Proceedings of the Sawtooth Software Conference

1992

Foreword

We are pleased to present these Proceedings of the fifth Sawtooth Software conference, held in Sun Valley, Idaho, in July, 1992.

These papers, which illustrate today's wide use of PC's in interviewing and sophisticatred analysis techniques, sparked lively discussion, challenges, and questions. Each presentation was followed by a prepared discussant, who helped foster audience participation by offering contrasting or complementary views. Many discussants chose to include their comments in this volume. We hope these Proceedings and the resulting interchange of ideas will form the basis for further research, papers, and meetings.

The papers and discussant comments are in the words of the authors; only light copy editing was performed. We thank all participants for their contributions to the advancement of the theory and practice of PC-based interviewing and analysis.

Margo Metegrano
Editor
October, 1992

Table of Contents

BEST PRACTICES IN DISK-BY-MAIL SURVEYS

Karlan J. Witt
IntelliQuest, Inc.
Steve Bernstein
Apple Computer

INTRODUCTION

The proliferation of personal computers and the development of personal computer-based survey software in the 1980s has created the opportunity to conduct disk-by-mail surveys (henceforth referred to as DBM).

This paper is organized into ten sections, beginning with a brief review of the history and background of DBM surveying. It then discusses when it is appropriate to use DBM, and describes factors that affect response rate. The next few sections provide respondents' evaluations of the disk-based survey task, outline its limitations, provide a brief examination of non-response, and conclude with cost and timing comparisons for DBM surveys and other data collection methodologies. The document then describes a case example from Apple Computer, including a discussion regarding the uses of the information. The paper then summarizes the best practices for DBM surveys, and ends with a look at the future of DBM surveying.

This paper focuses primarily on conducting DBM studies in a business-to-business environment. In specific cases, such as a customer satisfaction or new product follow-up for personal computer products, DBM can also be used successfully in the home market. However, because the penetration of personal computers in the home market is estimated to be only 32% (Source: Electronic Industries Association, 1992) and many of the home computer systems are not IBM-compatible, IntelliQuest does not suggest using DBM for the general home market at this time.

BACKGROUND OF DISK-BY-MAIL SURVEYS

A few companies and research agencies have been using DBM as a data collection methodology for years. Computer hardware, software, and telecommunications companies use disk-based surveys for product registration as well as custom research. Some of the types of research for which disk-by-mail is commonly used include:

- Product registration
- New product follow-up
- Customer satisfaction
- New product design
- Concept testing
- Pricing
- Product positioning
- Brand image/positioning
- Market segmentation

While for each of these types of research other methodologies may be used, DBM offers some advantage by collecting information in a manner which is superior to other data collection methodologies (see discussion next section).

IntelliQuest has used disk-by-mail surveys (DOS-based) since 1986, surveying 50,000+ respondents nationally and internationally. Apple Computer and IntelliQuest jointly developed MacSurvey software in 1990, designed specifically for DBM applications for the Macintosh computer.

WHEN TO USE DISK-BY-MAIL SURVEYS

DBM surveys offer unique opportunities for collecting data. They are not, however, a panacea for data collection. As with other mail surveys, DBM surveys can offer lower cost, respondent convenience, anonymity of the respondent, ability to administer lengthy surveys, and elimination of interviewer bias (Alreck and Settle, 1985; Joselyn, 1977; Kress, 1988; Lehman, 1979; Peterson, 1982; Rossi, Wright, and Anderson, 1983). Additionally, DBM may perform better on issues usually considered disadvantages of paper surveys, such as low response rates and slow turnaround times. This section of the paper describes when DBM is an appropriate data collection methodology.

1. **Research Sample.** One of the decisions that must be made prior to choosing DBM as a data collection methodology is the selection of the research population. The research population affects the choice of DBM in two ways:

- Access of respondents to personal computers to complete the survey
- Appropriateness of a disk-based survey for the target audience

As businesses and consumers continue to adopt technology, the availability of a personal computer is becoming less of a hindrance to DBM surveys. Still, it is likely that less than 100% of the target audience has access to a PC to complete the survey. In each study, this incidence rate needs to be addressed.

Internationally, DBM surveys offer a methodology to collect more consistent information across many countries without the systematic differences which are introduced when using other data collection methodologies. As with all mail surveys, regulations and customs vary from country to country. DBM surveys may not be accepted in all countries (Sawtooth Software, 1991).

