho_i} = w_i \left[\frac{J_i}{(1-\rho_i)(1+(J_i-1)\rho_i)} \right]$$

This derivative is never ≤ zero for permissible values of ρ_i and so, as long as ρ_i is not fixed, f cannot be minimized. However, the above derivative is not a function of any data from respondents, and so the value ρ_i might take on is not determined by the choice likelihood data we are modeling. Furthermore, in the only other derivatives in which ρ_i appears (for σ_i, $\mathbf{v}_i$, and **U**), it is multiplied by σ_i^2. Thus, while σ_i must be estimated so that its derivative is zero for whatever value ρ_i takes on, this estimation of σ_i assures that $\sigma_i^2(1 - \rho_i)$ is invariant under arbitrary changes in ρ_i. This invariance, in turn, assures that the estimates of $\mathbf{v}_i$ and **U** are also invariant. Furthermore, since the choice probabilities also contain $\sigma_i^2(1 - \rho_i)$, then the estimates of $\mathbf{p}_i$ will also be invariant. Consequently, the parameters of primary interest to the researcher—the choice probabilities, $\mathbf{p}_i$, and the partworths, $\mathbf{U}\mathbf{v}_i$—are not affected by the value of ρ_i and we may choose to set it to any value. It is particularly convenient to set ρ_i to 0, since it thereby disappears from all of the preceding equations for f, and from the derivatives with respect to σ_i, $\mathbf{v}_i$, and **U**. In particular, the f we now seek to minimize, conditional on the use of the maximum likelihood estimate of c_i and of $\rho_i = 0$, may be written as:

$$f = \sum_{i=1}^{N} w_i \left[\delta_y \left(J_i \ln(\sigma_i^2) + \frac{\tilde{\mathbf{e}}'_{yi} \tilde{\mathbf{e}}_{yi}}{\sigma_i^2} \right) + \delta_r \left(L_i \ln(\hat{\sigma}_{ri}^2) + \frac{1}{\hat{\sigma}_{ri}^2} \sum_{k=1}^{K_i} \tilde{\mathbf{e}}'_{rik} \tilde{\mathbf{e}}_{rik} \right) - 2\delta_x \mathbf{x}'_i \ln(\mathbf{p}_i) \right]$$

where θ_i in $\mathbf{p}_i$ is now defined as $\frac{\pi}{\sigma_i \sqrt{6}}$.

To find estimates of σ_i, $\mathbf{v}_i$, and $\mathbf{U}$ which minimize f, cycle over the following two steps.

1. Estimation of σ_i and $\mathbf{v}_i$

   Use numerical optimization of f to find σ_i and $\mathbf{v}_i$, conditional on all other parameters. In the optimization process, before any calculation of f (or, if used, of the above derivatives), compute estimates of a_{ik}, b_{ik}^2 ($k = 1...K_i$), c_i, and σ_{ri}^2 from equations (1) above so that all necessary simplifications apply. After optimization, re-calculate estimates of a_{ik}, b_{ik}^2 ($k = 1...K_i$), c_i, and σ_{ri}^2 so they are consistent with the optimized estimates of σ_i and $\mathbf{v}_i$.

2. Estimation of $\mathbf{U}$

   Use numerical optimization of f to find $\mathbf{U}$, conditional on all other parameters. After optimization, rescale $\mathbf{U}$ as described subsequently for $\mathbf{V}_T$ in equations (2a-b) to meet $\mathbf{U}$'s restrictions.

Cycle between steps 1. and 2. until convergence of the parameters is reached.

Note that the columns of each $\mathbf{U}_k$ may be recentered without affecting our estimates of $\mathbf{y}_i$ (since c_i and $\mathbf{v}_i$ can compensate exactly) or of $\mathbf{r}_{ik}$ (since a_{ik} will compensate) or of $\mathbf{p}_{is}$ (since additive constants in the exp function cancel out of the numerator and denominator of $\mathbf{p}_{is}$). Furthermore, any T × T nonsingular transformation matrix applied to $\mathbf{U}$ can be compensated for by multiplying $\mathbf{v}_i$ by the inverse of the transformation matrix. These invariance properties of $\mathbf{U}$ imply that there are T(L–K–T) uniquely identified parameters in $\mathbf{U}$, which implies that L-K must be ≥ T in order to uniquely determine $\mathbf{U}$. As a consequence, when T = L-K, there are no unconstrained elements to be estimated in $\mathbf{U}$. Setting T larger than L-K yields an indeterminate solution for $\mathbf{U}$.

In order to uniquely determine $\mathbf{U}$, it is necessary to restrict T(K+T) of its parameter values. It is most convenient for estimation if we directly fix certain elements to appropriate values. I suggest setting to zero the K_i rows of $\mathbf{U}$ which correspond to the first level of every attribute, and then to select T rows of $\mathbf{U}$ and set them equal to an identity matrix. The T rows to select may be any rows except the K_i first levels of every attribute. However, after some iteration of the estimation method has elapsed, I recommend that the T rows being constrained to $\mathbf{I}_T$ be re-selected based on the size of the corresponding pivot elements in a Cholesky decomposition of $\mathbf{U}'\mathbf{U}$. This assures that rows of $\mathbf{U}$ that have the largest elements in them are used for the constraints of 1.

During the iterative estimation procedure, only the unconstrained elements of **U** are estimated; the constrained elements of **U** are left alone.

Numerical optimization requires starting points for the parameter estimates on the first cycle. When $\delta_y = 1$, use as the starting points for σ_i and $\mathbf{v}_i$ their closed-form maximum likelihood estimates calculated when the profiled product ratings are the only one of the three types of rating obtained (see the **Conjoint (δ_y = 1, δ_r = 0, δ_x = 0)** sub-section below). If $\delta_y = 0$, σ_i is restricted to 1, as noted later: set $\mathbf{v}_i$ to **0** to start the first cycle. After the first cycle, the starting points are the values of $\mathbf{v}_i$ and σ_i from the prior cycle.

For the starting value required for **U,** use the method described in the following paragraphs.

If the attribute level ratings are present (i.e., $\delta_r = 1$), use as the starting value for **U** the eigenvectors of the matrix of covariances among the attribute level ratings, with the eigenvectors rescaled to satisfy the restrictions on **U**. Define the covariance matrix among the attribute level ratings as follows.

$$\mathbf{C}_{rr} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i'$$

where $\tilde{\mathbf{r}}_i$ is a column vector composed of the terms $\mathbf{r}_{i1} - \frac{1}{L_1}\mathbf{1}'_{L_1}\mathbf{r}_{i1}$, … , $\mathbf{r}_{iK} - \frac{1}{L_K}\mathbf{1}'_{L_K}\mathbf{r}_{iK}$, concatenated vertically, with mean replacement for missing elements of $\mathbf{r}_{ik}$, corresponding to zero rows of $\mathbf{G}_i$. Obtain $\mathbf{V}_T$ as the T eigenvectors of $\mathbf{C}_{rr}$ corresponding to its T largest eigenvalues.

If the attribute level ratings are not present (i.e., $\delta_r = 0$), select the rows to constrain as described earlier, and set to zero all other elements of **U**.

Alternatively, create a starting point based on the T largest principal components of an estimate of **U** obtained for a larger value of T. For example, I find it convenient to begin estimation with T = L-K, so that no parameters in **U** need be estimated. Then, I set T = L-K-1 and create a new starting point as just described.

The rows of $\mathbf{V}_T$ (however obtained) corresponding to one attribute are then re-centered so that the first row is **0.** This re-centering is accomplished by subtracting the first row from the remaining as is represented here:

(2a) $$\tilde{\mathbf{U}}_k = \left(\mathbf{I}_{L_k} - \mathbf{1}_{\mathbf{L_k}}\begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}\right)\mathbf{V}_{Tk}$$

Select the T rows of this recentered **U** that will be constrained to be an identity matrix. Call this collection of T rows **U***. Then carry out the following calculation to get the final starting value for **U**:

(2b) $$\mathbf{U}_k = \tilde{\mathbf{U}}_k \mathbf{U}^{*-1}$$

Note that the required constraints on **U** are imposed by (2a-b).

The model must be calculated for each T = 0,...,L-K and a value of T must be selected, although, as discussed later, it is probably not necessary to evaluate the model for T much greater than 10, regardless of the value of L-K. To aid in selecting T, consider the value for f at each T (call it $f^{(T)}$) and examine the size of the decrease in $f^{(T)}$ from one value of T to the next. A statistical significance test on this decrease can be carried out: because the parameter estimates are maximum likelihood estimates, it must follow that $f^{(T)} - f^{(T+1)}$ is asymptotically distributed in the χ^2 distribution with (N+L-K-2T-1) degrees of freedom.

It is also possible to select T using cross-validation. There are many possible ways to carry this out, but a simple approach is to choose that value of T which maximizes the accuracy of prediction of a hold-out choice probability task.

I should note that sampling variances of the parameter estimates cannot easily be computed (even asymptotic variances) since this requires the inversion of the negative of the matrix of second derivatives of f with respect to the parameters. However, sampling variances are of very limited usefulness in this model. First of all, most of the parameters are individual respondent parameters where significance testing is impractical due to the sheer number or parameters. Secondly, sampling variances on the aggregate parameter matrix, **U**, are of far less interest than hypothesis testing on whether **U** is significantly different from a zero matrix, which can be done readily using a slight extension of the chi-square statistic described above: $f^{(0)} - f^{(T)}$ is asymptotically distributed in the χ^2 distribution with degrees of freedom given by $T(N + L - K - T) + \delta_r \sum_i K_i$.

When a final model is selected, estimate and save for choice share simulation the information necessary for choice probability estimation, namely, the estimates σ_i^2, $\mathbf{v}_i$, and **U**, as well as the indicator matrix $\mathbf{G}_i$ (or enough data to reconstruct $\mathbf{G}_i$).

SPECIAL CASES OF THE FULL ICL MODEL

Some important simplifications of the estimation procedure can occur in special cases of the general model when one or two of the data types are excluded.

Conjoint ($\delta_y = 1$, $\delta_r = 0$, $\delta_x = 0$):

If profiled product ratings are all that are collected, the ICL model simplifies to a conjoint analysis model as follows. (The closed form estimates of σ_i^2 and $\mathbf{v}_i$ are used in the full ICL model as starting points for numerical optimization of those parameters.)

Parameters Estimated: σ_i, c_i, $\mathbf{v}_i$, **U**

Loss Function: $f = \sum_{i=1}^{N} w_i \left[J_i \ln(\sigma_i^2) + \frac{\tilde{\mathbf{e}}'_{yi} \tilde{\mathbf{e}}_{yi}}{\sigma_i^2} \right]$

Closed-Form Estimates:

$\hat{c}_i$ as in equation (1a)

$$\hat{\mathbf{v}}_i = \left(\mathbf{U}'\mathbf{G}_i'\mathbf{D}_i'\left(\mathbf{I}_{J_i} - \tfrac{1}{J_i}\mathbf{1}_{J_i}\mathbf{1}_{J_i}'\right)\mathbf{D}_i\mathbf{G}_i\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{G}_i'\mathbf{D}_i'\left(\mathbf{I}_{J_i} - \tfrac{1}{J_i}\mathbf{1}_{J_i}\mathbf{1}_{J_i}'\right)\mathbf{y}_i$$

$$\hat{\sigma}_i^2 = \frac{\tilde{\mathbf{e}}_{yi}'\tilde{\mathbf{e}}_{yi}}{J_i}$$

Partial Derivatives:

$$\frac{\partial f}{\partial \mathbf{U}} = -2\sum_{i=1}^{N}\left(\frac{w_i}{\hat{\sigma}_i^2}\right)\mathbf{G}_i'\mathbf{D}_i'\tilde{\mathbf{e}}_{yi}\hat{\mathbf{v}}_i'$$

Discrete Choice ($\delta_y = 0$, $\delta_r = 0$, $\delta_x = 1$):

If the only data collected are discrete choices among products, then the ICL model simplifies to a discrete choice model. Actually, the ICL model allows for "non-discrete" choice as well: e.g., the model allows for the use of choice probability ratings in the form of constant sum ratings. Therefore, referring to the present simplification as a discrete choice version of the ICL model is something of a misnomer. However, the term "discrete choice" is an often used term, and so it is convenient to use it as a short-hand reference to the present simplification of the general ICL model.

In this case, there is an indeterminacy in the loss function and hence in the parameter estimates. Referring back to the multivariate normal probability density function used in the multinomial probit integral defining the choice probabilities, $\mathbf{p}_{is}$, it can be shown that a change in σ_i can be exactly compensated for by an opposite change in $\mathbf{v}_i$ to leave the integral invariant. That is, if we multiply σ_i by some value, we can divide $\mathbf{v}_i$ by the same value, thereby leaving the choice probability estimates exactly the same.

To remove the above indeterminacy, it is sufficient to restrict the estimate of σ_i, as indicated below.

Parameters Estimated: $\mathbf{v}_i$, $\mathbf{U}$

Parameters Restricted: $\sigma_i = 1$

Loss Function: $f = -2\sum_{i=1}^{N} w_i \mathbf{x}_i' \ln(\mathbf{p}_i)$

Partial Derivatives:

$$\frac{\partial f}{\partial \mathbf{v}_i} = -2w_i\theta_i\mathbf{U}'\mathbf{G}_i'\mathbf{H}_i'\left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i\right)*\mathbf{x}_i\right]$$

$$\frac{\partial f}{\partial \mathbf{U}} = -2\sum_{i=1}^{N} w_i\theta_i\mathbf{G}_i'\mathbf{H}_i'\left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i\right)*\mathbf{x}_i\right]\mathbf{v}_i'$$

Hybrid Conjoint ($\delta_y = 1$, $\delta_r = 1$, $\delta_x = 0$):

Green *et al* (1981) created a version of conjoint analysis called "hybrid conjoint analysis" in which direct attribute ratings are combined with full-profile product ratings. The ICL version of this model is the special case when choice ratings are omitted, but profiled product ratings and attribute level ratings are included.

Parameters Estimated: σ_i, σ_{ri}^2, a_{ik} & b_{ik}^2 (k = 1...Ki), c_i, $\mathbf{v}_i$, $\mathbf{U}$

Loss Function: $$f = \sum_{i=1}^{N} w_i \left[J_i \ln(\sigma_i^2) + \frac{\tilde{\mathbf{e}}_{yi}' \tilde{\mathbf{e}}_{yi}}{\sigma_i^2} + L_i \ln(\hat{\sigma}_{ri}^2) + \frac{\tilde{\mathbf{e}}_{ri}' \tilde{\mathbf{e}}_{ri}}{\hat{\sigma}_{ri}^2} \right]$$

Closed-Form Estimates: $\hat{\sigma}_{ri}^2$, $\hat{a}_{ik}$, $\hat{b}_{ik}^2$, $\hat{c}_i$ as in equation (1)

$$\hat{\sigma}_i^2 = \frac{\tilde{\mathbf{e}}_{yi}' \tilde{\mathbf{e}}_{yi}}{J_i}$$

Partial Derivatives:

$$\frac{\partial f}{\partial \mathbf{v}_i} = -2 w_i \mathbf{U}' \mathbf{G}_i' \left[\frac{1}{\hat{\sigma}_i^2} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \frac{1}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} \right]$$

$$\frac{\partial f}{\partial \mathbf{U}} = -2 \sum_{i=1}^{N} w_i \mathbf{G}_i' \left[\frac{1}{\hat{\sigma}_i^2} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \frac{1}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} \right] \mathbf{v}_i'$$

Hybrid Discrete Choice ($\delta_y = 0$, $\delta_r = 1$, $\delta_x = 1$):

Combining direct attribute ratings with choice ratings, analogous to hybrid conjoint analysis, has, to my knowledge, never been suggested in publication. However, it is certainly conceptually possible to pose this combination as another special case of the general ICL model. Once again, to remove indeterminacy, it is necessary and sufficient to restrict the estimate of σ_i.

