Sawtooth Software

TECHNICAL PAPER SERIES

The ACA/Web v6.0 Technical Paper

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

ACA System for Adaptive Conjoint Analysis

Copyright Sawtooth Software, Inc.
Sequim, Washington USA
+1 360 681-2300
September, 2007

Introduction

"Adaptive Conjoint Analysis" (ACA) is a component within Sawtooth Software's SSI Web platform. ACA may be used in web-based data collection, or using computers not connected to the web (CAPI mode). The term "adaptive" refers to the fact that the computer-administered interview is customized for each respondent; at each step, previous answers are used to decide which question to ask next, to obtain the most information about the respondent's preferences.

ACA has two essential capabilities. First, it lets the researcher design a computer-interactive interview and administer the interview to respondents. The interview can consider many attributes and levels, paying special attention to those the respondent considers most important. Questioning is done in an "intelligent" way; the respondent's utilities are continually re-estimated as the interview progresses, and each question is chosen to provide the most additional information, given what is already known about the respondent's values. Respondent utilities are available upon completion of the interview.

Second, ACA lets the researcher simulate respondent preferences for new or modified products. The ACA simulator can be used to explore "what if" scenarios, such as changes in pricing, product formulation, or marketing activities. The researcher describes a group of hypothetical products by specifying each product's level on each attribute. Respondent utilities are used to estimate strengths of preference or buying likelihoods for each product, and results are cumulated over respondents to provide shares of preference or average estimated buying likelihoods for each product.

ACA has been the focus of many articles over the last three decades. Studies bearing on the validity of ACA are reviewed in a paper available from Sawtooth Software (Johnson, 1991), and additional sources are described below.

Background and History

Conjoint analysis made its first appearance in marketing research in the early '70s, and has had increasing use since that time. Cattin and Wittink (1982) surveyed firms that reported conducting several hundred conjoint studies during the '70s. An update of their survey (Wittink and Cattin 1989) showed that use of conjoint analysis had continued to expand in the '80s. Based on a 2007 survey of Sawtooth Software customers, we estimate that our customers alone were responsible for nearly 10,000 conjoint studies over the previous year. For an overview of the early history of conjoint analysis, we recommend the review by Green and Srinivasan (1990). In characterizing developments of the 1980s, they conclude:

> "As we reflect on the activities that characterize research in conjoint analysis, two key trends appear to have been the development of (1) standard microcomputer packages and (2) modified approaches to conjoint analysis for obtaining stable part-worth estimates at the individual level for problems involving large numbers of attributes."

ACA's "modified approach" to conjoint analysis is also its most unusual aspect. Its adaptive interview

provides a way of studying large numbers of attributes and levels using efficiencies of computer-administered interviews.

The process that ultimately led to the development of ACA began with a series of early experiences with large-scale conjoint projects. One such study, conducted in 1977 for the Department of Defense, sought ways to make military careers more attractive to potential recruits. That study involved conjoint analysis of 12 attributes of military careers, with a total of 38 levels. Approximately 1,300 young men completed the conjoint questionnaire, answering while seated at terminals connected by phone lines to a central computer. Computer-administered data collection was selected because it was considered most likely to produce data of high quality. However, the computer time and the telephone charges required for computer interviewing before the age of the PC were very expensive on a per-interview basis. The need to minimize cost focused attention on the issue of how to get the most information from the briefest possible interview, and suggested a line of questioning which eventually evolved into the approach used in ACA.

With the advent of microcomputers, it became possible to obtain the benefits of computer-administered interviews at much lower cost. Dozens of conjoint projects were undertaken between 1978 and 1985 in which respondents were interviewed using Apple II computers. Those studies permitted a series of informal experiments to test interview procedures for providing stable estimates of respondent utilities for many attributes and levels.

The first version of ACA benefited from this program of informal optimization. Although refinements were offered in Versions 2 and 3, the basic ideas underlying the interview process remained unchanged. However, several authors offered suggestions for improvements in ACA (for example, see Green, Krieger, and Agarwal, 1991). Some of those suggestions were incorporated as options in ACA Version 4. For versions 5 and 6, further refinements were made based on more recent research.

In the 1990s, ACA was the most widely used conjoint software in the world (Vriens, Huber, and Wittink, Vriens 1997). Based on our own internal surveys of customers, we believe it held that position until about 2000, when Choice-Based Conjoint (CBC) eclipsed the use of ACA. ACA has unique capabilities and is still often used, especially for product design, employee research, and segmentation studies where the number of attributes exceeds what can reasonably be handled with CBC and traditional conjoint methods.

The Problem of Too Many Attributes

Users of conjoint analysis would probably agree that their most serious practical problem is that of dealing with large numbers of attributes. The client often has a detailed understanding of the product category, and previous work has usually identified many issues of interest. The respondent, on the other hand, usually has less interest and enthusiasm, and is rarely willing to submit to a long interview. Thus the researcher is often in conflict. Should there be a long interview, risking unhappy respondents and data of questionable quality? Or should the researcher insist on narrower focus in the interview, providing the client with less breadth than desired?

Statisticians think of this as a problem in “degrees of freedom.” Suppose a dozen attributes are thought to be important, with an average of four levels each. (A dozen attributes is probably fewer than are found in the average commercial study today.) We may set the utility for the least-liked level of each attribute equal to an arbitrary value such as zero, and we may scale all levels so the greatest utility, across all attributes, is equal to any arbitrary value, such as 100. Thus, if there are n attributes with an average of k levels, we must estimate $n (k-1) -1$ utilities. With twelve attributes having an average of four levels each, there would be 35 values to estimate.

The questions asked of the respondent are likely to be difficult, perhaps requiring much thought. At the very minimum, the respondent would have to provide 35 answers. If there is any random component to the responses we would need more observations. As a rule of thumb, one might require three times as many observations as parameters being estimated, which would require our unfortunate respondent to answer 105 questions.

Accordingly, practical applications of conjoint analysis have often required one or another of these expedients:

1. Limit the number of attributes to what the client may regard as unrealistically few.

2. Limit the number of levels per attribute. This can be done for quantitative attributes where high and low levels may be studied and the researcher may interpolate to estimate mid-level values. But this is not useful with "categorical" attributes like brand or color.

3. Estimate values for groups rather than individuals. If the analysis is done by aggregating data for many individuals, each person may have to answer only a few questions.

4. Use "hybrid" models that estimate utilities for individuals, but augment each individual's data with information based on group averages.