Many populations, such as data processing professionals or purchasing agents in Fortune 500 companies, have ready access to PCs. Other populations may not use a PC on a regular basis, but have access if they need it. Still some groups have no immediate access to a PC, and to the extent that these respondents are systematically different than other respondents, a bias is introduced into the sample. While respondents may be pre-screened for access to a PC, the potential respondents without access need to be analyzed carefully. Later in this paper, the issue of non-response for DBM surveys is addressed.

Further, not all office employees who have access to PCs are comfortable using them. While one advantage of DBM surveys is the familiar environment they provide for many respondents, this may introduce another type of bias when using DBM in populations with respondents who are intimidated by using computers.

2. **Questionnaire Design**. The next major issue which may dictate the use of a disk-based survey — if not necessarily a disk-by-*mail* survey — is the questionnaire design. DBM surveys provide a superior methodology for collecting many types of information, such as Adaptive Conjoint Analysis (ACA System by Sawtooth Software), while making others, such as unaided awareness, more difficult.

At the questionnaire design stage, certain study objectives may make a DBM survey an attractive alternative, such as an adaptive conjoint design or a concept test where it is vital that respondents not look ahead at the survey,

Other objectives, such as unaided awareness, may be easier to capture on a DBM survey than on a traditional paper-by-mail survey, since the respondent cannot look ahead at an aided list. However, such responses will require additional coding, which is not necessary when collecting unaided awareness data over the phone, since interviewers are likely to have lists of possible answers.

Similarly, DBM surveys offer an opportunity for collecting open-end information where the respondents can elaborate and record their thoughts without the concern of what "the interviewer" might think, and without interviewer transcription errors. While many respondents may provide better information in this environment, others may provide unclear or less detailed information than is desired when there is no interviewer present to probe and clarify the responses.

3. **Data Collection Methodology**. After the sample and questionnaire topics have been developed, a decision must be made regarding the data collection methodology. This decision must weigh the issues discussed above, and evaluate the benefits of each methodology within the available project budget.

While DBM surveys are often used where other data collection methodologies would suffice, there are many instances when it is a superior methodology:

- Programmed automatic skip patterns give respondents only relevant questions
- Survey can incorporate adaptive modules, such as ACA, which are most easily self-administered (discussion next page)
- Open-end questions capture lengthy verbatim answers without interviewer bias
- Respondents perceive the survey to take less time to complete than it actually does
- Randomization reduces order bias within lists and across questions
- Less respondent fatigue than for a phone survey
- Respondents cannot look ahead, as they can in a paper survey that is too long or too complex

For concept-testing, or studies that require display of visual information or control of the respondent task, DBM surveys:

- Prevent respondents from looking ahead to concept or follow-up questions
- Use survey software designed to allow incorporation of graphical images
- Allow for a greater range of measurement (allows the researcher to use scales not possible to administer via telephone)

For the implementation of complex survey designs, DBM offers or supports:

- Conjoint or other multivariate or adaptive techniques
- Ability to show lengthy explanations on complex lists of responses

For example, using ACA, it is possible to collect powerful data for use in product design, pricing, market segmentation, and other applications of market structure data. The questions provided by ACA, like the example below, are more easily asked in a DBM self-administered survey rather than over the phone.

Figure 1

WHICH NOTEBOOK COMPUTER WOULD YOU PREFER?
Type a number from the scale below to indicate your preference.

2 hour battery life 8" x 10" 5 lbs Toshiba $3,100	OR	4 hour battery life 9" x 11" 7 lbs IBM $2,700

Strongly Prefer Notebook Computer On Left — Don't Care — Strongly Prefer Notebook Computer On Right

1----------2----------3----------4----------5----------6----------7----------8----------9

For populations that enjoy using technology, DBM surveys:

- Can improve response rates
- Create a very comfortable environment in which to work

For populations that are difficult to reach by phone and who have access to personal computers, DBM surveys:

- Allow respondents to complete the surveys at their convenience
- Produce response rates which tend to be higher than for paper-and-pencil surveys

For longitudinal studies, DBM surveys:

- Eliminate interviewer bias because survey administration is consistent across waves
- Permit complexly programmed surveys to be used repeatedly

Thus, the applications for DBM surveys are very broad, and are becoming more easily applied as the general business population adopts the use of personal computers.

FACTORS AFFECTING RESPONSE RATE

As with any mail survey, there are many factors in a DBM survey which affect response rate (Peterson, 1989). For DBM surveys, even more than for other types of data collection methodologies, non-respondents are potentially systematically different from respondents in at least one aspect: their access to personal computers. Although respondents may be screened for access to PCs, this may introduce a source of non-response bias.

It is critical to consider this and other sources of non-response in a study, and create a well-balanced approach, including such aspects as:

The Survey

There are several aspects of the survey itself which affect response rate for any mailed survey. Here are several considerations along with an explanation of how each uniquely affects a DBM survey.