Parameters Estimated: σ_{ri}^2, a_{ik} & b_{ik}^2 (k = 1...K_i), $\mathbf{v}_i$, $\mathbf{U}$

Parameters Restricted: $\sigma_i = 1$

Loss Function: $$f = \sum_{i=1}^{N} w_i \left[L_i \ln(\hat{\sigma}_{ri}^2) + \frac{\tilde{\mathbf{e}}_{ri}' \tilde{\mathbf{e}}_{ri}}{\hat{\sigma}_{ri}^2} - 2\mathbf{x}_i' \ln(\mathbf{p}_i) \right]$$

Closed-Form Estimates: $\hat{\sigma}_{ri}^2$, $\hat{a}_{ik}$, $\hat{b}_{ik}^2$ as in equation (1)

Partial Derivatives:

$$\frac{\partial f}{\partial \mathbf{v}_i} = -2 w_i \mathbf{U}' \mathbf{G}_i' \left[\frac{1}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} + \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i \right) * \mathbf{x}_i \right] \right]$$

$$\frac{\partial f}{\partial \mathbf{U}} = -2 \sum_{i=1}^{N} w_i \mathbf{G}_i' \left[\frac{1}{\hat{\sigma}_{ri}^2} \begin{bmatrix} \hat{b}_{i1}^2 \tilde{\mathbf{e}}_{ri1} \\ \vdots \\ \hat{b}_{iK_i}^2 \tilde{\mathbf{e}}_{riK_i} \end{bmatrix} + \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i \right) * \mathbf{x}_i \right] \right] \mathbf{v}_i'$$

Conjoint & Choice ($\delta_y = 1$, $\delta_r = 0$, $\delta_x = 1$):

This is another data combination that the ICL model subsumes which, to my knowledge, has never been published.

Parameters Estimated: σ_i, c_i, $\mathbf{v}_i$, $\mathbf{U}$

Loss Function: $f = \sum_{i=1}^{N} w_i \left[J_i \ln(\sigma_i^2) + \frac{\tilde{\mathbf{e}}_{yi}' \tilde{\mathbf{e}}_{yi}}{\sigma_i^2} - 2\mathbf{x}_i' \ln(\mathbf{p}_i) \right]$

Closed-Form Estimates: $\hat{c}_i$ as in equation (1)

Partial Derivatives:

$$\frac{\partial f}{\partial \sigma_i} = 2 w_i \left[\frac{1}{\sigma_i} \left(J_i - \frac{\tilde{\mathbf{e}}_{yi}' \tilde{\mathbf{e}}_{yi}}{\sigma_i^2} \right) + \theta_i (\mathbf{x}_i - \mathbf{p}_i)' \boldsymbol{\mu}_{zi} \right]$$

$$\frac{\partial f}{\partial \mathbf{v}_i} = -2 w_i \mathbf{U}' \mathbf{G}_i' \left[\frac{1}{\sigma_i^2} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i \right) * \mathbf{x}_i \right] \right]$$

$$\frac{\partial f}{\partial \mathbf{U}} = -2 \sum_{i=1}^{N} w_i \mathbf{G}_i' \left[\frac{1}{\hat{\sigma}_i^2} \mathbf{D}_i' \tilde{\mathbf{e}}_{yi} + \theta_i \mathbf{H}_i' \left[\left(\mathbf{1}_{M_i} - \mathbf{p}_i \right) * \mathbf{x}_i \right] \right] \mathbf{v}_i'$$

CHOICE SIMULATION

Given that product choice probabilities are an integral part of the ICL model, simulation of choices for new products is a relatively straightforward application of the model. Not all parameter estimates are needed, only σ_i, $\mathbf{v}_i$, and $\mathbf{U}$. In addition, information is needed about which attributes or levels are not applicable to each respondent (i.e., the matrix $\mathbf{G}_i$). In the terminology of conjoint analysis, the values corresponding to respondent partworths are $\mathbf{G}_i\mathbf{U}\mathbf{v}_i$.

For any scenario, s* (need not be one of the S_i scenarios used on respondent i), the researcher must specify the matrix $\mathbf{H}_{is*}$, which could be the same matrix for all respodents, the same for a group of respondents, or even different for every respondent. Use $\mathbf{H}_{is*}$ to estimate $\boldsymbol{\mu}_{zis*}$ and $\boldsymbol{\Sigma}_{zis*}$:

$$\hat{\boldsymbol{\mu}}_{zis^*} = \mathbf{H}_{is^*}\mathbf{G}_i\hat{\mathbf{U}}\hat{\mathbf{v}}_i$$
$$\hat{\boldsymbol{\Sigma}}_{zis^*} = \hat{\sigma}_i^2\mathbf{Q}_{M_{is^*}}$$

These parameter estimates are then used in the equicorrelated probit approximation to estimate choice probabilities $\mathbf{p}_{is^*}$. The necessary equation is reproduced here.

$$\hat{p}_{ism} = \frac{\exp(\hat{\theta}_i\hat{\mu}_{zis^*m})}{\sum_{m'=1}^{M_{is}}\exp(\hat{\theta}_i\hat{\mu}_{zis^*m'})}$$

As described earlier, this equation is modified when products are (near-)identical by splitting the probability equally across the (near-)identical products. Furthermore, for $\hat{\sigma}_i^2 \le \varepsilon$, the first choice model is used to estimate probabilities.

To estimate choice *shares* (the proportion of a population of choice decision-makers that will choose each product) from a sample of respondents, consider that a respondent's choice probability is an estimate of the proportion of times that respondent (or one like him/her) will choose a product (see the **Model** section). In the population, a product's choice share is the expected value of the product's choice probabilities over every individual. Thus, the best sample estimate of product m's choice share in scenario s* is the weighted sample mean of the individual choice probabilities for m:

$$\hat{P}_{s^*m} = \frac{\sum_{i=1}^{N} w_i\hat{p}_{is^*m}}{\sum_{i=1}^{N} w_i}$$

If no products are identical and $\hat{\sigma}_i^2$ is at least modestly large, then the above choice model is an IIA model at the individual respondent level. However, the model is non-IIA if some products are identical or $\hat{\sigma}_i^2 \to 0$. Furthermore, the ICL model's aggregate choice shares are not subject to IIA. For example, as $\hat{\sigma}_i^2 \to 0$ for all respondents, the ICL choice share estimates approach those of the first-choice model, a non-IIA model. In fact, only if $\hat{\sigma}_i^2 \to \infty$ for all i or $\frac{1}{N}\sum_{i=1}^{N}(p_{ism} - \bar{p}_{sm})^2 \to 0$ for every m (so that every respondent has the same choice probabilities) does the ICL model approach an IIA model. The ICL model departs from an IIA model to the extent that measurement error is small or parameters are heterogeneous across individuals.

DESIGN ISSUES

A respondent provides $\delta_y J_i + \delta_r L_i + \delta_x (M_i - S_i)$ unique ratings, from which we must estimate $\delta_y + \delta_r(2K_i + 1) + T + 1$ individual-level parameters. In addition, those respondent ratings must share in the estimation of $T(L - K - T)$ unique aggregate-level parameters. In total, the degrees of freedom (df) across all respondents is the difference between the total number of unique ratings across all respondents and the total number of unique parameters being estimated across all respondents:

$$df = \delta_y\left(\sum_{i=1}^{N} J_i - N\right) + \delta_r\left(\sum_{i=1}^{N}(L_i - 2K_i) - N\right) + \delta_x \sum_{i=1}^{N}(M_i - S_i) - N(T+1) - T(L - K - T)$$

The average degrees of freedom per respondent is df/N. To follow usual practice in ordinary conjoint analysis studies, the average degrees of freedom per respondent should be at least 5 or 6, with greater degrees of freedom producing a more reliable model with less sampling error in parameter estimates.

Profiled product and choice probability ratings are difficult and burdensome for respondents if they take the task seriously. This model can afford a savings in the number of such ratings needed per respondent.

To illustrate, consider a 3^6 main effects design with a sample size of 500. Assume that every respondent gets every level of every attribute so that $K_i = K = 6$ and $L_{ik} = L_k = 3$ for all $k = 1,...,K$. For a typical orthogonal full-profile conjoint exercise, respondents would rate 18 products, leaving 5 degrees of freedom per respondent. In the full ICL model with all three rating types, the respondent would do 18 (= 6·3) level ratings, and, depending on what we assume about the number of general respondent types (T), as few as 4 total profiled product plus unique choice probability ratings. The tradeoff between assumed T and $J_i+M_i-S_i$ (the total number of profiled product ratings plus *unique* choice probability ratings) is shown in the following table.

Average df for ICL Model (3^6 Design)

	Assumed T			
Ji+Mi-Si	**2**	**5**	**10**	**12**
4	**5**	2	--	--
7	8	**5**	--	--
12	13	10	**5**	3
14	15	12	7	**5**
18	19	16	11	9

The combination of T and $J_i+M_i-S_i$ which matches the typical full-profile conjoint average df is shown bolded and underlined.

If general respondent types are conceptually similar to segments in segmentation analysis, where we rarely exceed 10 segments, then it makes sense to take 10 as the typical maximum value for T. Remember that, to accommodate the restrictions on **U**, it is necessary that $T \leq L\text{-}K$. I suggest here that we not consider values of T much greater than 10 regardless of L-K, which in the present example is 18 - 6 = 12. The above table indicates that $J_i+M_i-S_i = 12$ would suffice. Bear in mind that the number of actual ratings implied by $J_i+M_i-S_i$ is J_i+M_i so that $J_i+M_i-S_i = 12$ implies as few as 12 actual ratings (no choice probability ratings) or as many as 24 (no profiled product ratings and all scenarios having only two products).

Attribute level ratings are perceived by respondents to be less tiresome than profiled product or choice probability ratings. In fact, level ratings are a beneficial warm-up for profiled product or choice probability ratings (Huber *et al*, 1993). I find that breaking up the conjoint into these three different kinds of ratings makes it easier for respondents to remain attentive in what is cognitively a very difficult task.

With the flexibility of the ICL model, it is difficult to give completely general guidance on the product design matrices to be used in data collection. However, some suggestions are offered.

The primary rule for design is that **for any given respondent, there must be adequate ratings to identify all of that respondent's parameters**. In particular, the following relationship must hold for any individual respondent:

$$J_i + L_i + M_i - S_i > \delta_y + \delta_r(2K_i + 1) + T + 1$$

As a simple planning rule, the left side of the previous inequality should be, on average, at least 5 more than the right side plus T(L - K - T)/N, this last term usually being a small fraction. If $\delta_y = 1$, J_i should be > 1 to allow estimation of c_i; if $\delta_r = 1$, L_i should be $> 2K_i + 1$.

In addition, I recommend the following.

- If possible, have the respondent provide attribute level ratings for all levels of every attribute before either of the other ratings tasks. This provides the useful warm-up mentioned above and is an easy way to get a lot of information on the relative attractiveness of levels within attributes.
- If any attribute is excluded from the profiled product or choice probability ratings of most respondents (as represented by the design matrices $\mathbf{D}_i$ and $\mathbf{H}_{is}$), then you might as well consider the attribute to be dropped from the entire study. Without those data for an attribute, it is impossible to determine the proper scale of the general respondent types' partworths for the attribute relative to that of the other attributes.

- If attribute level ratings are completed before either of the other types of ratings, carry forward into a respondent's profiled product or choice probability ratings tasks only two levels for each attribute : the two levels being those which are rated highest and lowest for the attribute. So long as the direct attribute level ratings are present ($\delta_r = 1$), there are sufficient data to estimate all partworths for an attribute, even if a level of the attribute is not included in *any* respondent's profiled product or choice probability ratings.
- If physical constraints in the data collection process (such as the size of a computer screen in computerized interviewing) dictate that some attributes must be dropped from the profiled product or choice probability ratings tasks, then consider using only those for which the largest and smallest level ratings for the attribute are different (unless some attributes are required by design to be forced into everyone's profiled product or choice probability ratings tasks). If still more attributes must be dropped, then randomly select the attributes to include with probabilities proportional to the difference between an attribute's largest and smallest level ratings.
- With this method, an attribute gets dropped from the study only if nearly everyone rates all its levels the same. Obviously, if attributes are dropped from a respondent's profiled product or choice probability ratings tasks in this way, then the
  respondent cannot have fewer than two attributes in those tasks.
- For the choice probability ratings task(s), do not use any more than three products in a scenario. More than that is burdensome on respondents (probably inducing poorly considered ratings) and yields no additional information needed by the model. This is consistent with findings of Huber & Hansen (1987) concerning the number of attributes to use in the adaptive conjoint analysis approach.
- While it would theoretically be useful data, I do not recommend confusing respondents by including the same product twice within the set of products rated in the profiled product ratings task or twice within one scenario of the choice probability ratings task.
- The ICL model has built-in design flexibility at the individual respondent level: i.e., J_i, K_i, L_i, M_i, $\mathbf{D}_i$, $\mathbf{G}_i$, and $\mathbf{H}_{is}$ are individually defined for each respondent. Although the model can use any product profiles that generate adequate degrees of freedom for the profiled product and choice probability ratings tasks, I would recommend that you create a full design with the set of all product profiles you would use in an ordinary conjoint analysis (see Addelman, 1962 as one example). Then, for each respondent, randomly select from that set the product profiles to be included. Because the number of attributes may vary by respondent, then, for each

of the combinations of attributes that will be present for any respondent, you may need to design a different set of product profiles from which to select.

ATTRIBUTE INTERACTIONS

To this point, I have treated the ICL model as a "main effects" model, that is, the effects of attributes are simply additive with no allowance for interactions between the attributes. It is possible to allow 2-way (first order) interactions in this model. The following changes are necessary.

- If attribute level ratings are used, have the respondent rate the pairwise combination of levels of any two attributes involved in an interaction. Note that different a_{ik} and b_{ik}^2 parameters must still be estimated for each interaction effect and that σ_{ri}^2 applies to all main effect and interaction effect ratings as well.
- Assure that all parameters are estimable for whatever design matrices are used in generating the products being rated in the profiled product ratings or the choice probability ratings tasks. This includes proper structuring of the design matrices ($\mathbf{D}_i$ and $\mathbf{H}_{is}$) as well as the general respondent types' partworths (in $\mathbf{U}$) to include both main effect partworths and interaction effect partworths. See Finkbeiner & Lim (1991) for a discussion of designs including interaction effects.
- Impose sufficient constraints on $\mathbf{U}$ to uniquely identify the partworths. The restrictions imposed by use of equations (2a-b) apply only to main effects and not to interaction effects. To describe restrictions for the interaction effects, the definition of a block ($\mathbf{U}_k$) of partworths must be extended: there are K main effects $\mathbf{U}_k$'s and κ interaction effects, so that k runs from 1 through K and then from K+1 through K+κ. Each 2-way interaction effect has a pair of attributes associated with it: γ_{k1} and γ_{k2} are the first and second attributes in the interaction, with $L_{\gamma_{k1}}$ and $L_{\gamma_{k2}}$ levels, respectively. The $\mathbf{U}_k$ for an interaction effect (k in the range K+1,...,K+κ) has $L_{\gamma_{k1}} \times L_{\gamma_{k2}}$ rows, with the row corresponding to level ℓ_1 of attribute γ_{k1} and level ℓ_2 of attribute γ_{k2} being row number $(\ell_1 - 1) \times L_{\gamma_{k2}} + \ell_2$.

  In order to adequately restrict the kth interaction effect block, $\mathbf{U}_k$, to create $\tilde{\mathbf{U}}_k$, all the rows of the block corresponding to level 1 of attribute γ_{k1} must be set to $\mathbf{0}$ and then all of the rows corresponding to level 1 of γ_{k2} must be set to $\mathbf{0}$. Then, apply equation (2b) to the $\tilde{\mathbf{U}}_k$'s (whether defined for main effects or for interaction effects) to get the $\mathbf{U}_k$'s; the full $\mathbf{U}$ matrix will now have the required restrictions for both the main effects blocks and the interaction effects blocks.
- The above restrictions needed for estimating interaction effects have implications for the dimensionality of $\mathbf{U}$. With interaction effects, there are now

$L + \sum_{k=1}^{\kappa} L_{\gamma_{k1}} L_{\gamma_{k2}}$ rows in **U**, with $K + \sum_{k=1}^{\kappa} \left(L_{\gamma_{k1}} + L_{\gamma_{k2}} \right)$ restrictions. Since we wish to further restrict **U** so that a submatrix of T of its rows form an identity matrix, then the following must hold: $T \leq L - K + \sum_{k=1}^{\kappa} \left(L_{\gamma_{k1}} L_{\gamma_{k2}} - L_{\gamma_{k1}} - L_{\gamma_{k2}} \right)$.