5. Customize the interview so each respondent is asked in detail only about those attributes of greatest relevance.

ACA solves this problem in the fifth way. The interview consists of an initial screening section that determines the relative desirability of each attribute level and the relative importance of each attribute. Then, using options chosen by the researcher, the interview can focus on just those attributes the respondent regards as most important, and only those attribute levels regarded as most relevant.

The ACA Questionnaire

ACA interviews can study up to 30 attributes, and each attribute can include up to 15 levels. Most ACA studies in practice have between 8 to 15 attributes, each described on no more than about 5 levels. Attribute levels are usually described using short phrases, but attribute levels can also be represented as graphics, sounds or videos.

Suppose an ACA study has the following attribute list:

Brand:
Dell
HP
Lenovo
Gateway

Processor Speed:
2.66 GHz (Dual Core Processor)
2.66 GHz (Quad Core Processor)
3.50 GHz (Quad Core Processor)

Memory (RAM):
1 GB RAM
2 GB RAM
4 GB RAM

Hard Drive:
100 GB Hard Drive
160 GB Hard Drive
250 GB Hard Drive
500 GB Hard Drive

Monitor:
17" Monitor
19" Monitor
21" Monitor

Keyboard:
Standard Keyboard
Ergonomic Keyboard

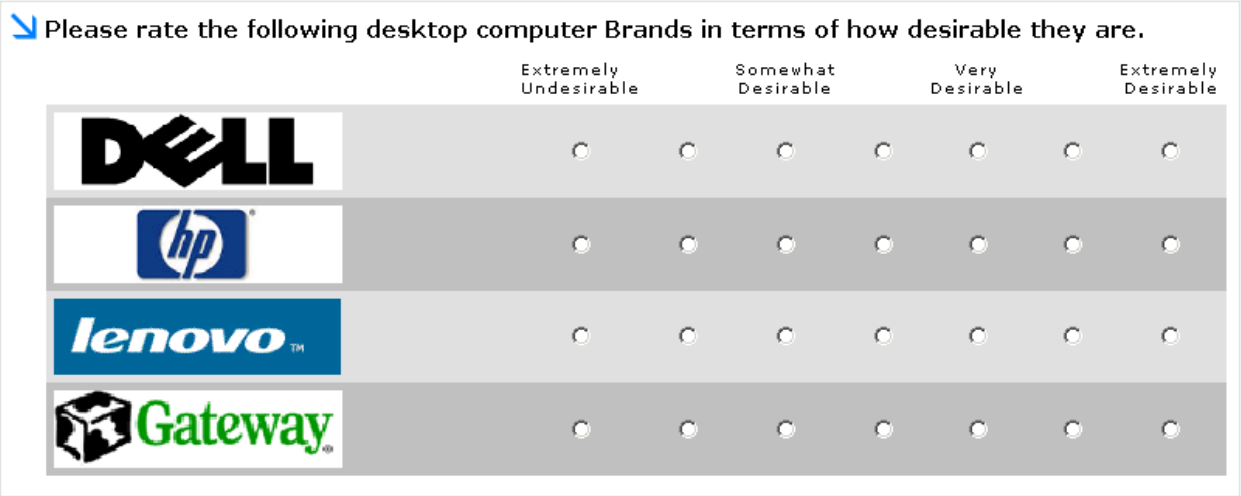
Productivity Software:
Microsoft Works
Microsoft Office Small Business (Basic + PowerPoint, Publisher)
Microsoft Office (Word, Excel, Outlook)
Microsoft Office Professional (Small Bus + Access database)

The ACA interview has several sections, each with a specific purpose. Below, we provide examples of the questions in each section, together with brief explanations of their purposes.

1) Preference for Levels

(The "ACA Rating" question type.)

In this section, the respondent rates the levels in terms of relative preference. This question is usually omitted for attributes (such as price or quality) for which the respondent's preferences should be obvious. (When you input the attributes and levels, you can specify that the order of preference for levels is "best to worst" or "worst to best" and the ratings question is skipped for such attributes.) The screen may look like the following:



The ratings question can be defined from 2 to 9 points. We suggest using at least 7 scale points. In any case, it is probably not wise to use fewer scale points than the number of levels in any one attribute for which the Rating question is asked. The respondent is required to check one radio button per attribute level.

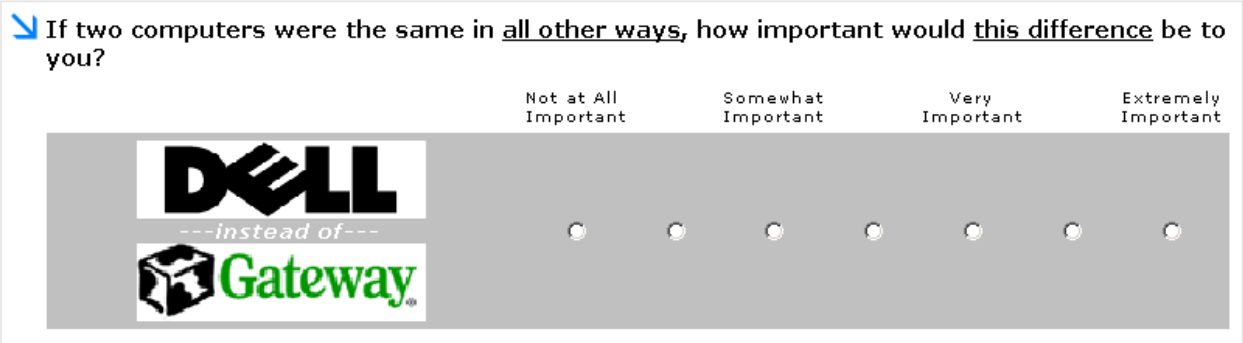
2) Attribute Importance

(The "ACA Importance" question type.)

Having learned preferences for the levels within each attribute, we next determine the relative importance of each attribute to this respondent. This information is useful in two ways. First, it may allow us to eliminate some attributes from further evaluation if the interview would otherwise be too long. Second, it provides information upon which to base initial estimates of this respondent's utilities.

As a matter of fundamental philosophy, we do not ask about attribute importance with questions such as "How important is price?" The importance of an attribute is clearly dependent on magnitudes of differences among the levels being considered. For example, if all airline tickets from City A to City B were to cost between $100 and $101, then price couldn't be important in deciding which airline to select. However, if cost varied from $10 to $1000, then price would probably be seen as very important.

Our questioning is based on differences between those levels the respondent would like best and least, as illustrated (assuming Dell is rated best and Gateway rated worst):



Recent research has suggested that the importance question in ACA can be problematic. If respondents

have a tendency to say that all attributes are important, this can bias the part-worth utilities toward too little discrimination. Also, if respondents aren't made familiar with the breadth of attributes that will be covered, they may have a difficult time using the rating scale well for attributes considered one at a time. Researchers have also found importance questions to have lower reliability than rating the preference for levels within attributes. (See King *et al.* 2004.)