1. **Saliency of 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 the type of respondent, introducing a bias into the study.

2. **Length of survey.** There are two components to survey length which elicit behavioral responses from the potential respondents. The first is the expected length of time to complete the survey, if reported to the respondent in the cover letter. This eliminates certain respondents who are unwilling to commit that 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 (Bahner, 1987).

It is important to note that on disk-based surveys, respondents' perception of elapsed time is less than the actual time lapsed. In a study conducted by IntelliQuest (in 1989), respondents estimated that the questionnaire took less time to complete than it actually did. On average, respondents stated that it took nineteen minutes to complete the survey, when the average time registered by the clock on the survey disk registered thirty-three minutes. Published research supports this finding (Higgins, Dimnik, and Greenwood, 1987).

In another study recently conducted by IntelliQuest (in 1992), the time reported to the respondent was varied. One-third of the respondents (randomly selected) were told the survey would take approximately 15 minutes, one third were told 20 minutes, and the last third were not given an expected time to complete the study. Figure 2 shows the response rate for each group. (A reported time estimate of 15 minutes produced a significantly higher response rate than either the 20 minute reported time estimate or no time estimate. A chi-squared goodness of fit analysis (using Yates correction) shows a significant difference, $\alpha = .10$.).

Figure 2

Stated Time to Complete Survey	Response Rate
15 minutes	44% returned survey
20 minutes	36% returned survey
None	38% returned survey

3. **Limited time demands**. The shorter the survey, typically the higher the response rate. This is a critical component to gaining an appropriate response from the over-surveyed populations and the respondents who place a high value on their time (discussed below).

4. **Respect for respondents' time; high professional ethics**. While there is evidence that respondents will respond to longer surveys using a DBM methodology, it is the responsibility of the researcher to always respect respondents' time.

The Sample

5. **Composition of research sample**. Certain populations, such as purchase decision influencers and senior executives, are frequently asked to participate in surveys, and others place a very high value on their time. Both of these groups typically demonstrate lower than average response rates in research studies.

6. **Access to personal computers**. The majority of DBM surveys are conducted on IBM-compatible personal computers. Whether a Macintosh survey software diskette is offered as an additional option depends largely on the target audience and the objectives of the research. In either case, respondents must be known to have — or must be screened for — access to a personal computer. IntelliQuest estimates that approximately half of the Fortune 1000 business sites have Macintosh computers installed, and nearly 100% have IBM-compatible computers installed. Penetration of computers varies with company size and industry.

Depending on the subject matter being measured, respondents without access may or may not be systematically different than respondents with personal computers. It is recommended that respondents without access to PCs be asked to respond to primary demographic and firmographic questions, as well as attitudinal questions about the subject being measured to analyze the potential for bias in the non-respondent sample.

It is also important to provide both 3.5" and 5.25" diskettes to let respondents take the survey using a disk which is compatible with their system. While it is possible to pre-screen the respondents for the preferred disk size, IntelliQuest has not found that the most effective procedure. Sorting disks and respondents adds administrative time to the project, and even though the respondents have been asked which size they prefer, a portion will not know the correct size, or will indicate an incorrect size.

7. **Convenience of taking the survey**. A DBM survey provides the convenience of completing the survey at a time of the respondents' choosing. This convenience provides an advantage of DBM surveys over telephone or other data collection methodologies, and produces a higher response rate overall. Additionally, providing all materials necessary for the respondent to complete and return the survey, such as the postage-paid return disk mailer, will increase response rate.

The Presentation of the Survey

8. **Sponsorship of survey disclosed.** One of the key factors impacting response rate is whether or not the sponsor of the research is disclosed. While it is clearly 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 (in 1989), 35% of respondents stated that their attitudes toward the sponsor improved as a result of receiving the survey from the sponsor (Zandan and Frost, 1989).

9. **Guarantee of anonymity or confidentiality**. Mailed surveys in general offer respondents some degree of anonymity; the lack of anonymity is often a source of non-response in other data collection methodologies. This anonymity helps both on an item non-response and a unit non-response level.

10. **Priority or First Class mail**. Respondents react in some way to a package as soon as it arrives. The packaging and professional appearance of the package and its contents will be the respondents' first impression. At this point the goal is to have the respondents complete the survey immediately, or at least to have them keep the survey. Even if the survey is not thrown away at this stage, the respondent still may not choose to respond at a later time.