- There are also implications for degrees of freedom of these new restrictions for interaction models:

$$df = \delta_y \left(\sum_{i=1}^{N} J_i - N \right) + \delta_r \left(\sum_{i=1}^{N} \left(L_i - 2K_i + \sum_{k=1}^{\kappa} L_{\gamma_{k1}} L_{\gamma_{k2}} - 2\kappa_i \right) - N \right) + \delta_x \sum_{i=1}^{N} \left(M_i - S_i \right)$$
$$- N(T+1) - T\left(L - K + \sum_{k=1}^{\kappa} \left(L_{\gamma_{k1}} L_{\gamma_{k2}} - L_{\gamma_{k1}} - L_{\gamma_{k2}} \right) - T \right)$$

EXAMPLE

To illustrate the ICL model, I use data from a study reported in Zwerina & Huber (1996). In this study, 50 MBA students were asked in a self-administered computerized questionnaire to provide a variety of choice likelihood ratings for laptop computers. The attributes used were:

Attributes	Levels
1. **Brand Name**	1. NEC
	2. IBM
	3. Toshiba
2. **Memory Size (RAM)**	1. 4 MB
	2. 8 MB
	3. 16 MB
3. **Hard Drive**	1. 250 MB
	2. 340 MB
	3. 510 MB
4. **Screen Type**	1. Passive Display (Mono)
	2. Dual Scan (Color)
	3. Active Display (Color)
5. **Price**	1. $3,995
	2. $3,149
	3. $2,459

In the first week of the study, respondents completed three tasks:

- Six holdout choice scenarios with three products per scenario.
- A direct level ratings task for each of the above 15 levels.
- A full-profile conjoint with 18 profiled product ratings.

In the second week of the study, respondents completed two tasks:

- The same six holdout choice task as in the first week.
- Thirty choice scenarios with three products in each.

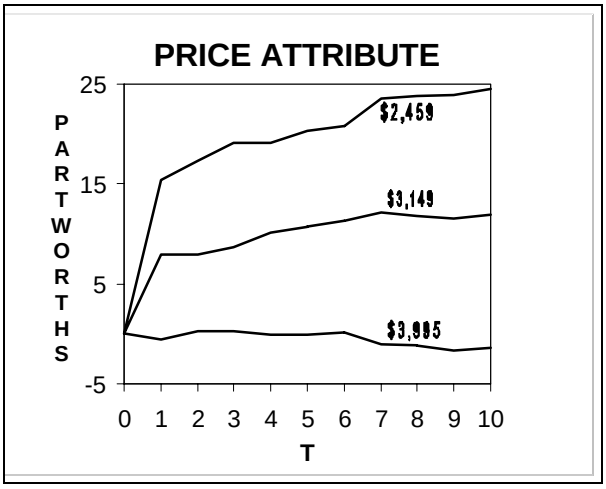
The same design was used for all respondents for the holdout and conjoint tasks. However, the designs for the choice tasks were individualized based on the direct level ratings, and were different for each respondent.

The repetition of the holdout choice tasks allowed estimation of the test-retest reliability of the holdout tasks by comparing one week's results to the other's. The reliability was reported as 77.3%, meaning that in 77.3% of the choice scenarios, respondents chose the same products from the same scenario in the second week as they had chosen in the first week.

In this problem, there are 93 (= 18+5×3+(30×3–30)) unique data points for each respondent, for a total of 4,650 across all 50 respondents. From these data, we must estimate 13+T (= 1+2×5+1+T +1) individual-level parameters for a total of 650+50T across all 50 respondents. We must also estimate T×(15–5–T) aggregate **U** parameters. Note that no **U** parameters are being estimated for T = 0 and T = 10; there is no point in going beyond T = 10 because **U** is not uniquely determinate. Thus, even for T set at its highest value of 10, there are plenty of data points from which to estimate all parameters: df = 3,500 and average df is 70 for T = 10. In fact, these data allow an ICL model with far more degrees of freedom than is typically provided for in most conjoint or choice tasks.

I estimated the full ICL model on these data for T (the number of general respondent types) ranging from 0 through 10. In general, the means of the partworths, $\mathbf{U}_{vi}$, tend to flatten out somewhat as T gets smaller. This is illustrated with the Price attribute in Figure 1.

Figure 1



This pattern is exactly the same as we see with "shrinkage" estimators in regression analysis, such as ridge regression or principal components regression. These methods also flatten (or "shrink") the regression model parameters closer together. In fact, the general respondent types used in the ICL model play exactly the same role as the principal components in principal components regression.

Since, in principal components regression with small samples, it is common that cross-validity improves with some number of principal components less than the maximum, it is not surprising that the ICL model produces better cross-validation on the holdout choice tasks with fewer than the maximum general respondent types. The table below shows that, similar to principal components regression, T = 10 produces the best fitting model to the data (with the smallest f value), but T = 6 or 7 produces the best cross-validation. (The "1st Ch. Validity" is the percent of individual-level first choices in all holdout tasks correctly predicted; the "MAE" is the mean absolute error in the aggregate choice shares vs. the holdout choice shares.)

Goodness of Fit & Validity

T	f	df	1st Ch. Validity	MAE
0	15,543	4,250	33.3%	21.0%
1	11,834	3,941	64.8%	9.6%
2	10,738	3,884	69.8%	8.0%
3	10,400	3,829	72.2%	7.8%
4	10,240	3,776	76.0%	7.1%
5	9,352	3,725	76.3%	6.9%
6	8,953	3,676	78.5%	6.6%
7	8,086	3,629	78.5%	6.6%
8	7,224	3,584	78.0%	7.4%
9	6,262	3,541	75.8%	7.8%
10	4,834	3,500	74.0%	7.6%

From the function values in the above table, we could compute the chi-square statistic to test for significance of difference for each successive value of T. All such differences are highly significant.

It is to be noted that the validities for T = 6 or 7 (78.5%) are higher than the test-retest reliability of 77.3%. This emphasizes the point that the holdout choice tasks themselves contain measurement error.

To provide an interesting comparison for the ICL results, consider some models presented by Zwerina & Huber. They report results for five different models, two of them being novel approaches to developing discrete choice models at the individual respondent level. These five models are:

- A self-explicated model using only the direct level ratings data
- A conjoint model, using only the profiled product ratings data
- An individual-level discrete choice model using only the choice data (but using the direct level ratings to select "best" designs for each respondent)
- A conjoint model with attribute level order constraints using both the level ratings and the profiled product ratings
- A choice model with attribute level order constraints using both the level ratings and the choice ratings

Zwerina & Huber's results, in comparison to the best ICL results (for T = 6), are shown in the table below.

Model Comparison

Model	1st Ch. Validity	MAE
Self-Explicated	65.3%	12.7%
Conjoint	64.7%	12.2%
Individual Choice	73.3%	3.2%
Constrained Conjoint	70.2%	8.0%
Constrained Ind. Choice	76.3%	4.1%
ICL (T = 6)	78.5%	6.6%

It is of interest to note that First Choice Validity improves with the inclusion of the direct level ratings; it improves even more with inclusion of all three data types. The Mean Absolute Error, however, shows that inclusion of direct level ratings worsens these aggregate errors for the individual choice model, but improves the conjoint model errors, while the ICL model (which includes all three types of data) falls almost exactly half-way between the constrained conjoint and constrained individual choice models. Given the error in the holdout tasks themselves, it is not clear what differences here are statistically significant ones.

While the inclusion of all three data types may have helped, it does not eliminate from the ICL model all preference order violations in the partworths. The last four of the five attributes (Memory Size, Hard Drive, Screen Type, and Price) have an expected preference ordering of their levels, which can be established based on judgment alone. The direct level ratings do not violate this expected preference ordering on any of the four attributes for any respondent. However, the ICL partworths violate the expected preference ordering for ten of the respondents (20% of the 50). Among the specific attributes, Memory Size is the best, with one respondent showing violation of expected order, and Price is the worst, with seven respondents showing violations of the expected order.

It must be pointed out that the ICL model can provide estimates with considerably less data than was provided in this study. There may be some loss in validity as a result. For example, I selected 8 profiled product ratings and 4 choice scenario ratings for each respondent (4 choice scenarios produce 8 unique choice ratings) and estimated the ICL model with T = 10. The total degrees of freedom in this case are 400, an average of 8 df per respondent. The First Choice Validity for this model is 66.3% and the Mean Absolute Error is 7.0%. (Using smaller values of T here worsens the cross-validity.)

This result suggests that the considerably smaller degrees of freedom results in worse individual-level validity for the ICL model, but very little degradation in its aggregate validity. The benefit of what turns out to be a nearly 80% savings in respondent time for completing the profiled product and choice scenario ratings may well be worth the relatively small price of slightly worse validity.

CONCLUSION

On the downside, the ICL model is a complex one that is computationally time-consuming in estimation. In a sense, this is a technical problem that will be resolved by waiting for the computer hardware to get faster. To put this in perspective, when I first started working on this problem over 10 years ago, the algorithm used in the present solution would have taken weeks to run even small problems on then existing desktop computers. It now runs in a few hours for all of the required solutions for T.

The violations of expected preference ordering on the partworths produced by the ICL model are somewhat troubling. While this model is not alone in producing such violations, I believe it would benefit from forcing the partworths to be rank-order consistent with the direct level ratings which do not violate expected ordering. That way, respondents would be allowed to prefer, for example, higher-priced items if they wished, without causing the researcher to worry that such a preference is not simply due to error from the use of partial designs or from small degrees of freedom.

The ICL model proposed in this paper is very general in that it subsumes several published conjoint and choice models and identifies additional submodels. Because of its generality, it is statistically efficient in making use of a variety of data types. The use of general respondent types, similar to the principal components in principal components regression, improves the stability of the estimates resulting in good predictive validity. The choice model component of the ICL model uses a special case of the multinomial probit model for individual respondents which is non-IIA and is very easy to implement for choice share simulation.

Because the model is very flexible in terms of the attributes, attribute levels, or products included in the ratings tasks, it is ideally suited to the analysis of the kind of individually tailored designs which computerized interviewing makes possible.

Furthermore, its use of multiple data types may create opportunities for decreased respondent burden by allowing reduction in the amount of time spent on any single data type. In the example, we saw good aggregate validity and adequate individual-level validity from a design which would require the typical respondent to spend about 2 minutes rating 15 attribute levels, then about a minute rating 8 profiled products and slightly less rating 4 choice scenarios. Because this exercise moves quickly from one type of task to another, it will feel fast and interesting to respondents. By contrast, the full data collected by Zwerina & Huber required the typical respondent to spend about 2 minutes

rating 18 profiled products and about 5½ minutes on 30 choice ratings, after spending the 2 minutes on 15 attribute level ratings. (Note: these times do not include time spent reading and understanding instructions which can be considerable.)

It would be of substantial interest to determine some of the cost/benefits tradeoffs of using different amounts of data of each type. In particular, it would be useful to compare the various submodels to one another on their efficiency and validity.

DATA NOTATION

γ_1,γ_2 Attributes 1 and 2 in the 2-way interaction k ($k = K+1,...,K+\kappa$)

δ_r = 1 if attribute level ratings are present; = 0 if not

δ_x = 1 if choice probability ratings are present; = 0 if not

δ_y = 1 if profiled product ratings are present; = 0 if not

ε A tiny positive number (like 10^{-8}) used to keep $\hat{\sigma}^2_{ri}$ positive and $\mathbf{\Sigma}_{zis}$ positive definite.

κ Number of 2-way interaction effects included in the model

$\mathbf{1}_n$ n-vector of ones

$\mathbf{D}_i$ $J_i \times L_i$ design matrix for respondent i (i = 1...N); each row is the design specification for a product profile; each column corresponds to one of respondent i's attribute levels; a cell of $\mathbf{D}_i$ is designated d_{ijkl} which is the value for respondent i, product j, attribute k, level l; if an attribute is dummy-coded, $d_{ijkl} = 1$ if respondent i's level l of attribute k is present in product profile j and = 0 if not; if an attribute is a continuous variable, d_{ijkl} is respondent i's value on that variable (attribute k) and l is 1; different respondents may have different products specified in $\mathbf{D}_i$

$\mathbf{G}_i$ $L_i \times L$ matrix indicating which attributes and levels are included in the entire ICL model for respondent i: formed by dropping from the $L \times L$ identity matrix any row corresponding to the level(s) dropped for respondent i (an entire attribute is effectively dropped by dropping all rows from $\mathbf{G}_i$ which correspond to all levels of that attribute) (i = 1...N)

$\mathbf{G}_{ik}$ $L_{ik} \times L$ matrix indicating which levels of attribute k are included for respondent i; formed by dropping from $\mathbf{G}_i$ all rows not corresponding to levels of attribute k ($i = 1...N$; $k = 1...K_i$)

$\mathbf{H}_i$ $M_i \times L_i$ design matrix comprised of $\begin{bmatrix} \mathbf{H}_{i1} \\ \vdots \\ \mathbf{H}_{iS_i} \end{bmatrix}$ (i = 1...N)

$\mathbf{H}_{is}$ $M_{is} \times L_i$ design matrix for the products in scenario s for respondent i, constructed in the same fashion as $\mathbf{D}_i$ ($i = 1...N$; $s = 1...S_i$)

$\mathbf{I}_n$ $n \times n$ identity matrix

J_i Number of profiled products rated by respondent i (i = 1...N)

K	Full number of attributes
K_i	Number of attributes for respondent i (i = 1...N)
L	Total number of levels across all attributes ($\sum_{k=1}^{K} L_k$)
L_i	Number of levels across all attributes for respondent i (i = 1...N; $\sum_{k=1}^{K_i} L_{ik}$)
L_{ik}	Number of levels in attribute k for respondent i (i = 1...N; k = 1...K_i)
L_k	Number of levels in the full attribute k (k = 1...K)
M_i	Number of products across all choice scenarios (i = 1...N; $= \sum_{s=1}^{S_i} M_{is}$)
M_{is}	Number of products represented in respondent i's choice scenario s (i = 1...N; s = 1...S_i)
N	Number of respondents
N_{is}	The number of hypothetical occasions on which respondent i chooses one of the M_{is} products in i's scenario s
$\mathbf{Q}_X$	Equicorrelation matrix of order X: $(1 - \rho_i)\mathbf{I}_X - \rho_i \mathbf{1}_X 1'_X$
$\mathbf{r}_{ik}$	L_{ik}-vector of attribute k's level ratings by respondent i (i = 1...N; k = 1...K_i)
S_i	Number of choice scenarios for respondent i (i = 1...N)
w_i	Sample weight for respondent i
$\mathbf{x}_i$	M_i-vector comprised of concatenated vectors $\mathbf{x}_{is}$ (i = 1...N; s = 1...S_i)
$\mathbf{x}_{is}$	M_{is}-vector with elements x_{ism}; rated choice probabilities in respondent i's choice scenario s; $0 \le x_{ism} \le 1$ and $\sum_{m=1}^{M_{is}} x_{ism} = 1$ (i = 1...N; s = 1...S_i)
$\mathbf{y}_i$	J_i-vector of profiled (full- or partial-profile) product ratings by respondent i (i = 1...N)