Starting with version 6, users have the option (if using ACA/HB for part-worth estimation) of omitting the importance questions. Importance information for designing the subsequent tradeoffs in the questionnaire may be inferred from the average part-worth utilities of previous respondents. We suggest ample sample size if omitting the importance question, as some information at the individual level is lost (King et al. 2004). We also recommend adding more paired-comparison trade-off questions in place of the omitted importance questions. As an advanced option, rather than omitting the importance questions, users may customize them using any SSI Web question type that leads to a numeric input.

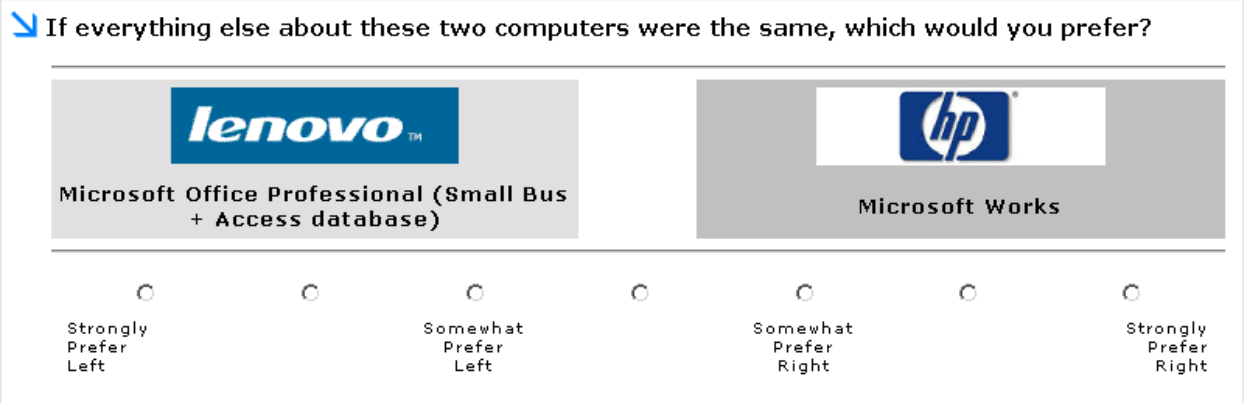
At this point we have learned which attributes are most important for this respondent and which levels are preferred. From now on, the interview is focused on those most important attributes and combinations of the levels that imply the most difficult trade-offs.

3) Paired-Comparison Trade-Off Questions

(The "ACA Pairs" question.)

Next, a series of customized paired-comparison trade-off questions is presented. Up to this point in the interview, we have collected "prior" information; no conjoint analysis has been involved. The Pairs section elicits the conjoint tradeoffs. In each case the respondent is shown two product concepts. The respondent is asked which is preferred, and also to indicate strength of preference.

The example below presents concepts differing on only two attributes. Although concepts may be specified on up to five attributes, simple concepts like these present an easy task for the respondent, and are a useful way to begin this section of the interview.



The number of attributes appearing in each concept is specified by the author, and can be varied during the interview. Concepts described on more attributes have the advantage of seeming more realistic. It is also true that statistical estimation is somewhat more efficient with more attributes.

However, with more attributes, the respondent must process more information and the task is more difficult. Experimental evidence indicates that a tradeoff occurs: as the number of attributes in the concepts is increased, respondents are more likely to become confused. It appears best to start with only

two attributes. Most respondents can handle three attributes after they've become familiar with the task. Preliminary evidence suggests that beyond three attributes, gains in efficiency are usually offset by respondent confusion due to task difficulty.

The computer starts with a crude set of estimates for the respondent's utilities, and updates them following each pairs question. The crude estimates are constructed from the respondent's preference ranking or rating for levels, and ratings of importance of attributes. Each pairs question is chosen by the computer to provide the most incremental information, taking into account what is already known about this respondent's utilities. The interview continues in this manner until a termination criterion (specified by the author) is satisfied.

Every time the respondent completes a pairs question, the estimate of the respondent's utilities is updated. Updating the utilities improves the quality of subsequent pairs questions.

4) Calibrating Concepts (Optional Section)

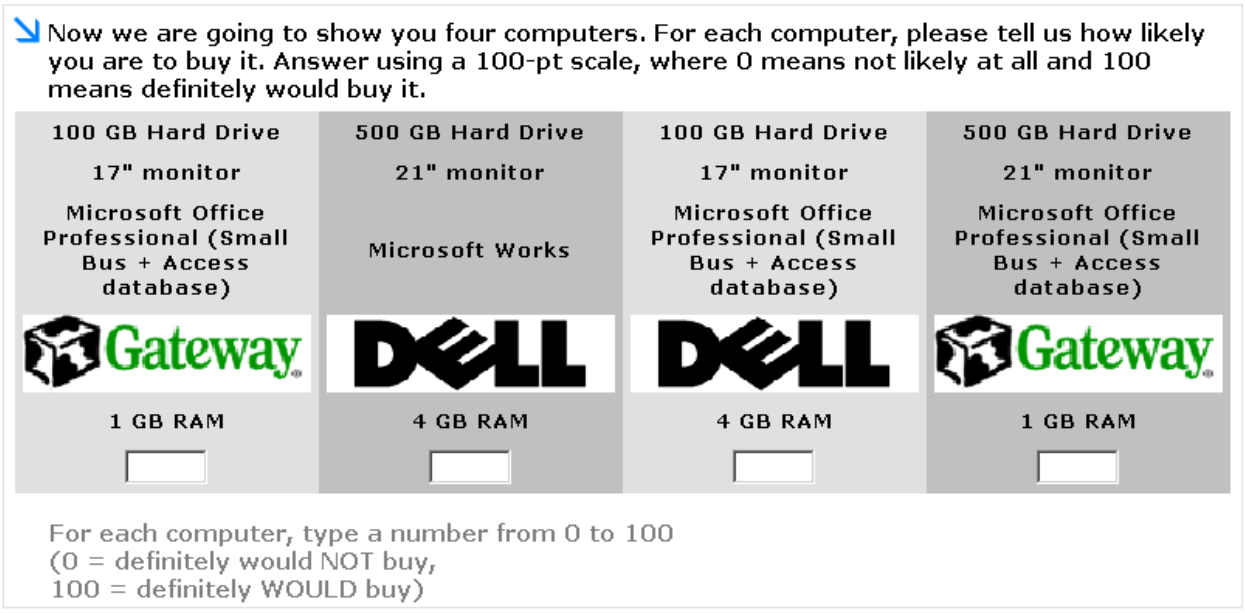
(The "ACA Calibration" question type.)