In one IntelliQuest study (in 1988), a split sample was used to test the effect of First Class vs. bulk rate postage on response rate. The response rate from the sample using First Class postage was 32%, while the response rate for the bulk rate sample was 27% (Pilon and Craig, 1988). Note that the response rates overall were low for this study, resulting from the nature of the sample and the length of the survey.

In debriefing with IntelliQuest respondents from another study (in 1992), it was found that faster mailing methods (for example, Federal Express or USPS Priority Mail) connote that the survey is of great importance to *the sponsor of the research*, and the respondents are therefore more likely to respond, and respond soon after receiving the survey.

11. **Personalized cover letters and envelopes.** This is a specific illustration of the packaging discussion above. The more professional the packaging and presentation from the research sponsor, the higher response from the sample. While personalized cover letters increase response rate, even small typographical errors in the cover letter may have an adverse effect on response rate.

One difficulty with using personalized cover letters is the availability of an appellation for the names in the research sample. It is difficult for the respondents to believe they are part of a select group when they are incorrectly addressed by assuming a "Mr." or "Ms." based on name instead of personal observation or self-reported data. A recommendation is that the respondent be addressed as "Dear Pat Jones" instead of "Dear Mr. Jones" or "Dear Ms. Jones" if no appellation is available.

12. **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.

To examine the effect of offering an incentive, IntelliQuest performed an experiment (in 1989) where potential respondents were randomly assigned to one of two groups. One of these groups was offered a coffee mug as an incentive for responding. The other was not offered an incentive. The promised incentive increased the response rate from 45% to 54% (Zandan and Frost).

Selection of incentives may also impact response. Incentives should be appealing and motivating to the target respondents. Incentives may include job-related incentives such as an executive summary of the research results, or a chance to win office equipment, or a personal incentive such as a chance to win cash, a trip, or other such prizes. IntelliQuest has found that a choice of prizes is effective, particularly when the choices consist of targeted prizes. For instance, early adopters of technology respond to high-end technology gadgetry.

It should be noted that incentives are not appropriate in all instances. Interviewing respondents in the public sector may require alternative strategies to increase response rate.

In targeting incentives, be cautious not to offend the intended respondents. In a debriefing of Fortune 500 senior executives, IntelliQuest found that some respondents felt the use of a $1 bill was insulting to them, considering the value of their time; conversely many thought $1 communicated that the survey was important to the survey sponsor.

Also, the law which applies to survey incentives is related to the one which governs lotteries and contests such as the *Publisher's Clearing House* drawing. In many cases, incentives must be offered to all potential respondents, not just to those who complete the survey. The practice of allowing non-respondents to write their names and addresses on a postcard and mail it in generally meets the legal requirements. IntelliQuest typically has less than one percent of respondents pursue this option.

13. **Printed supplementary materials such as a glossary of terms.** In addition to providing information within the text of the disk-based survey, it is often helpful to provide respondents with supplementary materials. These materials may be anything from simple definitions of terms (an example is provided in the Appendix), to elaborate illustrations of product concepts. These materials should be professionally presented and easy to interpret.

Logistics

14. **Timing of the survey (time of year mailed).** In an analysis of response rates from IntelliQuest DBM studies over the last six years, for the U.S., December is the worst time of year to mail, since many people take vacations or are too busy with other activities to respond. The second worst time, at least in a business-to-business environment, is summer, when again a large percentage of the population takes vacations. Here IntelliQuest has found that even when the desired individuals can be reached, they are often busy covering for co-workers who are on vacation, and hence do not have time to participate in a research study.

When conducting DBM research in other countries, as with any international study, consider each country's holiday schedule and incorporate it into the timeline of the data collection phase of the project.

15. **Pre-notification/pre-screening**. In many studies it is necessary to contact respondents in advance of the mailed 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 a personal computer
- Verify address

Even in instances where it is not necessary to conduct a pre-screening call for the reasons stated above, IntelliQuest has found that it increases response rate to pre-notify respondents, either by mail or phone, prior to the receipt of the DBM 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 (Pilon and Craig, 1988).

It is important for respondents to receive the survey package soon after the pre-notification. For a telephone pre-notification, IntelliQuest has found it most effective for respondents to receive the package within two to three days. With written pre-notification (letter or postcard), IntelliQuest has found it most effective for the package to be received approximately five to seven days after the notification.

16. **Second mailing or follow-up postcard or phone call**. As with pre-notification, a reminder call or postcard 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. In one IntelliQuest DBM study, the use of reminder phone calls almost doubled the response rate with a difficult-to-survey population.