PARAMETER NOTATION

θ_i	$\frac{\pi}{\sqrt{6\sigma_i^2(1-\rho_i)}} = \frac{\pi}{\sigma_i\sqrt{6}}$ when $\rho_i = 0$
$\boldsymbol{\mu}_{yi}$	$\mathcal{E}(\mathbf{y}_i) = c_i \mathbf{1}_{J_i} + \mathbf{D}_i \mathbf{G}_i \mathbf{U} \mathbf{v}_i$
$\boldsymbol{\mu}_{rik}$	$\mathcal{E}(\mathbf{r}_{ik}) = a_{ik} \mathbf{1}_{L_{ik}} + b_{ik}^2 \mathbf{G}_i \mathbf{U} \mathbf{v}_i$
$\boldsymbol{\mu}_{zi}$	M_i-vector comprised of concatenated vectors $\boldsymbol{\mu}_{zis}$ (i = 1...N; s = 1...S_i)
$\boldsymbol{\mu}_{zis}$	$\mathcal{E}(\mathbf{z}_{is}) = \mathbf{H}_{is} \mathbf{G}_i \mathbf{U} \mathbf{v}_i$
ρ_i	Respondent i's error correlation across all products for $\mathbf{y}_i$ and $\mathbf{z}_{is}$ (i = 1...N; s = 1...S_i)
σ_i^2	Respondent i's error variance for $\mathbf{y}_i$ and $\mathbf{z}_{is}$ (i = 1...N; s = 1...S_i)

σ_{ri}^2 Respondent i's error variance for every $\mathbf{r}_{ik}$ (i = 1...N; k = 1...K_i)

$\boldsymbol{\Sigma}_{yi}$ Respondent i's error covariance matrix for $\mathbf{y}_i$ (i = 1...N)

$\boldsymbol{\Sigma}_{rik}$ Respondent i's error covariance matrix for $\mathbf{r}_{ik}$ (i = 1...N; k = 1...K_i)

$\boldsymbol{\Sigma}_{zis}$ Respondent i's error covariance matrix for $\mathbf{z}_{is}$ (i = 1...N; s = 1...S_i)

$\Phi(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ = multivariate normal probability density function

$$= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $|\boldsymbol{\Sigma}| > 0$ and n = number of elements of $\mathbf{x}$ or $\boldsymbol{\mu}$

$\Psi(\mathbf{x}|\mathbf{p},N)$ = multinomial probability density function

$$= \left(\frac{N!}{\prod_{i=1}^{n}(x_i!)}\right)\prod_{i=1}^{n}p_i^{x_i} \text{ where } 0 \le p_i \le 1,\ \mathbf{1}_n'\mathbf{p} = 1,\ x_i \ge 0,\ \mathbf{1}_n'\mathbf{x} = N,\ n = \text{dimension of } \mathbf{p} \text{ or } \mathbf{x}$$

a_{ik} Additive constant in model for $\mathbf{r}_{ik}$, one constant for each of respondent i's attributes (i = 1...N; k = 1...K_i)

b_{ik}^2 Coefficient for attribute k in model for $\mathbf{r}_{ik}$ (must be ≥ 0) (i = 1...N; k = 1...K_i)

c_i Additive constant in model for $\mathbf{y}_i$

$\mathbf{e}_{ri}$ L_i-vector comprised of concatenated vectors $\mathbf{e}_{rik}$ (i = 1...N; k = 1...K_i)

$\mathbf{e}_{rik}$ L_{ik}-vector of error terms for modeling $\mathbf{r}_{ik}$ (i = 1...N; k = 1...K_i)

$\mathbf{e}_{yi}$ J_i-vector of error terms for $\mathbf{y}_i$, respondent i's profiled product ratings (i = 1...N)

$\mathbf{e}_{zis}$ M_{is}-vector of error terms for $\mathbf{z}_{is}$, respondent i's product total utilities in scenario s (i = 1...N; s = 1...S_i)

$\mathcal{M}(\mathbf{p},N)$ Multinomial distribution with probabilities $\mathbf{p}$ and number of choice occasions N and with probability density function $\Psi(\mathbf{x}|\,\mathbf{p},N)$

$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and with probability density function $\Phi(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$

$\mathbf{p}_i$ M_i-vector comprised of concatenated vectors $\mathbf{p}_{is}$ (i = 1...N; s = 1...S_i)

$\mathbf{p}_{is}$ M_{is}-vector with elements p_{ism} = respondent i's probability in scenario s that product m has the largest z_{ism} (i = 1...N; s = 1...S_i; m = 1...M_{is})

P_{sm} Choice share for product m in scenario s = $\mathcal{E}_i(p_{ism})$ where the expectation is over individuals

T Number of general respondent types (a.k.a. latent classes)

$\mathbf{U}$ — $L \times T$ matrix of partworths on all L possible attribute levels for each of T general types of respondent (a.k.a.: points of view or latent classes); $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_K \end{bmatrix}$; some collection of T rows of $\mathbf{U}$ (not the first one for each attribute) is restricted to be $\mathbf{I}_T$; note the change in restrictions and structure for $\mathbf{U}_k$ and $\mathbf{U}$ when interaction effects are included (see the **Attribute Interactions** section)

$\mathbf{U}_k$ — $L_k \times T$ matrix of partworths on the L_k possible levels of attribute k for each of T general types of respondent (a.k.a.: points of view or latent classes); first row is restricted to be **0;** note the change in restrictions and structure for $\mathbf{U}_k$ and $\mathbf{U}$ when interaction effects are included (see the **Attribute Interactions** section)

$\mathbf{v}_i$ — T-vector of coefficients for combining the T general respondent types' partworths into the partworths for respondent i ($i = 1...N$)

$\mathbf{z}_{is}$ — M_{is}-vector of product total utilities plus error terms in scenario s for respondent i ($i = 1...N$; $s = 1...S_i$)

REFERENCES

Addelman, S. "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments," *Technometrics*, 4, 1962, 21–46.

Bock, D. Multivariate Statistical Methods in Behavioral Research, McGraw-Hill, New York, 1975.

Clark, C.E. "The Greatest of a Finite Set of Random Variables," *Operations Research*, 9, 1961, 145–162)

Cooper, L.G. & Finkbeiner, C.T. "A Composite MCI Model for Integrating Attribute and Importance Information," *Advances in Consumer Research*, Association for Consumer Research, Provo, UT, 11, 1984, 109–113.

Efron, B. The Jackknife, the Bootstrap, and Other Resampling Plans, Society for Industrial And Applied Mathematics, *CBMS-NSF Regional Conference Series in Applied Mathematics*, 38, Philadelphia, 1982.

Finkbeiner, C.T. "Individual Differences Probit," *ORSA/TIMS Marketing Science Conference*, ORSA/TIMS, Nashville, March, 1985.

Finkbeiner, C.T. "Simplified Multinomial Probit," *ORSA/TIMS Marketing Science Conference*, ORSA/TIMS, Dallas, March, 1986.

Finkbeiner, C.T. "Comparison of Conjoint Choice Simulators," *Sawtooth Software Conference Proceedings*, Sawtooth Software, Ketchum, ID, 1988, 75–103.

Finkbeiner, C.T. “Alternative Applications of Preference Models to Customer Satisfaction Research,” *Sawtooth Software Conference Proceedings*, Sawtooth Software, Ketchum, ID, 1992, 127–159.

Finkbeiner, C.T. & Lim, P.C. “Including Interactions in Conjoint Models,” *Sawtooth Software Conference Proceedings*, Sawtooth Software, Ketchum, ID, 1991, 271–298.

Green, P.E., Goldberg, S.M., & Montemayor, M. “A Hybrid Utility Estimation Model for Conjoint Analysis,” *Journal of Marketing*, 45, 1981, 33-41.

Green, P.E., Krieger, A.M., & Agarwal, M.K. “Adaptive Conjoint Analysis: Some Caveats and Suggestions,” *Journal of Marketing Research*, 28, 1991, 215–221.

Green, P.E., & Srinivasan, V. “Conjoint Analysis in Consumer Research: Issues and Outlook,” *Journal of Consumer Research*, 5, 1978, 103-123.

Green, P.E. & Wind, Y. “New Way to Measure Consumers’ Judgments,” *Harvard Business Review*, July, 1975, 107–117.

Gumbel, E.J. “Bivariate Logistic Distributions,” *Journal of the American Statistical Association*, 56, 1961, 335–349.

Hagerty, M.R. “Improving the Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis,” *Journal of Marketing Research*, 22, 1985, 168–184.

Huber, J. & Hansen, D. “Testing the Impact of Dimensional Complexity and Affective Differences in Adaptive Conjoint Analysis,” *Advances in Consumer Research*, Association for Consumer Research, Provo, UT, 14, 1987, 159–163.

Huber, J. & Moore, W. “A Comparison of Alternative Ways to Aggregate Individual Conjoint Analysis,” *Educators’ Conference Proceedings*, American Marketing Association, Chicago, 1979, 64–68.

Huber, J.C., Wittink, D.R., Fiedler, J.A., & Miller, R. “The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice,” *Journal of Marketing Research*, 30, 1993, 105–114.

Johnson, R.M. “Trade-Off Analysis of Consumer Values,” *Journal of Marketing Research*, 11, 1974, 121–127.

Johnson, R.M. “Adaptive Conjoint Analysis,” working paper, Sawtooth Software, Ketchum, ID, 1987a.

Johnson, R.M. “Adaptive Conjoint Analysis,” *Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis, and Computer Interviewing*, Sawtooth Software, Ketchum, ID, 1987b, 253–266.

Kamakura, W., & Srivastava, R.K. “Predicting Choice Shares under Conditions of Brand Interdependence,” *Journal of Marketing Research*, 21, 1984, 420-34).

Lakshmi-Ratan, R., Chaiy, S., & May, J. "Mathematical Modeling of Contextual Effects on Individual Choice Behavior: Axiom and Model on Contextual Choice," working paper, University of Wisconsin, Graduate School of Business, September, 1984.

Lazarsfeld, P.F. & Henry, N.W. Latent Structure Analysis, Houghton Mifflin, Boston, 1968.

Leigh, T.W., MacKay, D.B., & Summers, J.O. "On Alternative Methods for Conjoint Analysis," *Advances in Consumer Research*, Association for Consumer Research, Provo, UT, 11, 1984, 317–322.

Louviere, J.J. Analyzing Decision Making: Metric Conjoint Analysis, Sage, *Series: Quantitative Applications in The Social Sciences*, 67, 1988.

Louviere, J.J. & Woodworth, G.G. "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20, 1983, 350–367.

Luce, R.D. Individual Choice Behavior. Wiley, New York, 1959.

McFadden, D. "Conditional Logit Analysis of Qualitative Choice Behavior," In P. Zarembka (Ed.), Frontiers in Econometrics, Academic Press, New York, 1970, 105-142.

Oliphant, K., Eagle, T.C., Louviere, J.J., & Anderson, D.A. "Cross-Task Comparison of Ratings-Based and Choice-Based Conjoint," *Sawtooth Software Conference Proceedings*, Sawtooth Software, Ketchum, ID, 1992, 383–404.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. Numerical Recipes in C, 2nd edition, Cambridge University Press, Cambridge, 1992.

Suppes, P. & Zinnes, J.L. "Basic Measurement Theory," In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), Handbook of Mathematical Psychology, Vol. I, Wiley, New York, 1963.

Thurstone, L.L. "The Prediction of Choice," *Psychometrika*, 10, 1945, 237–253.

Tucker, L.R & Messick, S. "An Individual Differences Model for Multidimensional Scaling," *Psychometrika*, 28, 1963, 333–367.

Zwerina, K. & Huber, J. "Deriving Individual Preference Structures from Practical Choice Experiments," unpublished manuscript, 1996.

NEURAL NETWORKS AND STATISTICAL MODELS

Tony Babinec
SPSS Inc.

INTRODUCTION

Neural networks are often presented as "simulated biological intelligence" or some such thing that "learns the patterns in your data." One sometimes gets the impression that applying neural networks is like starting your car—you just do it. In reality, as in conventional statistical modeling, one must invest a lot of "sweat equity" and think through one's problem when applying neural nets. I much prefer getting very far away from the brain and learning analogies, and instead prefer to think of neural networks as a flexible form of regression analysis or discriminant analysis. What is meant by this will be illustrated below.

We will proceed as follows. The Overview will develop the one hidden layer Multi-Layer Perceptron. Next, we will present an extended regression example. After that, we will briefly look at discriminant analysis and time series prediction. Finally, we will make a few summary points.

OVERVIEW

There are many types of neural networks. In this paper we restrict our attention to supervised neural networks, which are nonlinear mappings from an input **x** to an output **y** = **f(x; w)**. The parameters of the neural net are denoted by **w**. Of the various supervised neural nets, we will focus on the well-studied Multi-Layer Perceptron (MLP).

The *1-hidden layer MLP* can be expressed in the following form (MacKay, 1995):

$$hj = f^{(1)}\left(w_0^{(1)} + \Sigma_k w_{jk}^{(1)} x_k\right) ; \; y_i = f^{(2)}\left(w_0^{(2)} + \Sigma_j w_{ij}^{(2)} h_j\right) \qquad (1)$$

where, for example,

- $f^{(1)}(a) = \tanh(a)$,
- $f^{(2)}(a) = a$,

and

- the x are input variables,
- the y are output variables,
- the h are hidden nodes,
- the weights w are parameters to be estimated.

In words, the above model form says: *The response variable is a linear combination of nonlinear functions of linear combinations of the input variables.* The nonlinearity of $f^{(1)}$ at the "hidden layer" gives the neural network its computational flexibility. While the model might strike the reader as strange, there is precedent in the statistical literature for approximating a functional form by the sum of basis functions such as polynomials of increasing order or fourier components. The "hidden nodes" are basis functions that may or may not be interesting in their own right, but enable the modeler to fit arbitrary functional forms.

The 1-hidden layer MLP turns out to be a flexible model form that subsumes some well-known methods as special cases. For example, consider the part of equation (1) before the semicolon and replace *h* by *y*. Then, with $f^{(1)}$ linear, the model form is the usual linear regression. With $f^{(1)}$ sigmoidal, then the model form is logistic regression or the generalized linear model. Or, consider equation (1) as shown. If both $f^{(1)}$ and $f^{(2)}$ are linear in form, then equation (1) is the same as linear regression, though with "redundant" parameters. Finally, with $f^{(1)}$ nonlinear in form, the 1-hidden layer neural net becomes a flexible regression tool.

The reason that neural nets have garnered interest among data modelers is that various existence theorems have shown that the 1-hidden layer MLP in the form of equation (1) is a "universal approximator" (Bishop, 1995). That is, these networks can approximate arbitrarily well any functional continuous mapping from one finite-dimensional space to another, provided the number of hidden units is sufficiently large. An important corollary of this is that, in the context of a classification problem, this network can approximate any decision boundary to arbitrary accuracy. Thus model form (1) is also a form of nonlinear discriminant analysis. Having said that, a practical difficulty in the use of neural nets is that there is little theoretical guidance for choosing the number of hidden nodes in a particular data analytic context.

Given data in the form of values for *x* and *y*, and given a particular specification of (1), the network is "trained" to fit a data set by minimizing an error function such as the residual sum of squares, that is, the sum of squares of the differences between the observed *y* values and those predicted by the model. This requires starting values for the weights; typically, these are small random numbers. The error function is minimized by using some optimization method that makes use of the gradient of the error function, which can be evaluated using "backpropagation" (the chain rule). The details are skipped here. The interested reader can find a treatment of this in Bishop, 1995 or Warner and Misra, 1996. A practical issue is that gradient-based search methods are "greedy," and as the error surface in general can be complicated in shape, there is potential for the neural network to find a local minimum and not a global minimum.

Because neural nets are universal approximators, they can fit any data arbitrarily well. However, in general, the goal of network training is not to learn an exact representation of the training data itself, but rather to build a statistical model of the process which generates the data. This is important if the network is supposed to generalize well, that is, make good predictions for new inputs. In general, a neural network with too few hidden nodes will miss important features in the data, while a neural network with too many hidden nodes will fit some of the noise in the training data. This is often expressed as *the bias-variance tradeoff*. A model which is too simple or inflexible will have large *bias*, while one which is too flexible given the data at hand will have large *variance*. The best generalization is attained when there is balanced attainment of the goals of small bias and small variance, and to do so, one must control the complexity of the model. Failure to do so can result in overfitting, wherein a model performs well on the data at hand but poorly on like data.