Finally, the computer composes a series of "calibrating concepts" using those attributes determined to be most important. These concepts are chosen to occupy the entire range from very unattractive to very attractive for the respondent. The respondent is asked a "likelihood of buying" question about each. The likelihood can be expressed using a slider scale, or by typing a numeric value into a box.

We first present the concept we expect the respondent to like least among all possible concepts, and the second is the one we expect to be liked best. Those two concepts establish a frame of reference. The remaining concepts are selected to have intermediate levels of attractiveness.

This information can be used to calibrate the utilities obtained in the earlier part of the interview for use in Purchase Likelihood simulations during analysis. Conjoint utilities are normally determined only to within an arbitrary linear transformation; one can add any constant to all the values for any attribute and multiply all utilities by any positive constant. The purpose of this section is to scale the utilities non-arbitrarily, so that sums of utilities for these concepts are approximately equal to logit transforms of the respondent's likelihood percentages.

One possible screen format, showing multiple products per screen is as follows:



You can also show each product on a separate screen.

If you plan only to conduct share of preference simulations, and particularly if you are using the ACA/HB system for hierarchical Bayes estimation, you may consider not including calibration concepts in your ACA survey.

Estimating Respondent Utilities

Initial utility estimates are based on the respondent's desirability ratings for attribute levels together with ratings of attribute importance. Utility estimates are updated during the interview to incorporate the answer to each paired-comparison question. As the respondent progresses through the paired comparison section, the initial estimates become less influential.

In earlier versions of ACA, the final utilities were true least squares estimates, with the same weight being applied to the prior estimates and the information from the paired-comparison questions. Starting in Version 4 those two components could be given unequal weights chosen to best fit responses to the calibration concepts, although the user could choose to estimate utilities in the same way as earlier ACA versions. Version 5 and later incorporates elements of both earlier approaches, estimating utilities separately for the priors and pairs and combining the two sets of utilities based on the amount of information contributed by each section. (Details of utility estimation are provided in Appendix A.)

It is worth mentioning that a superior method of utility estimation is available through hierarchical Bayes estimation. For details regarding HB estimation for ACA, please see the ACA/HB Technical Paper.

Choosing the Next Paired-Comparison Question

After a question has been answered and the utility estimates have been updated to incorporate that response, a decision must be made whether to ask another question and, if so, which one. Both of those issues are handled heuristically.

The general recommendation for the total number of questions to be asked is three times the number of parameters being estimated (although the author may specify that fewer questions are to be asked). We

count as “questions” those in the level ratings, importance, and paired-comparison sections.

In deciding which question to ask next, one would clearly want to ask questions for which there remains most uncertainty about the respondent's answers. In that spirit, one might estimate the expected variance of the response to each possible question, and then ask the question with the largest expected variance.

Unfortunately, when using least squares, although the relative sizes of those variances do depend on the particular questions previously asked, they are independent of the respondent's previous answers. Therefore the strategy of minimizing variances is not one that can benefit from the respondent's previous answers.

However, useful information is available. If one were asking binary choice questions rather than paired comparisons with a graded scale, one would want to select questions in which the respondent had nearly equal probability of choosing each alternative. Similarly, in this context it makes sense to ask questions for which we expect a response in the middle of the scale, which means that we should present concepts with nearly equal utilities. Such an approach has these benefits:

- It gives the respondent the impression that the system is paying attention to his or her answers, and it seems to be asking increasingly insightful questions.

- It keeps the respondent operating within the defined range of a response scale rather than at its ends.

- It provides data on “tight” inequalities if estimation is later to be done by nonmetric methods.

Accordingly, ACA presents pairs of concepts to the respondent that are as nearly equal as possible in estimated utility. At the same time, constraints are imposed to ensure that the overall design is nearly orthogonal. Within concepts, each pair of attributes is presented with equal frequency, and within each attribute, each pair of levels is presented with equal frequency. In addition, if the paired-comparison questions show only two attributes at a time, further steps are taken to insure that the overall design is “connected.”

Simulating Respondent Preferences

So far we have been concerned with data collection rather than analysis. ACA includes a powerful Market Simulator. The Market Simulator lets the researcher model a hypothetical “market” by specifying each product's level on each attribute. The file of respondent utility values is read, and a computation is made of each respondent's relative utility for each hypothetical product. There are five options for the way the utilities are used:

1. *First Choice*: Each respondent is allocated to the product having highest overall utility.

2. *Share of Preference*: Each respondent's “share of preference” is estimated for each product. During the calibration section of the interview, the utilities are scaled so that their antilogs may be treated as relative probabilities. The simulator sums the utilities for each product and then takes antilogs to obtain relative probabilities. Those are then percentaged for each respondent to obtain shares of preference. Shares of preference are averaged for all respondents, and those averages are used to summarize preferences for the respondents being analyzed.

3. *Share of Preference with Correction for Product Similarity*: The “share of preference” model

is widely used, perhaps because of its simplicity and because many regard it as an intuitively reasonable model of reality. However, it has a widely recognized disadvantage. Unlike the First Choice model, if an identical product is entered twice, the share of preference model may give it as much as twice the original share of preference. This problem is sometimes referred to as the "red bus/blue bus" or "IIA" (independence from irrelevant alternatives) problem. Our third option was any earlier attempt to overcome this problem. It examines the similarity of each pair of products and deflates shares of preference for products in proportion to their similarities to others. This ensures that the share of two identical but otherwise unique products together will equal what either product alone would get. This technique is available in the software for historical purposes, as a newer technique seems to generally perform better. We generally recommend using the newer fifth option for correcting for product similarity, called Randomized First Choice.

4. *Likelihood of Purchase*: The first three options all assume a set of competitive products, and are concerned with estimating the share a product might receive in that competitive context. However, sometimes no competitive products are available. For example, the researcher might be concerned with a new product category, and might want to estimate absolute level of interest in the category rather than share of preference within the category. Our fourth option does this. The utilities are scaled so that an inverse logit transform provides estimates of purchase likelihood, as expressed by the respondent in the calibration section of the questionnaire. The simulator estimates how each respondent might have answered if presented with any concept in the calibrating section of the interview.