TYPICAL RESPONSE RATES ON DBM STUDIES

Response rates on IntelliQuest DBM studies have ranged from 35% for an over-surveyed group conducted during the summer, to 70% when a high profile client was disclosed as the sponsor. With follow-up phone calls, a response rate of over 40% was achieved for the first group (in a 1991 study). With this potential 2X difference in response rate, it is important to heed all factors affecting response rate.

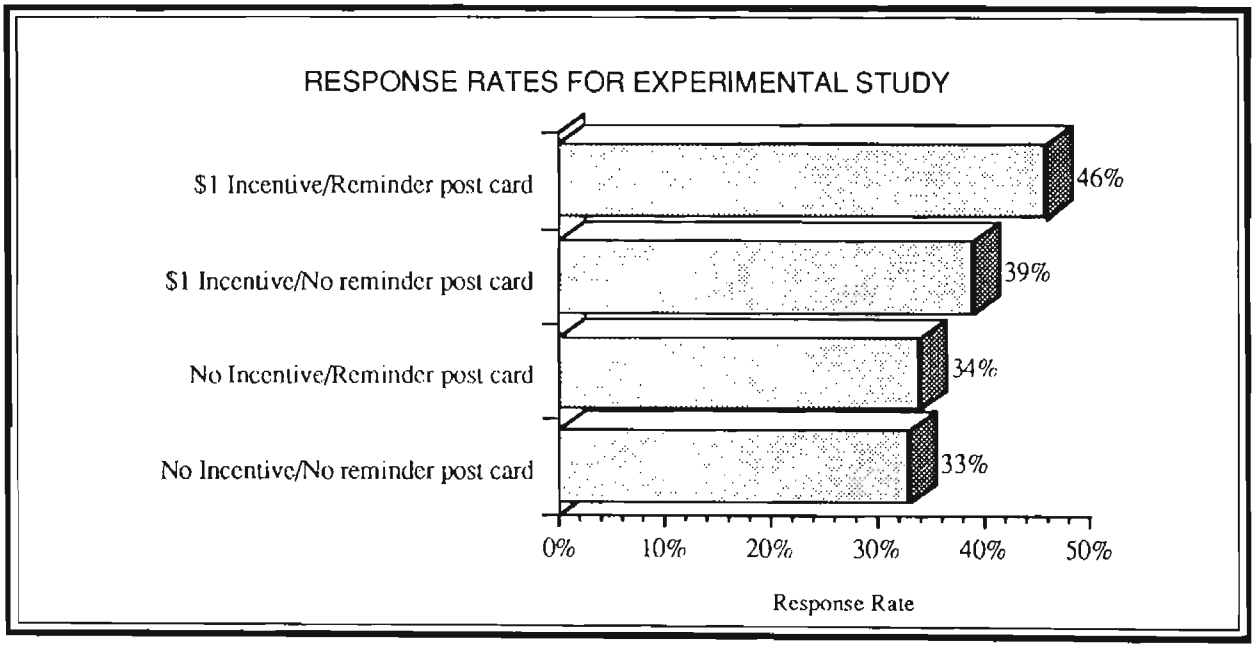
IntelliQuest typically aims for a minimum of a 40% to 50% response rate on DBM studies. Despite reports of other methods such as phone and traditional paper-by-mail having declining response rates, IntelliQuest has not experienced a decline in response rates overall on DBM surveys over the past seven years.

IntelliQuest has determined that the most influential factor to impact response rates is the disclosure of a highly respected corporate sponsor. The next most important aspect is the sample itself. Very senior executives, decision makers, and employees with select functions, for instance, will produce lower response rates. Additionally, the reported survey length affects the response rate dramatically.

To study the effects of incentives and reminder postcards on response rate, IntelliQuest conducted a study in 1987 where four groups were selected to receive a combination of a $1 incentive/no incentive and a reminder postcard sent five days later/no reminder postcard.

As shown in the following graph, the group that received both the $1 incentive and the reminder postcard had a 46% response rate. The group that received neither had a 33% response rate. The group that received $1 incentive and no reminder had a 39% response rate, and the group that did not receive an incentive, but did receive a reminder postcard had a 34% response rate. For this study, it seems that a $1 incentive worked well by itself and better in conjunction with the reminder card. The reminder card, when used alone, increased response rate only slightly (Pilon and Craig).

Figure 3



Other factors in the process, including incentive, pre-notification, reminder phone calls or postcards, a second mailing, packaging, and time of year mailed, each work to increase or decrease the response rate slightly. IntelliQuest recommends using a combination of these to accomplish the highest response rate within the project budget.