The neural network literature, and active research today, have addressed the issue of model complexity. To avoid overfitting, one can:

- Use models with fewer degrees of freedom—A smaller network.
- Limit the number of iterations—Stop early.
- Change the objective function—Regularization.
- Penalize complexity—Penalty functions.

The first bulleted point suggests a *model comparison* approach. Use networks of differing degrees of complexity and observe the reduction in error at each step.

The second bulleted point suggests the method of *early stopping* or *stopped training*. During a typical training session, for a given network, the error for the *training* data typically decreases as a function of the number of iterations. However, the error measured with respect to independent data, generally called a *validation* set, often shows a decrease at first, followed by an increase as the network starts to overfit the training data. Training can therefore be stopped at the point of smallest error in the validation set. While the error in the validation set is not an estimate of true error, the network weights at this point would be expected to have the best generalization performance.

Early stopping is relatively fast. The researcher must decide what proportion of data to allocate to the validation set, and should use some random mechanism for choosing cases. If an estimate of generalization error is desired, some data must also be set aside in a true holdout sample. The use of sample splitting might strike the statistician as being inefficient, but in some application areas there is a lot of data, which makes splitting seem less problematic.

The third bulleted point suggests adding a penalty function to the usual error sum of squares. One choice here is the "weight decay" term involving the sum of squares of the parameters in the neural net. The effect of adding this term into the usual error term is to force smaller-sized weights and smoother network mappings.

The fourth bulleted point suggests the use of criteria in the mold of the Akaike Information Criterion, that is, combine the training error with some term which grows as model complexity grows. Thus, if the model is too simple the penalty criterion will be large due to the residual training error, while if the model is too complex the penalty criterion will be large due to complexity term.

A REGRESSION EXAMPLE

The ideas expressed in the previous section are now illustrated with a "toy" data example.

First, we generated 141 data points using the generating function:

$$y = 1.1*(1-x+2*x^2)*\exp^{(-x*x/2)}; \text{ x from -2 to 5 by 0.05.} \quad (2)$$

Second, we added noise with a mean of 0 and a standard deviation of 0.15 to the *y* values.

Third, we randomly allocated the data into a training data set of 100 observations and a validation data set of 41 observations.

Fourth, we generated a holdout data set of 100 observations using the same generating function as for the training data, but with *x* values chosen randomly (uniform distribution) over the interval [-2, 5]. In this way, we insured that the holdout data did not have the same *x* values (in general) as the training and validation data.

Figure 1 shows the data generated by equation (2).

Figure 2 shows the data with added noise.

Figure 3 shows the holdout data along with the training and validation data. Note that the holdout data do not have added noise.

The form of the data is: A general trending down from upper left to lower right in the plot, with two noticeable humps.

Figure 4 shows the linear regression fit to the training data. Linear regression in this instance estimates two parameters. This simple model form captures the downward trend in the data, but also shows bias.

Figure 5 shows the fit of a 4^{th} order polynomial in *x* to the data. Here again, the model captures the downward trend in the data, but misses the humps.

We next fit MLPs of increasing order of complexity to the *training data only*. The results applied to the *training data* are shown in Figures 6 through 13. In these figures, we see a characteristic pattern. The one hidden node MLP fits a monotonically decreasing s-shaped curve to the data. The two hidden node MLP fits two humps, but misses the pattern in the data. The three hidden node MLP turns out to fit the training data rather well. When viewed against the plot of the generating function in Figure 1, the progression of Figures 9 through 12 shows increased overfitting, that is, the neural network progressively fits the noise in the data as the number of nodes goes up. See Table 1 for the values of the root mean square error for each of these fits. Table 1 also shows the number of network weights estimated for each MLP. Of course, the neural net worked with the observed training data shown in Figure 2, and not the theoretical data in Figure 1. Figure 13 superimposes the generating function, the observed data, and the 20 node MLP to show the overfitting.

The next set of figures, Figures 14 through 20, shows the results of applying the *training data only* fitted model to the *holdout sample*. Again, we see that the one-node and two-node networks do not fit. The progression of Figures 16 through 20 shows that the higher-node networks fit the data less well. This shows the relatively poorer generalization of the overfitted neural networks. See Table 1 for the values of the root mean square error for each of the fits.

Figures 21 through 28 repeat the analysis of Figures 6 through 13, however this time the method of early stopping is used. The training and validation sets are shown in Figure 2. We fit MLPs of increasing order of complexity to the training data, and use the error in the validation set at each iteration as an indication of "best" fit. Figures 21 through 28 show the fit on the training data when early stopping is used. Visually, the 3-node and 4-node networks fit well. The higher-node networks do not appear to exhibit much overfitting, with the exception of the 20-node network.

Figures 29 through 35 show the fit to the holdout data when stopped training is used. In general, our conclusions here are the same as for the preceding set of figures.

In sum, our example shows that a neural network can be used to fit a regression line (conditional mean line) when the functional form of the data is not known. To do this well, one must employ approaches that control the complexity of the fitted model. We partitioned the data into sets – training, validation (which is really an "adjunct" to training when early stopping is used), and holdout. The combination of a judicious choice of number of hidden nodes plus the use of early stopping led to a "good" network. Sarle, 1995 looked at early stopping and other methods for preventing overtraining, and found that while early stopping worked okay, the use of a weight decay term in the error worked even better.

CLASSIFICATION

Discriminant analysis is often used in classification. Discriminant analysis makes the assumptions that the data cases are independent, the data are multivariate normal within groups, and the data have homogeneous covariance matrices across groups. Discriminant analysis finds a linear combination of the discriminating variables that maximally separates the group centroids. As a byproduct, when assumptions are met, discriminant analysis finds simple near-linear boundaries that separate groups in such a way as to minimize the total number of misclassifications. Here, we present two well-known examples where groups do not separate well in their means.

Figure 36 shows a version of the "x-or" (exclusive or) data with jittered values. The points in the "1" group should be seen as a diagonal ridge that splits the "0" group. The point of this example is that there is no single line that separates the groups well. Figure 37 shows group assignment when linear discriminant analysis is applied to the data. While the two groups are theoretically centered at the centroid (0.5, 0.5), their group centroids are in fact not coincident due to noise. Therefore, discriminant analysis can find a direction of separation, though it does not work very well.

Figure 37 shows the fitted function that results from applying a one-hidden layer MLP with two hidden nodes to the same data after partitioning into training, validation, and holdout sets. Note that the neural network fits the diagonal ridge that we see in the data. In effect, there are two near-linear boundaries for group assignment, not one.

A second example often used in classification benchmarks is two nested spherical distributions. Figure 38 shows a two-dimensional version of this. Again, the idea is that the two groups do not differ in their centroids, however there is a very definite boundary between the groups. Once again, linear discriminant analysis does not work well (results not shown). What's more, other classification approaches such as CART do not work well, though in fairness to CART, the recent "bagging" approach will work well in classification. Again, although we do not show the figure, a one-hidden layer MLP can find the nonlinear boundary between the groups.

TIME SERIES

A chaotic time series can be thought of as a nonlinear deterministic series with or without added noise. These series exhibit bounded behavior, and can show pseudo-periodicities. In a way that we will shed light on below, these series also exhibit sensitive dependence on initial conditions. The key point here is that traditional linear methods do not necessarily work well on these time series. Use of such identification tools as the autocorrelation function might wrongly lead the researcher to conclude that there is no exploitable pattern in the data. In other words, while a white noise series produces an autocorrelation function with negligible autocorrelations, one cannot infer from negligible autocorrelations that the series is white noise. This is illustrated via the well-known logistic series.

Figure 40 shows 200 observations of the logistic series:

$$y_t = 4*y_{t-1}*(1-y_{t-1}) \; ; \; y_0 = 0.99$$

The series is bounded and appears to have some pattern to it.

Figure 41 shows the autocorrelation function plot for this series. The autocorrelations are all negligible and fall within the +/- 2 standard error limits.

Figure 42 shows the "phase plot," that is, the plot of y_t versus y_{t-1}. It is easy to see the deterministic relationship between *y* at time *t* and *y* at time *t-1*.

We held out the last 20 points, divided the first 180 into training and validation sets, and fit an MLP with 2 hidden notes. We then forecast forward from point 180 through the next 17 points in the holdout sample. Figure 43 shows that the forecast function tracks the holdout data well for about 5 or 6 points, and then the forecast and the holdout data divurge. This is an illustration of the fact that chaotic series may or may not be forecastable in the short term, but in the long term they are unforecastable—that is, there is divergence between the forecast function and the observed holdout data.

DISCUSSION

The above "toy" examples show contexts where neural networks could be expected to outperform conventional methods in prediction or classification. The commercial researcher ought to consider using neural networks when:

1. The functional form relating input variables to the response variable is not known or well understood, but is not thought to be linear.
2. There is a large sample of data.
3. A premium exists for better prediction that makes it worth the added effort to fit a well-tuned neural network.

Regarding the first point, there exist difficult real world problems for which theory is not well-articulated. The researcher might be willing to argue that certain variables are predictive of the response, but might not be willing to assume a linear functional form. On the other hand, regression and related parametric methods perform better when theory or experience indicate an underlying functional form.

Regarding the second point, because there is a lot of data, parameter-rich modeling approaches can and ought to be considered. The researcher might possibly be able to identify and model subtle patterns in the data. On the other hand, don't use a complicated model when a simple one will do. What's more, apply one or more simpler statistical approaches and compare results. Also, parametric methods such as regression can perform better for small sample sizes.

Regarding the third point, in commercial contexts better predictions lead to monetary results and might be worth the effort. And, there will be effort. Neural networks have the disadvantage that convergence to a solution can be slow, can depend on the network's initial conditions, and is not guaranteed to find a global optimum. This point also raises the issue of prediction "versus" explanation. From the standpoint of "structural" understanding, the estimated weights in a neural network are generally not of interest, and also are generally not easily interpreted. For one thing, the parameters in a one-hidden layer neural network are unidentifiable. This is known as the phenomenon of "weight-space symmetry." That is, for M hidden units, given any set of weights, there are $M!2^M$ equivalent set of weights. In the end, neural networks are therefore more useful for prediction than for understanding.

In conclusion, one should not blindly apply neural networks. There is no substitute for exploratory data analysis. When establishing causality is of interest, there is no substitute for designed experiments. In the examples we showed, the use of holdout data was illustrated. Cross-validation is important, or one risks believing a model that has merely found the noise in one's particular data. Finally, architecture choices are still an issue. There is no such thing as running a neural net "in general." The researcher must make decisions about the number of hidden nodes, the number of layers, the allocation of data to sets, and so on.

REFERENCES

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford: Clarendon Press.

MacKay, David J.C. (1995). Bayesian non-linear modeling for the Prediction Competition. In G. Heidbreder (Ed.), Maximum Entropy and Bayesian Methods, Santa Barbara 1993. Dordrecht: Kluwer.

Sarle, Warren S. (1995). Stopped Training and Other Remedies for Overfitting. To appear in Proceedings of the 27th Symposium on the Interface, 1996.

Warner, B. and Misra, M. (1996). Understanding Neural Networks as Statistical Tools, The American Statistician, November 1996, Vol. 50, No. 4, 284-293.

Table 1. Root mean square errors of various neural nets

# hidden nodes	# network weights	Naïve training training error	Naïve training holdout error	Stopped training training error	Stopped training holdout error
1	4	0.3860	0.3500	0.3860	0.3501
2	7	0.2504	0.1618	0.2504	0.1620
3	10	0.1334	0.05714	0.1686	0.08820
4	13	0.1289	0.05946	0.1324	0.04817
5	16	0.1289	0.06007	0.1479	0.06852
10	31	0.1253	0.06255	0.1347	0.04300
20	61	0.1196	0.08171	0.1671	0.08140

Standard deviation of added noise = 0.1391

Figure 1. The theoretical function for the regression problem

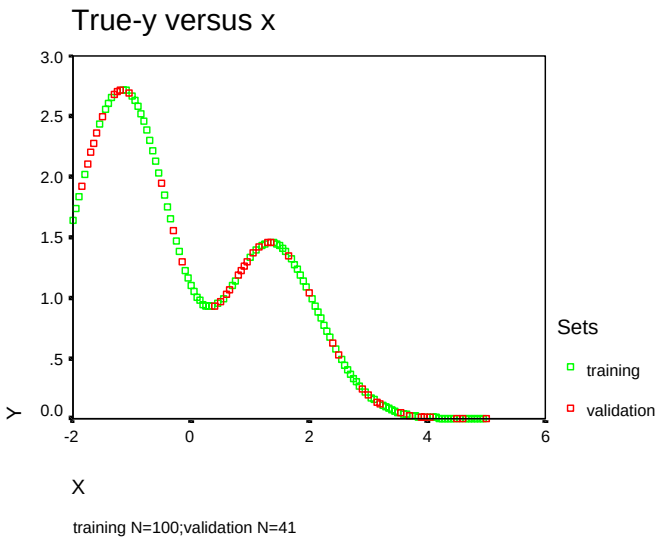


Figure 2. Function with added noise

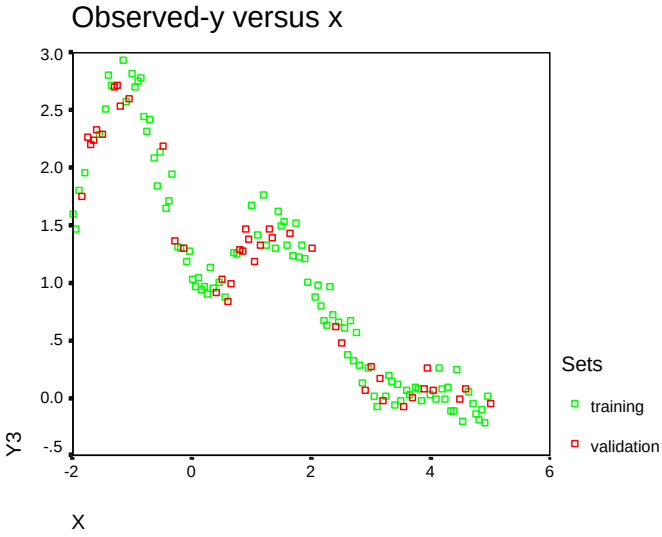


Figure 3. Holdout data (noiseless)