5. *Randomized First Choice*: The Randomized First Choice (RFC) method combines many of the desirable elements of the First Choice and Share of Preference models. As the name implies, the method is based on the First Choice rule, and significantly reduces IIA difficulties. Rather than use the utilities as point estimates of preference, RFC recognizes that there is some degree of error or variance around these points. The RFC model adds unique random variance to each part-worth (and/or product utility) and computes shares of preference in the same manner as the First Choice method. Each respondent is sampled many times to stabilize the share estimates. The RFC model results in a correction for product similarity due to correlated sums of variance among products defined on many of the same attributes.

The ACA Market Simulator provides these capabilities:

- The researcher provides a "base case" containing initial product specifications.
- Each product's base share of preference is first evaluated, and then product specifications can be modified and simulations can be done to estimate the impact of specific modifications.
- The effects of all possible one-attribute-at-a-time modifications can be evaluated automatically.
- The simulator can interpolate between attribute levels.
- Subsets of respondents can be analyzed, and average utilities as well as shares of preference are computed for each subset. Standard errors are also reported.
- Respondents can be weighted using demographic or other data.
- The results can be directly saved to Excel™ files, or cut-and-pasted into most any Windows-

based program.

Evidence of Usefulness and Validity

The general impression of those who witness ACA interviews is that computer-interactive methods capture and hold the respondent's attention more successfully than paper-and-pencil methods.

One of the earliest studies relevant to the validity of ACA was done with an "ACA-like" interview several years before ACA became available commercially. MacBride and Johnson (1979) reported a study in England and Germany in which equivalent samples of industrial buyers of office equipment received "ACA-like" and paper and-pencil interviews. Each respondent was interviewed about 10 attributes which the respondent had previously indicated as being important. Those respondents receiving paper-and-pencil interviews filled out tradeoff matrices. The "ACA-like" interview took slightly less time. When asked to estimate the length of the interview, those receiving the "ACA-like" version tended to underestimate the time actually spent. Each respondent was also asked to rate the interview experience for both "interest" and "ease." The "ACA-like" interview was preferred in both cases. Finally, utilities were used to predict each respondent's choice among holdout concepts. Again, the "ACA-like" interviews were more successful.

Several more recent studies have examined the validity of ACA, many of which were reviewed by Johnson (1991). Those studies produced mixed findings; some indicated that ACA performed quite well and others suggested that it had shortcomings. In reviewing those studies, Johnson proposed three guidelines:

1. Respondents should be like real customers, rather than just members of a convenience sample. (Several of the reviewed studies used students as respondents, and one used professional practitioners of conjoint analysis.)

2. If prediction of choice among holdout concepts is used to assess validity, then it is essential to have a measure of the consistency of those choices. If respondents are not consistent in their choices, then it is not possible for any method to predict those choices. Without a measure of reliability of holdout choice, one cannot assess the level of prediction produced by conjoint analysis.

3. If the same respondents perform multiple conjoint tasks (such as ACA and full profile) it is critical that task order be balanced. Respondents learn from conjoint interviews, and their answers are likely to be better in the second task than in the first.

We shall briefly describe studies conducted between 1986 and 1991 that did not violate these three principles:

Finkbeiner and Platz (1986) compared ACA with the full profile method of data collection in a six-attribute study of checking accounts. Their respondents were recruited from office building and mall locations, and were qualified by having checking accounts and not working for banks. They found:

- The results produced by the two methods were similar both in average utilities and accuracy of prediction.

- ACA took longer than full profile with six attributes, but with large numbers of attributes ACA would be expected to have a timing advantage.

- ACA permitted much faster data analysis; results were available in 1 to 2 days for ACA, but 1 to 2 weeks for the full profile method.

Finkbeiner and Platz suggested that ACA be used rather than full profile if the number of attributes and levels would require more than about 32 full-profile judgments.

Tumbusch (1991), shortly after retiring from Procter & Gamble, described a large validation study. In 1984, ACA was used to study four categories of frequently purchased products. Respondents were screened and recruited by phone to central location facilities in several cities. Sample sizes for the different product categories ranged from 319 to 706.

A year later, several concepts from each product category were tested with questionnaires mailed to other individuals. Each respondent rated a single concept on a purchase likelihood scale. At least 300 respondents rated each concept.

The average ACA utility value for each concept was then correlated with its "percent top box" purchase likelihood rating made by other respondents a year later. The resulting correlations, one per product category, ranged from .72 to .81.

Since different respondents were involved in the conjoint and validation tasks, some of the differences between ACA utilities and purchase likelihood estimates were due to sampling error. Also, many months intervened between the two measurements, and any changes that might have occurred in respondents' values would have decreased the correlations. For these reasons, Tumbusch points out that this study constituted a severe test of ACA.

McLauchlan (1991) responded to a suggestion by Green, Krieger, and Agarwal (1991) that attribute desirabilities and importances be obtained with rating scales having more categories. He used three custom-made versions of ACA which differed in the scales used for attribute desirabilities and importances.

His product category was grocery stores, described with 10 attributes having an average of three levels. He used 604 respondents screened to be "primary grocery shopper," between 18 and 65, "not competitively employed" and not having participated recently in a research project. Interviewing took place in six shopping centers.

Each respondent also rated five grocery store concepts using a 100-point "likelihood to shop" scale. The ratings were done twice by each respondent, before and after the ACA task.

McLauchlan found no differences in accuracy of prediction for the three versions of ACA, suggesting that ACA would not be improved by using rating scales with more categories. However, he also noted that all three versions were relatively unsuccessful at predicting ratings of holdout concepts. He found a problem with the consistency of the holdout data: about 40% of all respondents were inconsistent in terms of which of the five holdout concepts was most highly rated. In comparing ACA predictions to store ratings, he found that more than half of the error variance was due to unreliability of the ratings themselves. He recommended that in validity studies researchers pay more attention to obtaining reliable measures of that which is to be predicted by conjoint analysis.

Huber, Wittink, Fiedler, and Miller (1993) compared ACA with full profile analysis using an experimental design that examined several factors. The product category was refrigerators, and 400 respondents were recruited in shopping malls in 11 cities. All interviews, including full profile

interviews, were administered by computer. They used these experimental treatments:

- Each respondent provided both full profile ratings and ACA judgments. Half did ACA followed by full profile, and half did full profile followed by ACA.

- Half the respondents dealt with five attributes, and half with nine (which included the basic five).

- The number of intermediate levels was varied for four of the attributes. This was done to study the effect of the number of levels on attribute importances.

- For the full profile tasks, order of presentation of attributes within profiles was varied. (No systematic order effects were observed.)

Each treatment had two variations, and a factorial design was used with 25 respondents in each of the 16 cells.