Achieving a high response rate is beneficial in two ways:

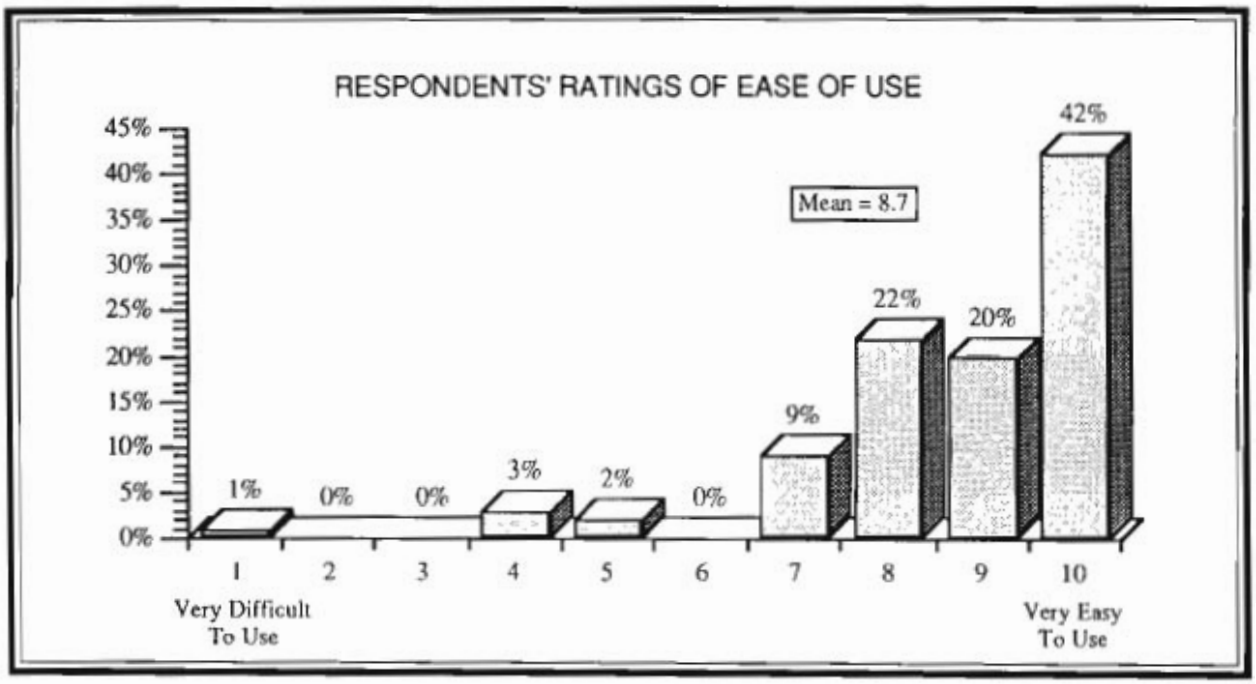
- increases representativeness of survey results
- decreases cost per completed interview

IntelliQuest has found that incentives typically pay for themselves because the increased response rate requires fewer survey packages to be mailed to achieve the same number of completed interviews.

REACTIONS TO DISK-BASED SURVEYS

Reactions to disk based surveys are generally favorable as compared to paper-by-mail and telephone methodologies (Morrison, 1988; Zandan and Frost). In a 45-minute self-administered IntelliQuest disk-based survey in 1989 of computer users, 84% rated the survey an 8, 9, or 10 on a scale of 1 to 10 with 1 meaning "very difficult to use" and 10 meaning "very easy to use."

Figure 4



Additionally, respondents indicated that the disk-based surveys were interesting, and that they were more likely to respond to a disk-based survey than to a paper survey. The novelty of disk-based surveys is one factor which produces higher response rates in DBM as compared to paper-by-mail and telephone surveys.

NON-RESPONSE FOLLOW-UP

Response rate drives much of the cost of a DBM data collection methodology. In addition to providing an economically attractive methodology to clients, the primary reason for examining response rate is to examine the extent of potential non-response bias.

What is non-response bias? Non-response bias occurs if those respondents who do not respond are systematically different from those who do respond, and if the differences affect what is being measured by the study.

Reasons for non-response bias. To examine the non-response biases for DBM surveys in 1989, IntelliQuest conducted a telephone non-response follow-up study which measured key demographic and firmographic variables, attitudinal information about the product being studied, and reasons for non-response (Zandan and Frost). The reasons given for not responding included:

- No time to complete survey
- Do not participate in surveys
- Suspect sales pitch instead of research
- Concerned about computer virus

The most sizeable of these was "do not participate in surveys," at 37%. Approximately 10% were concerned about a computer virus. Some respondents who fear computer viruses call IntelliQuest offices to confirm the validity of the survey, then proceed to complete and return the survey. In more recent non-response studies, IntelliQuest has found a small number of companies who forbid employees from using diskettes from outside the organization in the company's computer systems. To date, these companies have not been found to be systematically different from other companies in the populations studied.