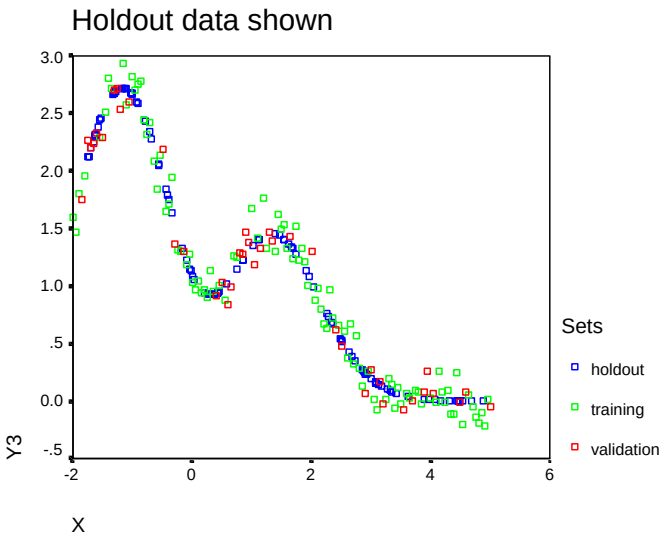


Figure 4. Simple regression shows bias

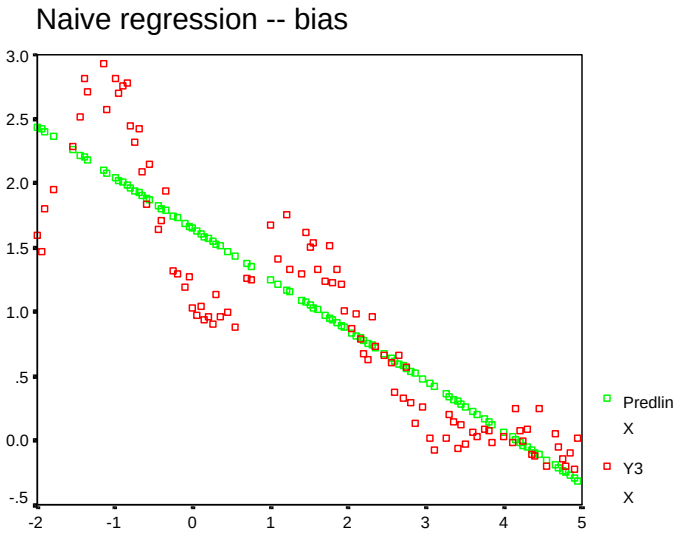


Figure 5. Quartic equation in X shows bias

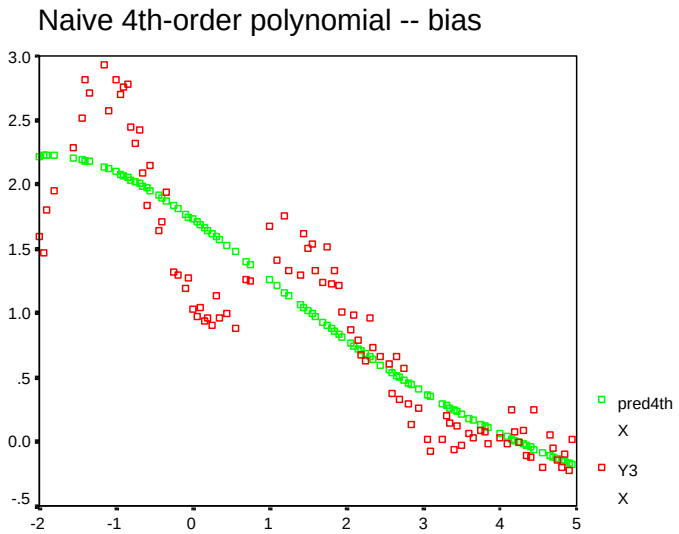


Figure 6. Train on training data, fit to training data

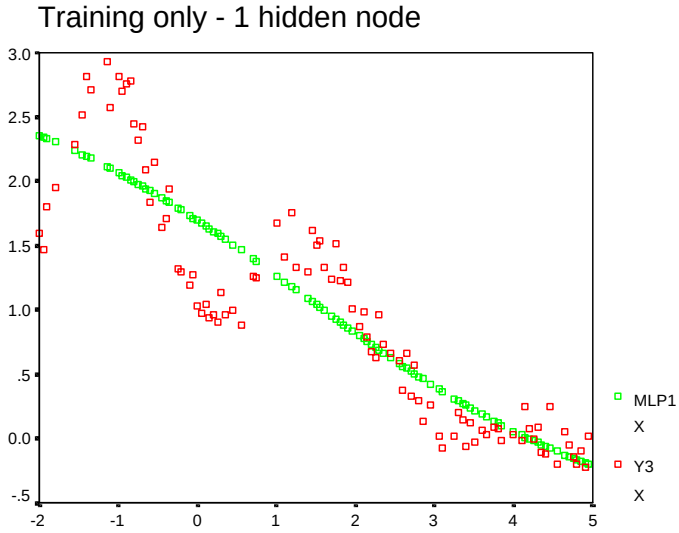


Figure 7. Train on training data, fit to training data

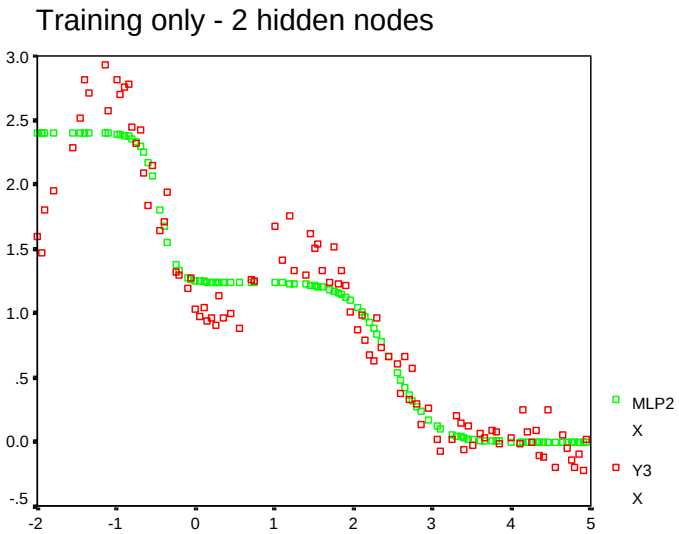


Figure 8. Train on training data, fit to training data

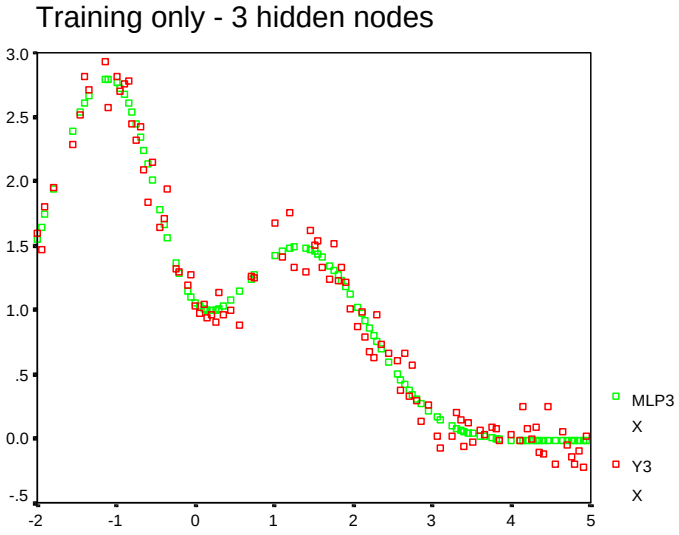


Figure 9. Train on training data, fit to training data

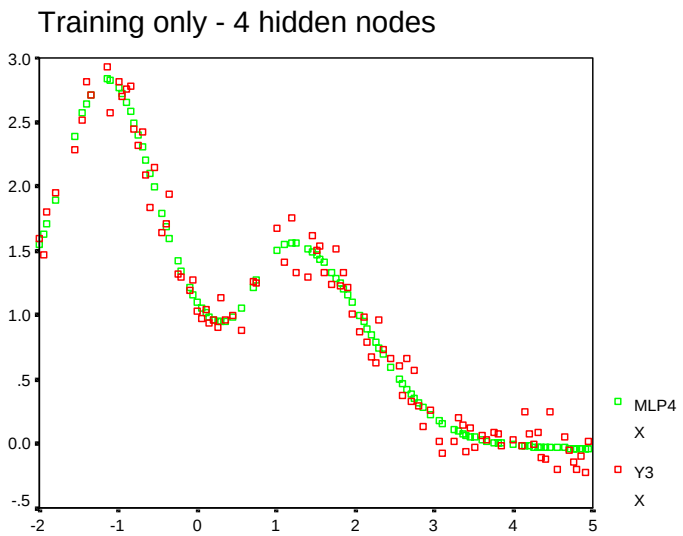


Figure 10. Train on training data, fit to training data

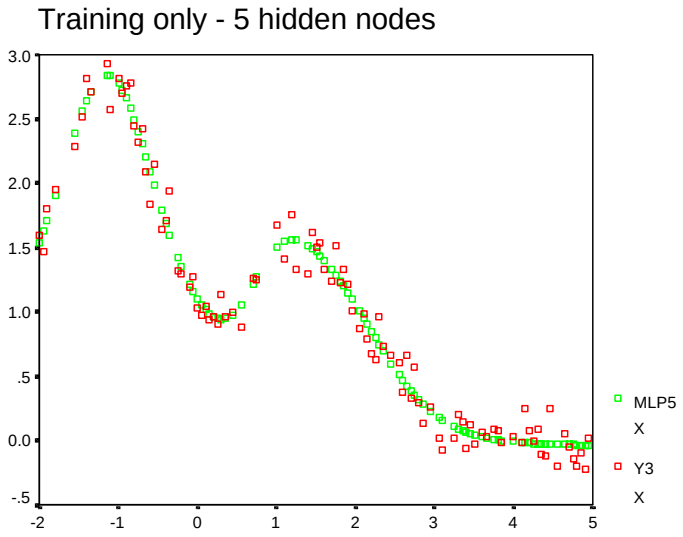


Figure 11. Train on training data, fit to training data

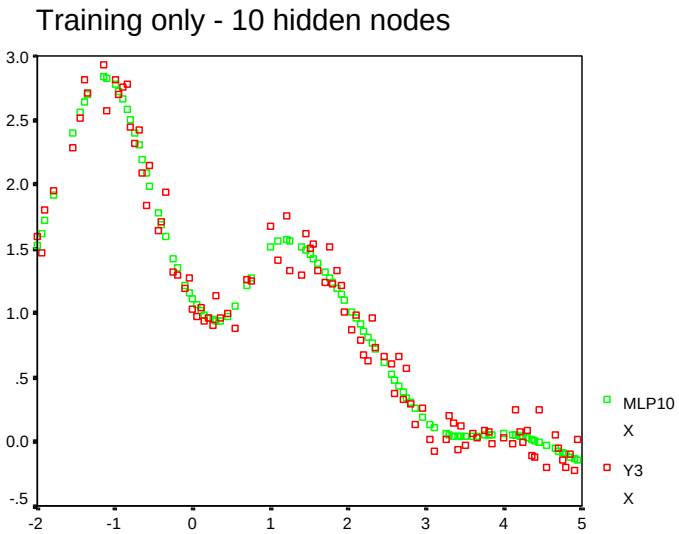


Figure 12. Train on training data, fit to training data

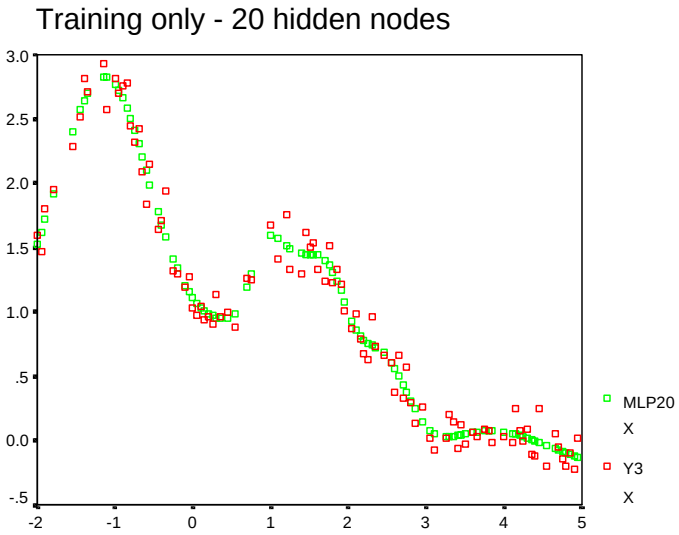


Figure 13. Train on training data, fit to training data

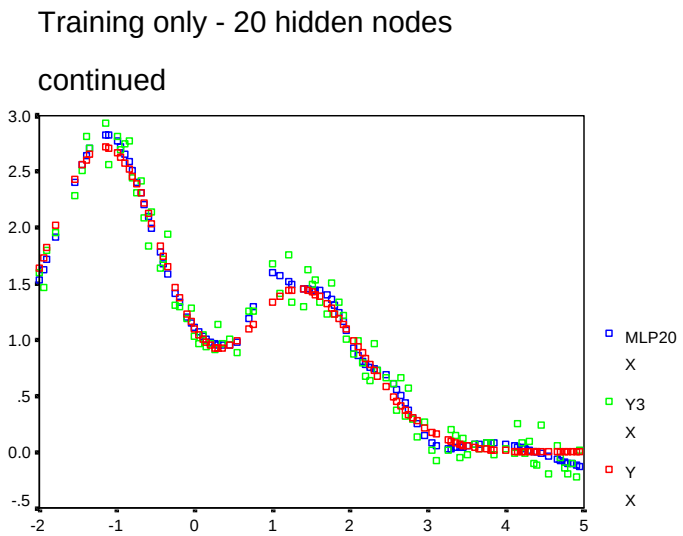


Figure 14. Train on training data, fit to holdout data

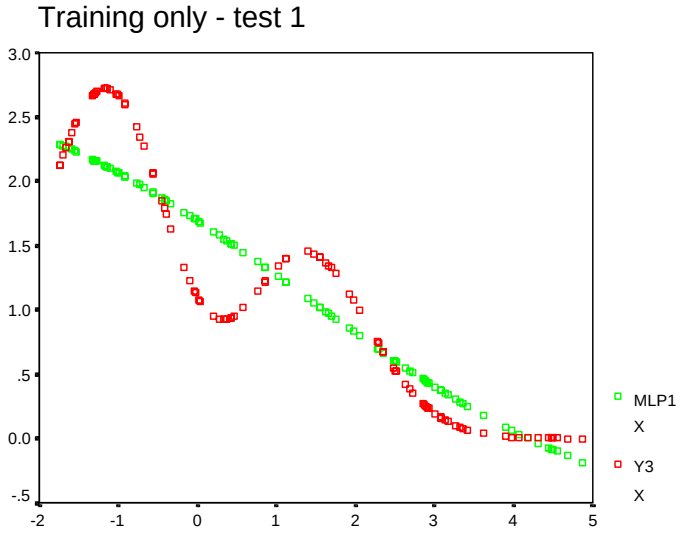


Figure 15. Train on training data, fit to holdout data

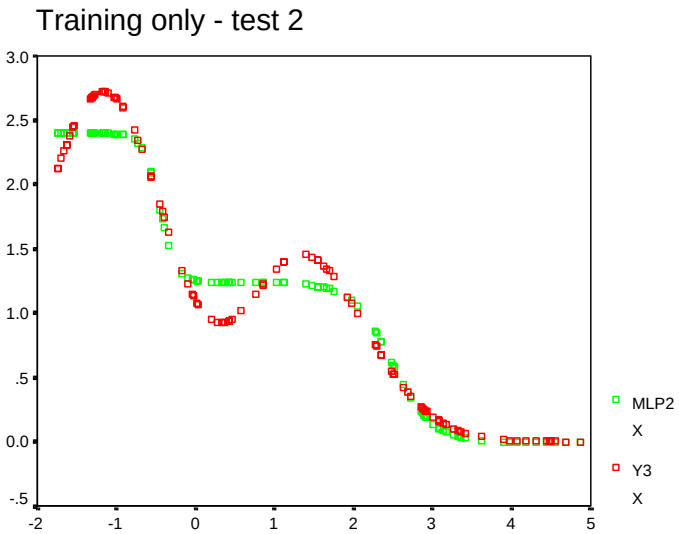


Figure 16. Train on training data, fit to holdout data

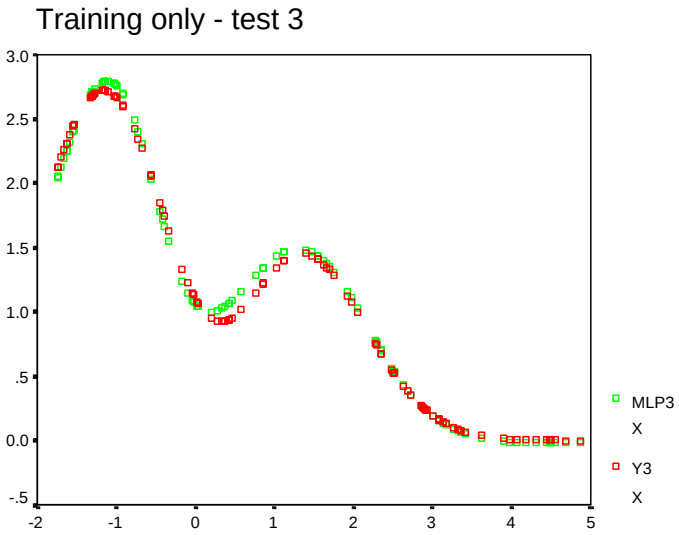