Each respondent also performed four validation tasks, which were administered twice during the interview. Two of the tasks presented pairs of refrigerators and asked which the respondent would be most likely to buy. The other two tasks presented triples of refrigerators, and asked which the respondent would be most and least likely to buy. The validation tasks were identical for all respondents, and used only the most extreme levels of the basic five attributes, which were included in all respondents' conjoint tasks. Data from each triple were transformed into three inferred paired-comparison responses, and the choice data were regarded as two replications of eight paired comparisons.

Overall, the percentage of consistent choices in paired-comparison tasks was 80%. ACA had an overall hit rate of 73% when predicting binary choices, and full profile had an overall hit rate of 68%. Both ACA and full profile performed better with five than with nine attributes (ACA had a 4% or 5% advantage in each case). Also, a strong order effect was observed. ACA had a hit rate of about 73% regardless of whether it was the respondent's first or second conjoint task. However, full profile had a hit rate of only 64% when it was the first conjoint task, but improved to 74% when it followed ACA. The authors concluded that ACA provided a training opportunity that enhanced the quality of subsequent full profile evaluations.

Respondents also rated each conjoint task on nine descriptive scales. The tasks were perceived similarly for the most part, but there were two significant differences: ACA was perceived as more enjoyable, and full profile was perceived less favorably in terms of taking long to do.

Although both ACA and full profile performed well in this study, the comparative results favored ACA. In discussing possible reasons why ACA dominated full profile, the authors commented that full profile may have had a disadvantage in being administered by computer. Respondents saw only one concept at a time, and that may have made the task more difficult than if they had been able to sort cards into piles. Unfortunately a similar study has not been done in which full profile was administered with a card sort.

Recent research with hierarchical Bayesian (HB) estimation (ACA/HB) demonstrates that even better ACA results can be obtained. HB estimation for ACA has been available since the late 1990s and generally produces better part-worth utility estimates than the standard ordinary least-squares (OLS) utility estimation included in the base ACA package (see ACA/HB Technical Paper).

Recommendations for Choice of Method

Many methods are available for collecting and analyzing conjoint data, and the researcher contemplating a conjoint study must choose among them. We at Sawtooth Software have had many years of direct experience with these methods, as well as the benefit of many conversations with users of our own and other software. Based on that experience, we offer the following suggestions:

The *Full Profile Method* was the original conjoint method introduced to the marketing research community, and it remains a standard. Green and Srinivasan (1990) recommend use of the full profile method when the number of attributes is six or fewer. We agree that six is a useful guideline, but the actual number of attributes that can be used in full-profile depends on the length of attribute text and the familiarity of respondents with the product category. We think respondents are likely to become overloaded and confused when confronted by large numbers of lengthy profiles. Our experience is that, when there are more than about six attributes, and pricing research is not the goal, ACA works better. We also think the weight of evidence shows that ACA works at least as well as full profile when there are fewer than six attributes (for example, see Huber et al., 1993) and pricing research is not the goal, although with few attributes ACA has no compelling advantage.

The *ACA System* was developed specifically for situations where there are many attributes and levels. Most of ACA's questions present only small subsets of attributes, so questions do not necessarily become more complex when there are many attributes in the study. With more than six attributes, we think ACA is likely to be the more appropriate method when pricing research isn't the goal.

Like most full profile applications, ACA is a "main effects only" model, and assumes there are no interactions among attributes. Many conjoint practitioners agree that one must remain alert for the possibility of interactions, but that it is usually possible to choose attributes so that interactions will not present severe problems. Like other conjoint methods, ACA can deal with interactions in a limited way by defining composite variables. For example, we could deal with an interaction between car color and body style by cross-classifying the levels:

Red Convertible
Black Convertible
Red Sedan
Black Sedan

However, if the attributes in question have many levels, or if an attribute (such as price, for example) is suspected of having interactions with many others, then composite attributes will not be enough. In that case too many parameters must be estimated to permit analysis at the individual level, and the most common solution is to evaluate interactions by pooling data from many respondents. ACA has been shown to have weaknesses in pricing research, where it often underestimates the importance of price. We generally recommend that either CVA or CBC (described below) be used if pricing research is the main purpose of your study. Some researchers include price as an attribute in ACA, but adjust the price utilities using information gained from a secondary full-profile conjoint exercise or series of holdout tasks. For a useful method of adjusting the weight of price in ACA, see "Calibrating Price in ACA: The ACA Price Effect and How to Manage It" at www.sawtoothsoftware.com/techpap.shtml.

ACA, like other conjoint techniques involving a self-explicated section and/or partial profile trade-offs, can be especially problematic if the attributes are not independent. Attributes that are perceived to imply one another or have similar meaning can be biased upward in importance due to "double counting."

The *CVA System* is conjoint software first introduced by Sawtooth Software in 1990 for traditional full-

profile conjoint analysis. It is a good technique when the number of attributes is about six or fewer. It often does a better job than ACA in pricing research. CVA uses a paired-comparison interview that can be administered either by computer or with paper and pencil.

The *CBC System* is conjoint software first introduced by Sawtooth Software in 1993 to administer and analyze "Choice based Conjoint" studies. CBC conducts paper- or computer-administered interviews in which the respondent sees a series of choice tasks. Each task displays several concepts and asks which the respondent would choose from that set. Optionally, a "would choose none" option may be offered. Attribute levels in each concept are varied in such a way that values similar to conjoint utilities can be estimated for each attribute level. Analysis can be done at the group level with logit, which is included with the base CBC system. Additionally, latent segment-based utilities can be generated using Latent Class. Individual-level utilities can be estimated from choice data using hierarchical Bayes (CBC/HB).

We think CBC provides three potential advantages over other conjoint methods:

1. It presents tasks that may be more "realistic" than other conjoint methods. In the real world, buyers express their preferences by choosing one product or another, rather than by rating or ranking them.

2. By including the opportunity for the respondent to choose "None of these," CBC may be able to deal more directly with questions relating to volume (rather than just share). By contrast, ACA models volume using "Likelihood of Purchase" simulations, based on responses to Calibrating Concepts.

3. Because CBC analysis can be done for groups rather than for individual respondents, sufficient information is available to measure interactions as well as main effects.

However, CBC has the disadvantage of being an inefficient way of collecting data. The respondent must read and process several full profile concepts before giving each answer. To keep the respondent from becoming overloaded and confused, we suggest using no more than about six attributes with CBC (again, depending on the length of attribute text and familiarity of respondents with the product category). CBC should be considered when there are few attributes and when interactions are likely to occur, both of which are often true of pricing studies.

This paper presents a brief overview of conjoint techniques in general and ACA in particular. Literature about conjoint techniques abounds in current journals and text books. We urge the reader to consult the papers referenced in this paper and to consult additional sources, such as those listed in our Technical Papers Library at www.sawtoothsoftware.com/techpap.shtml.