The responses during a follow-up telephone interview with the DBM non-respondents were compared to those of the survey respondents, and no differences were found to exist between responders and non-responders with regard to:

- Overall satisfaction with manufacturer
- Job title
- Gender
- Income
- Education
- Age
- Company size (sales and number of employees)
- Computer expertise

Further comparisons of respondents and non-respondents did reveal two differences:

- Non-respondents were more likely to state they thought the sponsoring company was trying to "sell something" through the survey

- Respondents were more likely to express the belief that the survey would have a positive effect on the sponsoring company's customer relations

In the absence of non-response follow-up, a standard practice is to compare early returns versus late returns. Those who return surveys later tend to be more like non-responders than those who return early. If no significant differences are found between early responders and late responders, non-response bias is less likely to be a problem.

Later, this paper discusses best practices for DBM studies. These guidelines incorporate practices to minimize non-response bias.

DISK-BY-MAIL SURVEY LIMITATIONS

Although DBM surveys offer many benefits, the researcher should approach DBM studies with three cautions in mind.

1. **Abuse of medium**. DBM is subject to the same misuses other data collection methodologies have experienced, as well as some misuses unique to the medium. In particular, some potential 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 are meaningless

2. **Added complexity**. Changes in questionnaire content and flow after a questionnaire has been programmed cost time and money, and introduce possibilities for error.

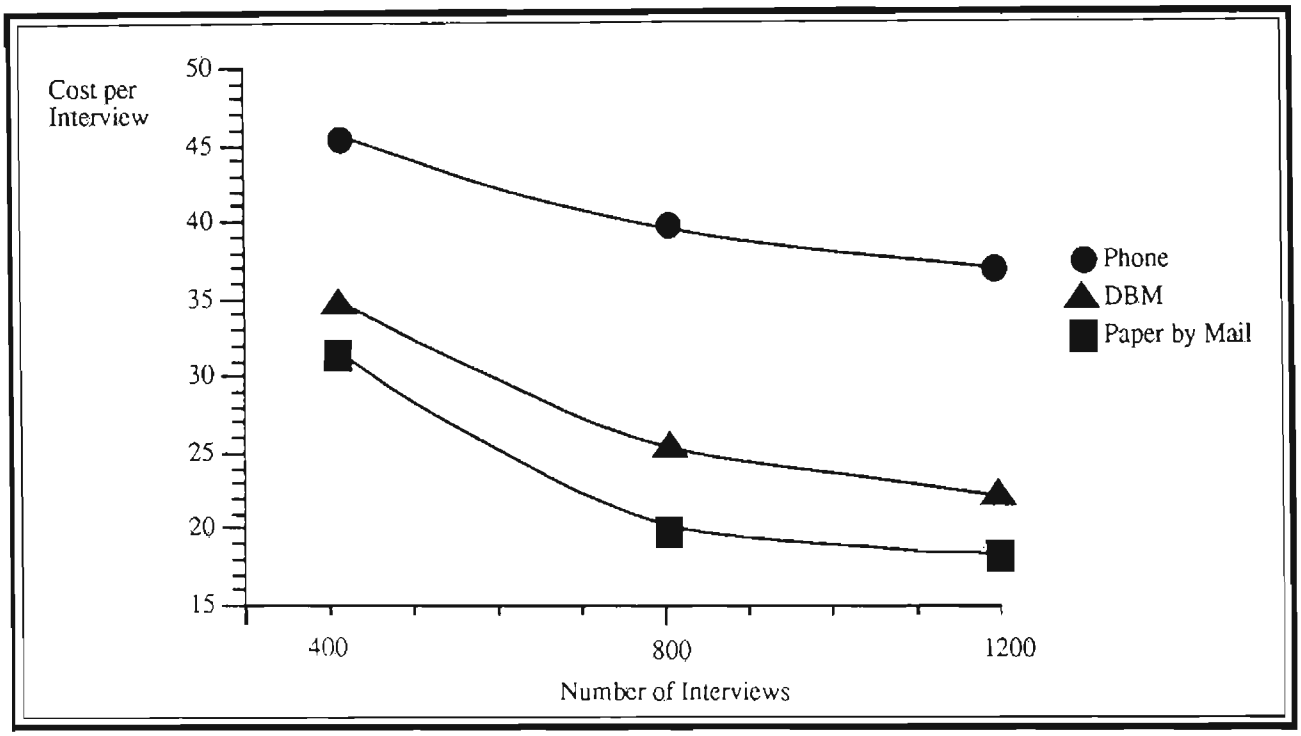
3. **Respect for the respondent**. Respondents value their time, and the researcher must provide the respondents with surveys that are professional in presentation, and, as with all surveys, ask important, relevant questions so that respondents do not feel that completing the survey is a waste of their time.