Figure 17. Train on training data, fit to holdout data

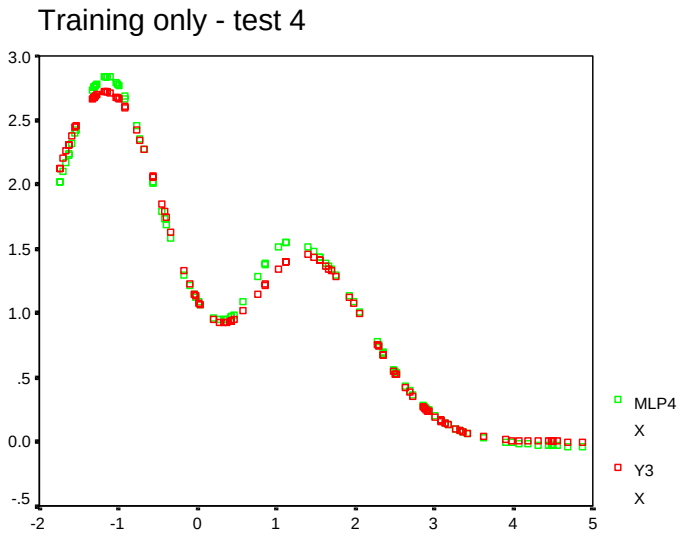


Figure 18. Train on training data, fit to holdout data

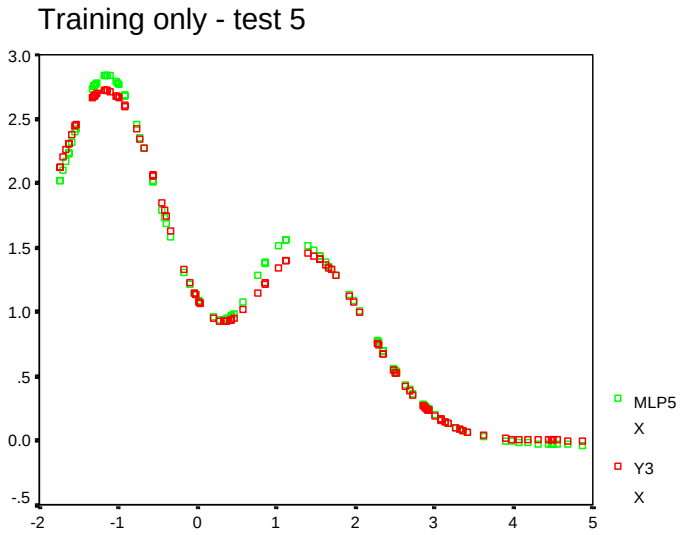


Figure 19. Train on training data, fit to holdout data

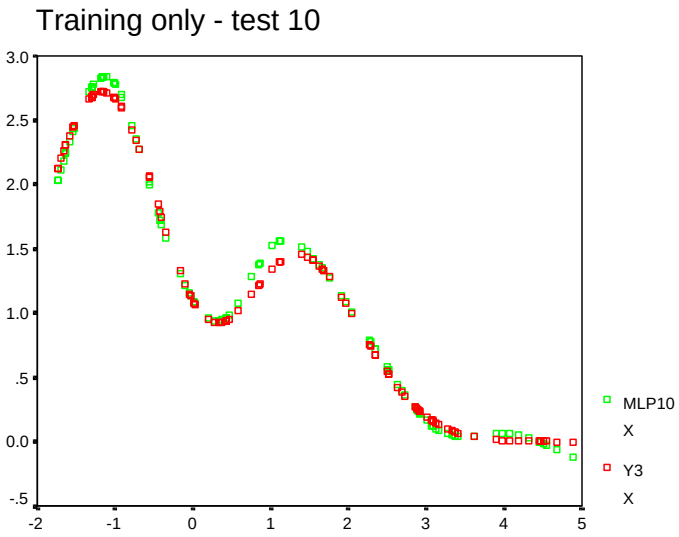


Figure 20. Train on training data, fit to holdout data

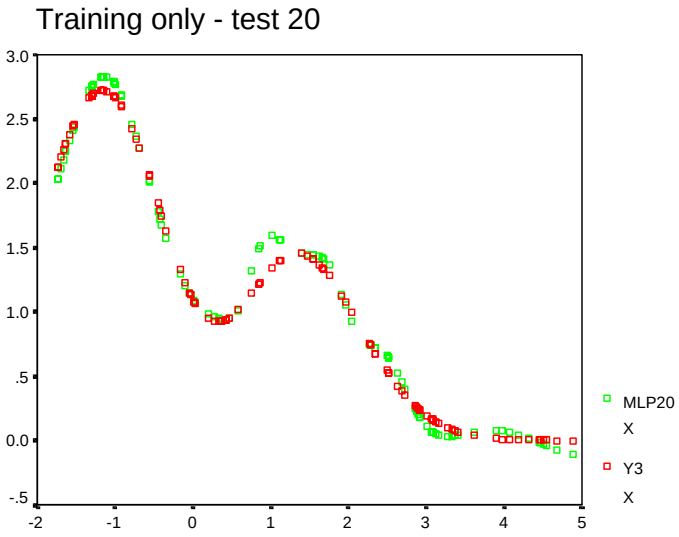


Figure 21. Train via stopped training, fit to training data

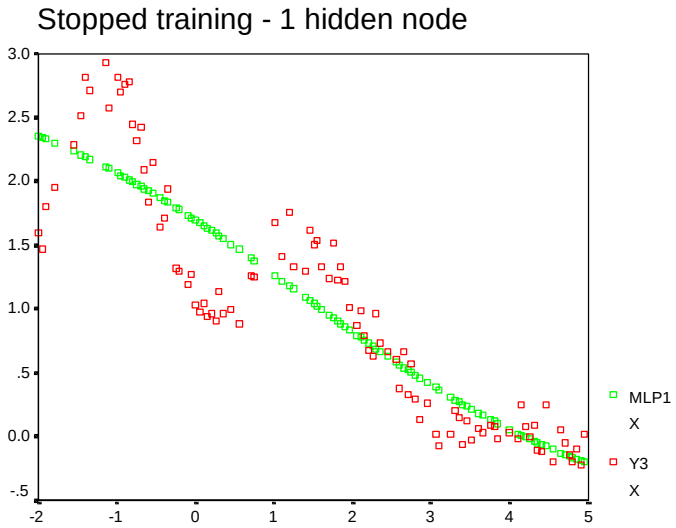


Figure 22. Train via stopped training, fit to training data

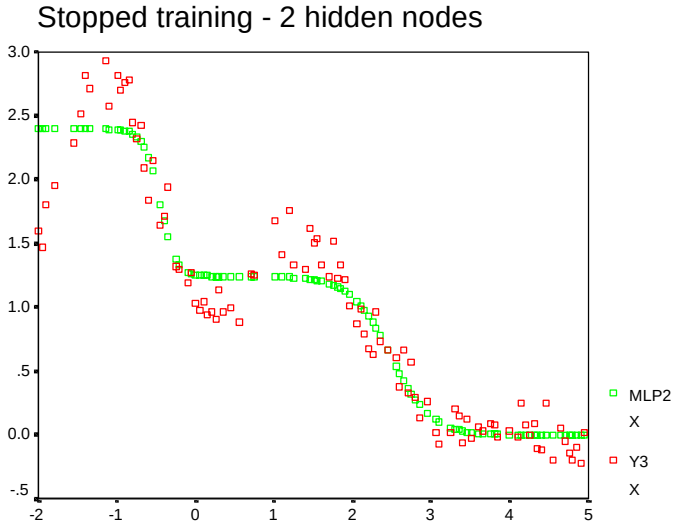


Figure 23. Train via stopped training, fit to training data

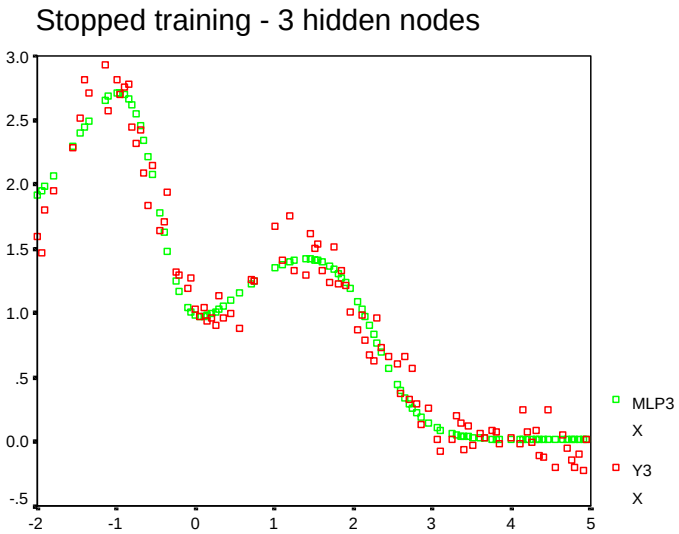


Figure 24. Train via stopped training, fit to training data

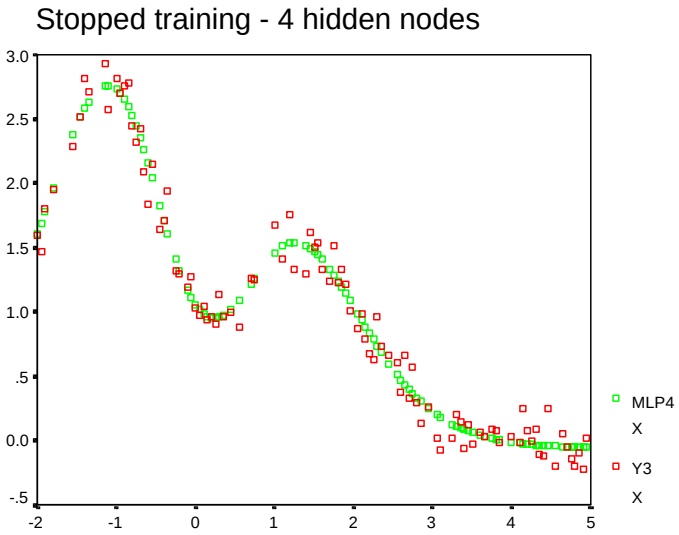


Figure 25. Train via stopped training, fit to training data

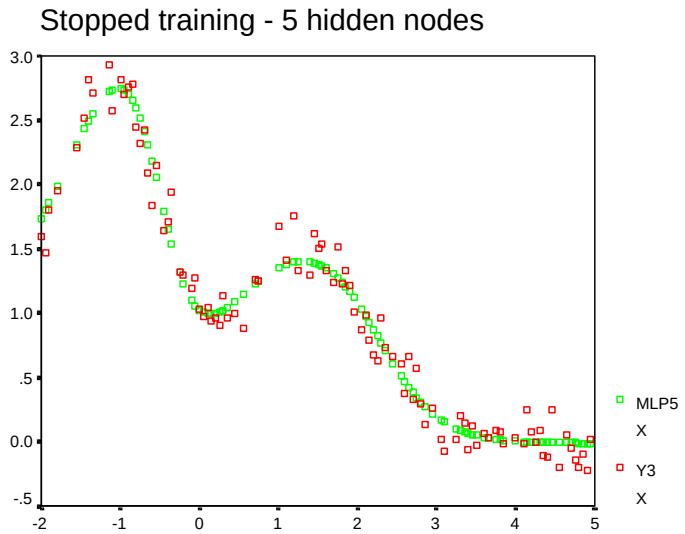


Figure 26. Train via stopped training, fit to training data

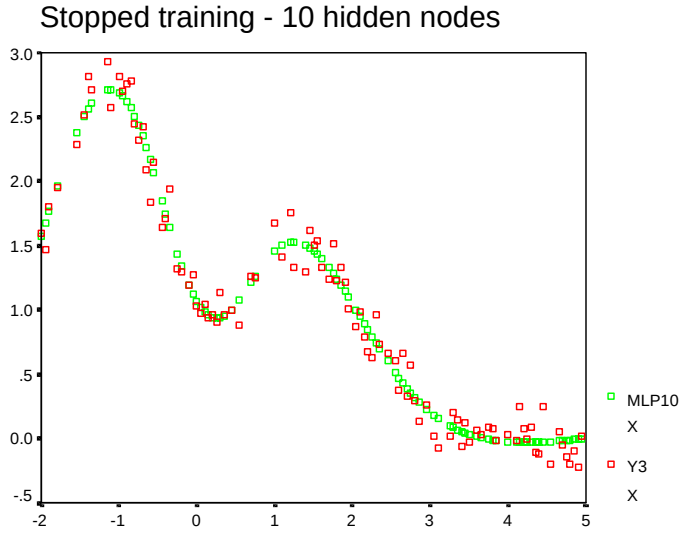


Figure 27. Train via stopped training, fit to training data

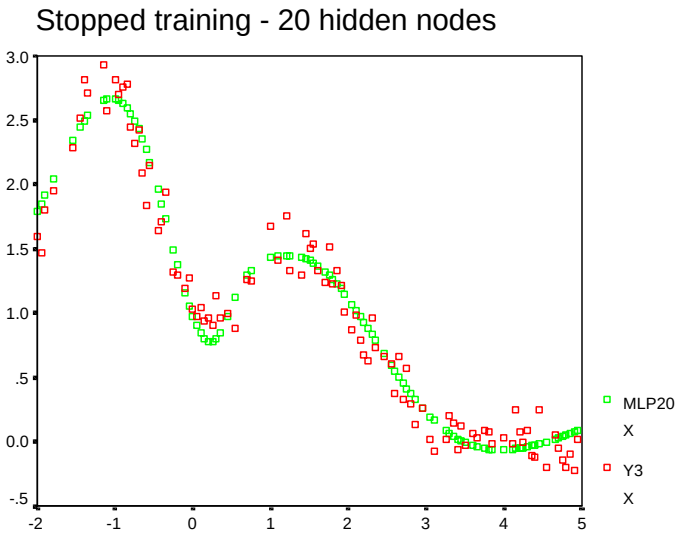


Figure 28. Train via stopped training, fit to training data

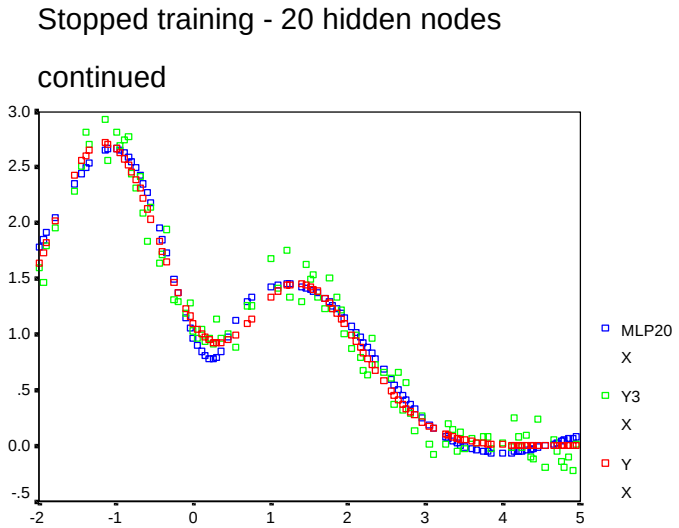


Figure 29. Train via stopped training, fit to holdout data

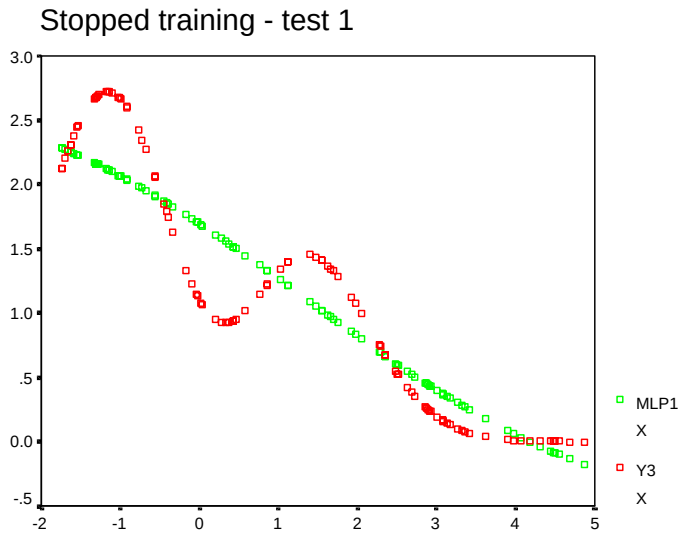