Appendix A: Estimating Respondent Utilities

Initial Estimates

Before the first paired-comparison question we construct "prior" utility estimates for each attribute level as follows:

> If rank orders of preference are asked we convert them to relative desirabilities by reversing them. For example, ranks of 1, 2, and 3 would be converted to values 3, 2, and 1, respectively. If desirability ratings are asked, those are retained "as is."
>
> The average for each attribute is subtracted to center its values at zero. For example, desirability values 3, 2, and 1 would be converted to 1, 0, and -1, respectively.
>
> The values for each attribute are scaled to have a range of unity. For example, desirability values of 1, 0, and -1 would be converted to .5, 0, and -.5.
>
> The importance ratings for each attribute are scaled to range from 1 to 4, and then used as multipliers for the unit-range desirability values. Thus, if an attribute has desirabilities of .5, 0, and -.5 and an importance of 3, we get -1.5, 0, and 1.5.

The resulting values are initial estimates of utilities, with these characteristics:

> For each attribute the range of utility values is proportional to stated importance, and attribute importances differ by at most a factor of 4.
>
> Within each attribute the values have a mean of zero, and differences between values are proportional to differences in desirability ratings or rank orders of preference.

Updating

Estimates of the respondent's utilities are updated after each paired-comparison response. First consider the general case of how least squares regression coefficients can be updated to include the effect of an additional observation.

> Let X be a matrix of predictor variables with a row for each of n observations and a column for each variable.
>
> Let y be a vector of responses for the first n observations.
>
> Let z' be a row vector of predictor values for a new observation, appended as a row to X.
>
> Let r be a response for the new observation.

Considering only the first n observations, we have the regression equation: $X b_n \sim y$

where $$b_n = (X'X)^{-1} (X'y) \qquad (1)$$

is the vector of coefficients that would be obtained by least squares estimation based on the first n observations.

Now consider adding one observation. The expanded layout is:

$$\begin{bmatrix} X \\ z' \end{bmatrix} b_{n+1} \sim \begin{bmatrix} y \\ r \end{bmatrix} \qquad (2)$$

where

$$b_{n+1} \sim = (X'X + z'z)^{-1}(X'y + zr)$$

is the least squares estimate based on n+l observations. Suppose we already have b_n , X, y, z, and r, and we want to obtain b_{n+1}. First consider an identity. Let

$$v = (X'X)^{-1} z. \qquad (3)$$

Then it can be shown that

$$(X'X + zz')^{-1} = (X'X)^{-1} - \frac{v v'}{1 + v'z} \qquad (4)$$

Substituting into equation (2), we get

$$b_{n+1} = b_n + v \frac{r - z' b_n}{1 + v'z} \qquad (5)$$

Equation (5) gives a formula for updating the estimate of utilities following each response, a relatively easy computation since the numerator and denominator on the right are both scalars. We must also update the inverse as in equation (4). That is also fairly easy since the vector v is already available. If we are dealing with k attribute levels, then an updating cycle requires about $2k(k+1)$ multiply and add operations. This is a significant savings when compared to the cost of re-estimating "from scratch" after each response, and the final results are identical.

Now consider how this scheme is applied to the specific situation in ACA:

> Before the first updating we set X equal to the identity matrix and both b_n and y equal to the inital utility estimates.
>
> The vector z consists of plus and minus 1's and 0's. An element equals 1 if the corresponding attribute level appeared in the concept on the right of the screen, -1 if in the concept on the left of the screen, and 0 if that level did not appear in either concept.
>
> The response r is coded so that +4 means "strongly prefer right," -4 means "strongly prefer left," and 0 means indifference.

Pairs Utilities:

An independent variable matrix is constructed with as many columns as levels taken forward to the pairs questions. If a level is displayed within the left concept, it is coded as -1; levels displayed within the right-hand concept are coded as +1. All other values in the independent variable matrix are set to 0.

A column vector is created for the dependent variable as follows: the respondents' answers are zero-centered, where the most extreme value for the left concept is given a -4, and the most extreme value on the right +4. Interior ratings are fit proportionally within that range.

Each pairs question contributes a row to both the independent variable matrix and dependent variable column vector. Additionally an n x n identity matrix is appended to the independent variable matrix, where n is the total number of levels taken forward to the pairs questions. An additional n values of 0 are also appended to the dependent variable matrix. The resulting independent variable matrix and dependent variable column vector each have t + n rows, where t is the number of pairs questions and n is the total number of levels taken forward to the pairs questions. OLS utility estimates of the n attribute levels are computed by regressing the dependent variable column vector on the matrix of independent variables.

Combining the Priors and Pairs Utilities

The priors and pairs utilities are normalized to have equal sums of differences between the best and worst levels of each attribute across all attributes. (Note that the procedures described above automatically result in zero-centered utilities within attribute for both sets of utilities.) The priors utilities for levels also included in the pairs questions are multiplied by n/(n+t), where n is the total number of levels used in the pairs section, and t is the number of pairs questions answered by the respondent. Any element in the priors that was not included in the pairs section is not modified. The pairs utilities are multiplied by t/(n+t). The two vectors of utilities (after multiplication by the weights specified above) are added together. These are the final utilities, prior to calibration.

As a final step, the utilities are calibrated. It is widely recognized that the utilities arising from most conjoint methods are scaled arbitrarily, and that the only real information is contained in the relative magnitudes of differences among them. So far, that is true of ACA as well.

However, the calibration concepts permit scaling of utilities in a non-arbitrary way. In any product category, some respondents will be more interested and involved than others. We attempt to measure each respondent's degree of involvement by asking "likelihood of buying" questions for several concepts that differ widely in attractiveness. The data obtained from those concepts is useful in three ways:

Correlations between utilities and likelihood responses may be used to identify unmotivated or confused respondents. Respondents whose likelihood responses are not related to their utilities should probably not be included in subsequent preference simulations.

The level of likelihood responses may identify respondents who are more or less involved in the product category. Respondents who give low likelihood responses even to concepts custom-designed for them should probably be treated as poor prospects in simulations of purchase behavior.

Variation in likelihood responses may also identify respondents who are "tuned in" to the product category. A respondent who gives a low likelihood rating to the least attractive concept and a high rating to the most attractive should be made to respond sensitively in preference simulations, whereas someone who gives every concept similar likelihood values should be made insensitive in simulations.