Figure 30. Train via stopped training, fit to holdout data

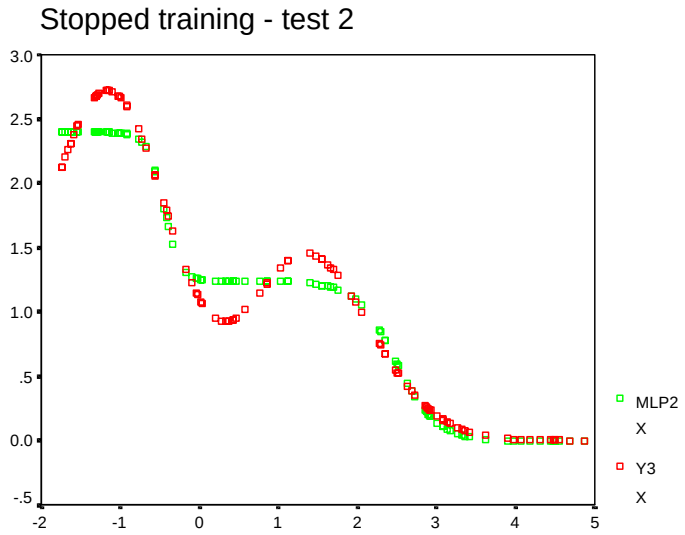


Figure 31. Train via stopped training, fit to holdout data

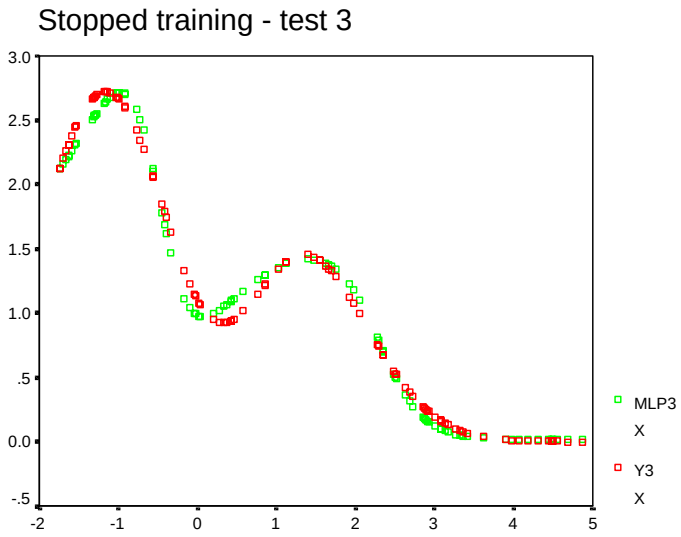


Figure 32. Train via stopped training, fit to holdout data

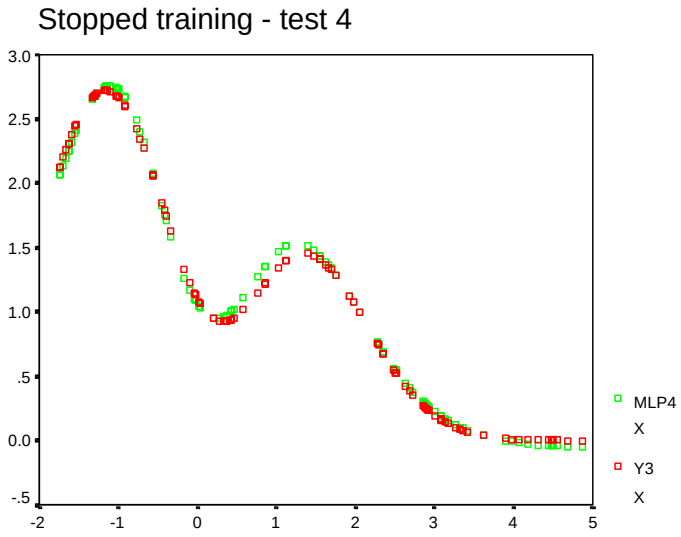


Figure 33. Train via stopped training, fit to holdout data

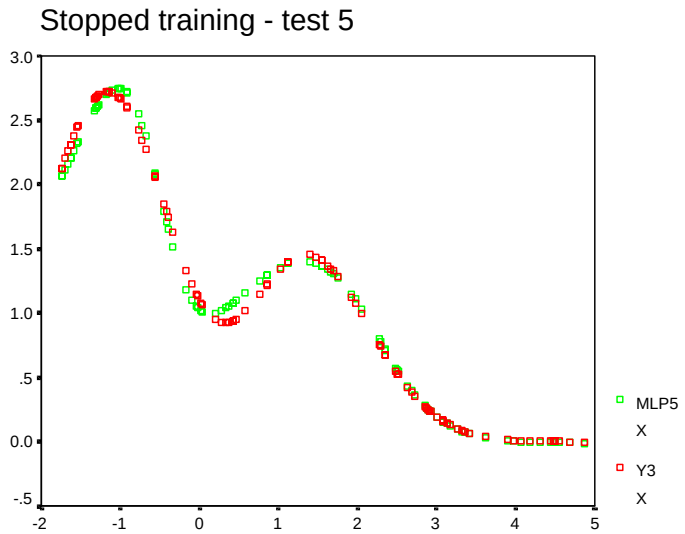


Figure 34. Train via stopped training, fit to holdout data

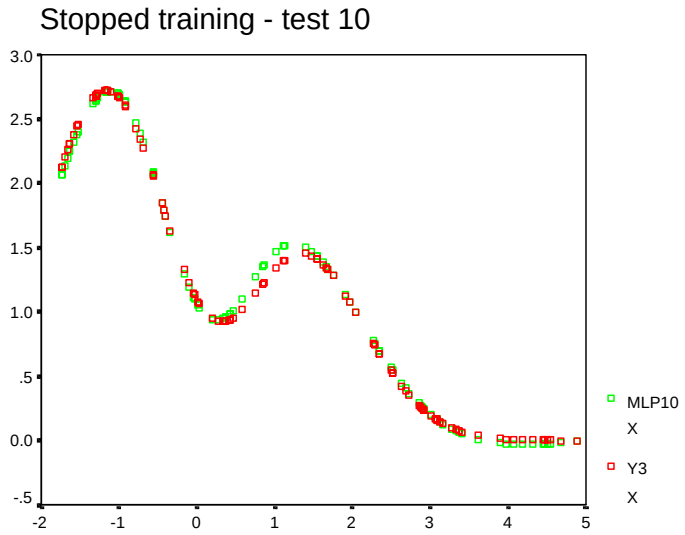


Figure 35. Train via stopped training, fit to holdout data

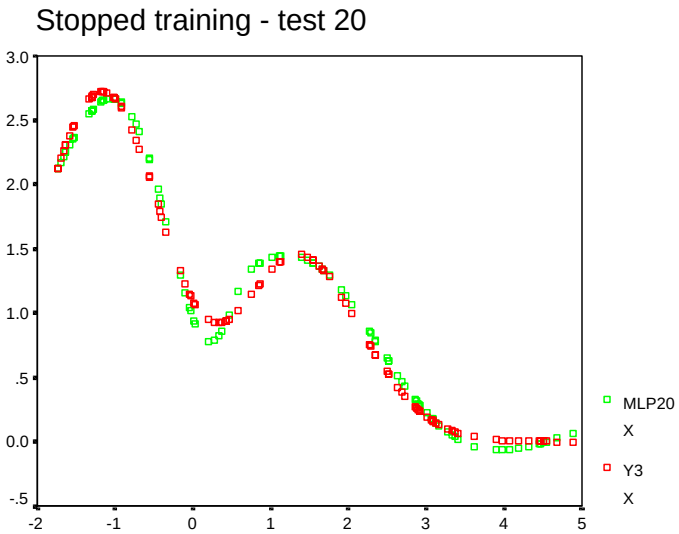


Figure 36. Two group problem, not linearly separable

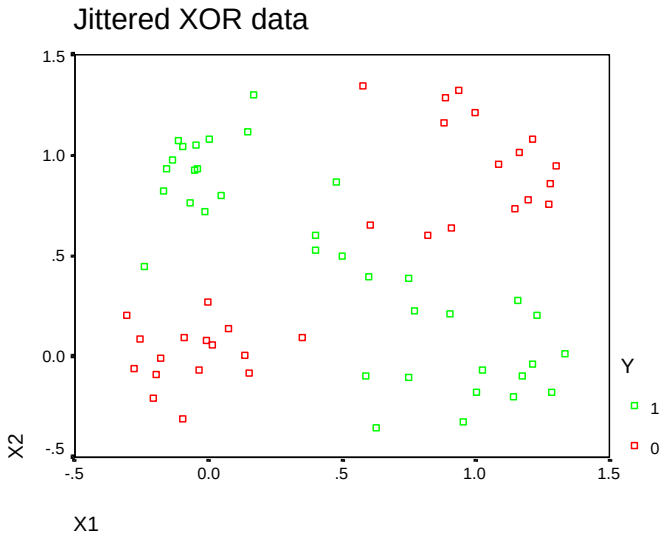


Figure 37. “Best” linear boundary

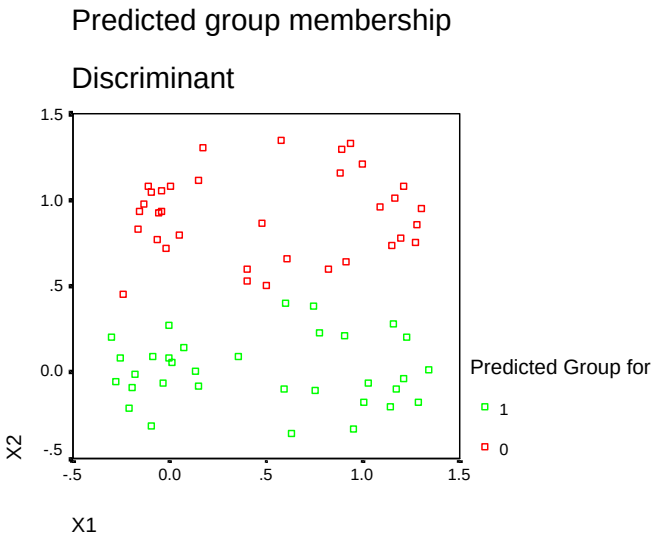


Figure 38. Nonlinear function fit to jittered XOR data

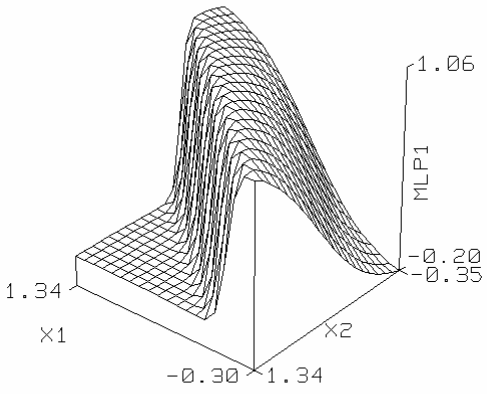


Figure 39. Two group problem, no linear boundary

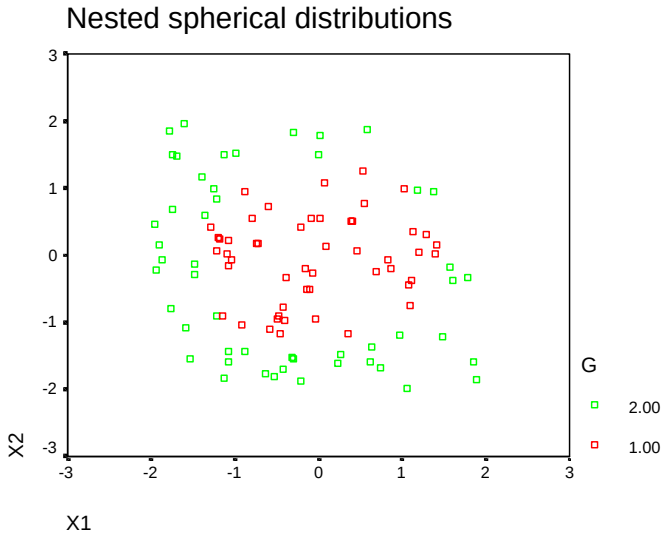


Figure 40. Chaotic time series

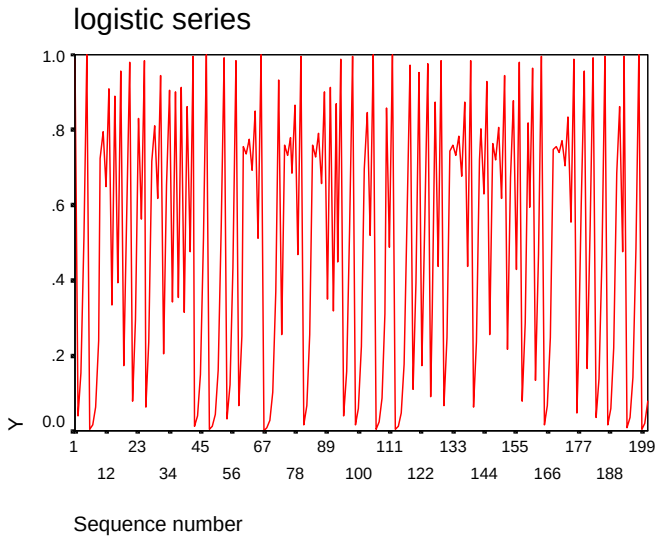


Figure 41. ACF plot for chaotic series

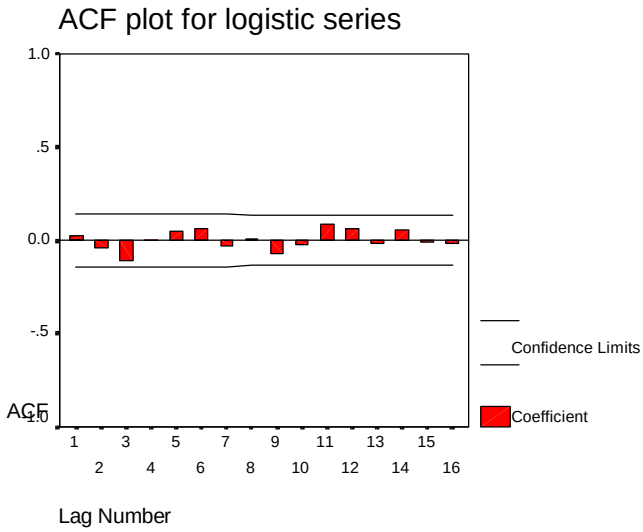


Figure 42. Phase plot for chaotic time series

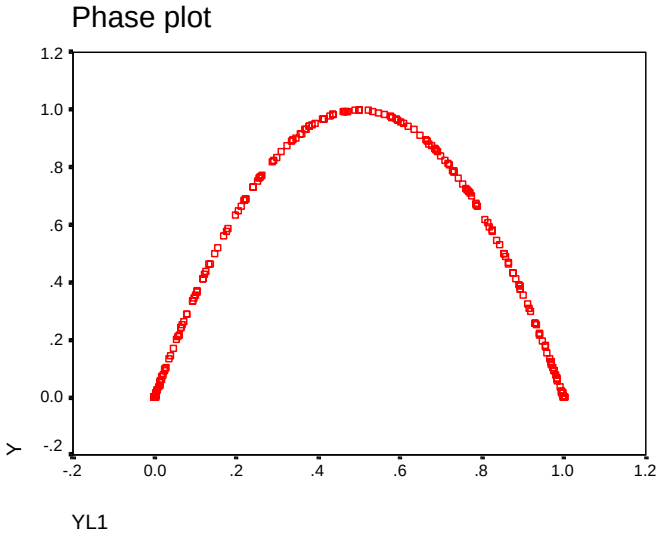


Figure 43. Neural net forecast into holdout sample of last 20 points