Each respondent is first shown what should be the least attractive possible concept, followed by the most attractive possible concept, as determined from his or her own answers. Those two concepts establish a frame of reference. The remaining concepts are of middling attractiveness. We determine an intercept and one regression coefficient to apply to utilities to best predict logits of likelihood responses. Those parameters are then used in a final scaling of utilities, which are therefore no longer arbitrarily scaled. The procedure is as follows:

Let: p = the predicted likelihood of buying a concept
x_1 = the concept's utility based on the final "uncalibrated" utilities
b_1 = the coefficient used to weight the utilities
a = an intercept parameter

The actual likelihood response is a single digit on a scale with n points. Responses are trimmed to the range of 5 to 95. Then, using the logit transformation, we model buying likelihood as a function of the respondent's utilities as:

$$\ln [p / (100 - p)] \sim a + b_1 x_1$$

If the regression coefficient is negative we assume the estimation is faulty and use a conservative positive value. The r-squared (measure of fit) reported in the data file is set to 0 in such cases. If the calibration concepts section is not included in the interview, the respondent is assumed to have answered 0 and 100 to the worst and best concepts, respectively, and 50 to the other concepts.

To calibrate the utilities, each is multiplied by b1. The intercept a is divided by the number of attributes, and the quotient added to the utility for every attribute level. The utilities can be added up and antilogs of their sums are predictions of odds ratios for claimed likelihood of purchase of any concept, just as though that concept had been included in the calibration section of the questionnaire.

A Note about Hierarchical Bayes Estimation

OLS has been successfully used in ACA calculations for over two decades. However, a relatively new technique called hierarchical Bayes (HB) estimation provides a more theoretically satisfying way of combining information from priors and pairs. The results are also usually better from the practical standpoint of improved predictions of holdout questions. We recommend that the interested reader investigate ACA/HB by downloading the technical paper from our Web site (www.sawtoothsoftware.com).

APPENDIX B: Choosing Paired-Comparison Questions

The first part of an ACA interview is concerned with screening the attribute levels and learning enough about the respondent's preferences to construct initial utility estimates. After that is done we begin the paired-comparison section, in which pairs of concepts are shown and preference questions are asked. Following each response we update our estimates of utilities and then decide what pair of concepts to present next.

The number of possible concepts is very large, and we need some reasonably efficient procedure to choose a pair of them at each stage that will be most beneficial in some way. There are several principles to keep in mind when thinking about how to choose concepts.

> Concepts should be chosen by a method that gives the author as much control as possible over the interview, in terms of the complexity of the concepts and the number of questions asked.
>
> The design should be as "balanced" as possible. Observations should be spread as evenly as possible over all attribute levels, and the columns of the design matrix should be as nearly orthogonal as possible.
>
> We should ask the respondent questions that require careful consideration. There is no point in asking questions for which we already know the answer, such as "High quality at a low price" versus "low quality at a high price." We learn more if we choose concepts nearly equal in attractiveness.

Our procedure addresses these points. The author may specify the number of attributes to appear in each concept. The range is from two to five. It is possible to start with only two attributes per concept and, after the respondent has gained experience, to increase their complexity.

The concepts in a pair always have different levels of the same attributes. Our procedure for choosing those concepts is:

> Count the number of times each pair of attributes has appeared together in any concept. Pick a set of attributes at random from among those whose members have previously appeared together the fewest times.
>
> For each of the chosen attributes, repeat similar logic to find levels that have been paired least frequently.
>
> Examine all possible ways of combining these levels into concepts (with just two attributes there are only two possible ways; with 5 attributes there are $2^5 = 32$ ways). Find the pair of concepts most nearly equal in attractiveness, using the current estimates of the respondent's utilities.
>
> Randomly determine which concept will appear on each side of the screen.

ACA lets the author specify certain pairs of attribute levels that must not appear together in the same concept. The procedure described above is modified slightly to take account of such prohibitions. When concepts are described on only two attributes, ACA chooses the first few questions in a slightly different way. (When the concepts are described on only two attributes, it would be possible to blunder into a design in which the attributes would be divided into subsets in such a way that those in one subset would never be paired with those in another subset. Such designs would provide no information about the

relative importance of attributes in different subsets, and ACA automatically corrects the design in such a situation.)

ACA's designs usually have good statistical efficiency, although they are not strictly orthogonal. Statistical efficiency is increased as more attributes are used in each concept, and it is also possible to produce concepts more nearly equal in attractiveness when there are more attributes with which to work. However, using larger numbers of attributes has the unfortunate consequence of making the questions more complicated, and respondents are more easily confused.

Both anecdotal and experimental evidence has shown that it is usually best to start with only two attributes per concept and, after a few pairs, to increase the number of attributes to three. Beyond three attributes, gains in efficiency are usually offset by respondent confusion due to task difficulty.

References

Cattin, Philippe, and D.R. Wittink (1982), "Commercial Use of Conjoint Analysis: A Survey," Journal of Marketing, 46 (Summer), 44-53.

Finkbeiner, Carl and P.J. Platz (1986), "Computerized Versus Paper and Pencil Methods: A Comparison Study," paper presented at the Association for Consumer Research Conference, Toronto (October).

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.

Green, Paul E., A.M. Krieger and M. K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Cautions and Caveats," Journal of Marketing Research, 28, (May), 215-22.

Huber Joel C., D.R. Wittink, J.A. Fiedler, and R.L. Miller (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," Journal of Marketing Research, 30 (February), 105-114.

Johnson, Richard M (1991), "Comments on Studies Dealing With ACA Validity and Accuracy, With Suggestions for Future Research," published by Sawtooth Software.

King, W. Christopher, Aaron Hill, and Bryan Orme (2004), "The 'Importance' Question in ACA: Can It Be Ommitted?" Sawtooth Software Conference Proceedings, Sequim, WA: Sawtooth Software.

MacBride and Johnson (1979), "Respondent Reaction to Computer-Interactive Interviewing Techniques," paper presented at the ESOMAR Conference.

McLauchlan, William G. (1991), "Scaling Prior Utilities in Sawtooth Software's Adaptive Conjoint Analysis," 251-68 Sawtooth Software Conference Proceedings, Ketchum, ID: Sawtooth Software.

Mehta, Raj, W.L. Moore, and T.M. Pavia (1992), "An Examination of the Use of Unacceptable Levels in Conjoint Analysis," Journal of Consumer Research, (December), 470-76.

Tumbusch, James J. (I 99 1), "Validation of Adaptive Conjoint Analysis Versus Standard Concept Testing," 177-83, Sawtooth Software Conference Proceedings, Ketchum, ID: Sawtooth Software.

Vriens, Marco, Huber, Joel, and Dick 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., and P. Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," Journal of Marketing, 53 (July), 91-6.