COST AND TIMING COMPARISONS

DBM surveys are most efficient for collecting complex data and for administering lengthy surveys. For comparison, Figure 5 shows a per-interview cost comparison for a lengthy survey which could be administered by phone, paper-by-mail, or DBM (centralized interviewing and personal interviews are excluded from this comparison for purposes of simplification). Data collection estimates are for a survey which would take 20 minutes by phone. Estimates are based on the following assumptions:

- For phone interviews, 1 completed interview per interviewer hour, including programming the disk-based survey
- For disk-by-mail interviews, a 40% response rate, $1 incentive, providing both 3.5" and 5.25" disks, including programming the disk-based survey in color
- For paper surveys, a 25% response rate, $1 incentive, 6-page (3 page duplex) survey, including data entry of coded data, but not verbatim responses
- All estimates are for a business-to-business survey

Figure 5



These costs include data collection costs only. If the lengthy survey also collects complex information, the alternative methodologies are further limited to telephone or DBM. It should be noted that when a pre-screen interview is required, DBM will likely be more expensive than telephone interviewing. Also, a DBM survey needs to be programmed in color to increase respondent interest. *DBM should not be used simply to save money on data collection.*

In published research available comparing DBM and paper surveys, Higgins et al found a significant difference in response rates between the two methodologies, holding other factors affecting response rate constant. In this study, 78% returned the disk-based survey, and 63% returned the paper survey.

Additionally, while split sample comparisons are not available, IntelliQuest has generally experienced a quicker response time on DBM surveys as compared to paper surveys. Higgins et al also examined response speed. In their study, the average response time for a paper-by-mail survey was 8.85 days. Their average response time for DBM was a significantly lower 6.68 days.

MAXIMIZING RESPONSE RATE IN APPLE COMPUTER'S RECENT MAC BUYERS STUDY

As an example of an application of DBM surveying, Apple Computer is in the midst of completing the third wave of a disk-by-mail survey of recent Macintosh buyers. Since by definition every member of the population of recent Macintosh buyers has a Macintosh, this provides an ideal application for DBM.

The software being used, discussed earlier, is called MacSurvey. While it does not have adaptive conjoint or perceptual mapping capabilities, it is very easy for the respondent to use, taking full advantage of Macintosh's mouse and graphical features. Its capabilities include constant sum, sorting, and analog rating scale measurements that are virtually impossible to use in phone surveys, and are often difficult in self-administered paper-and-pencil surveys. The software can also include sounds and graphics in the questionnaire, though Apple really has not taken full advantage of this yet.

In the recent buyer study, Apple is clearly — some might say aggressively — identified as the sponsor of the survey. Since the survey questions reveal the sponsorship from the beginning, there is no point in trying to conceal Apple as the sponsor, so it is played to the hilt.

The source for the sample is Apple's database of returned customer registration cards. A separate research project recently showed that Apple's registration card returns have very little non-response bias along the dimensions that matter in this study.

Respondents first received a postcard from IntelliQuest announcing the survey and asking them to participate. The postcard came from IntelliQuest, but Apple's logo was prominently displayed. Several days later, a cover letter arrived, signed by Steve Bernstein, a market research manager at Apple, on Apple letterhead, thanking the respondents for their recent purchase and asking them to complete and return the survey on the enclosed diskette. Included was an IntelliQuest 800 number respondents could call if they had any problems. Finally, IntelliQuest sent a follow-up postcard if the diskette was not returned within about two weeks.

A pilot test showed that the questionnaire required between 20 and 30 minutes to complete.

In wave I, an experiment was conducted to see if an offer to participate in a drawing would increase response rate. Half the respondents were told that if they returned the completed questionnaire, their names would be entered in a drawing for a StyleWriter (ink jet) printer. The other half received no offer. The offer had no influence on response rate.

The response rate in wave I, after six weeks in the field, was 70%. In wave II, data collection was cut off after three weeks, with a 57% response rate. Needless to say, Apple was very happy with the response.

As a safety check, a follow-up phone survey with non-respondents was conducted to see if they differed from the respondents in any meaningful way. They did not. Further, with such a high response rate, it is unlikely that non-respondents' answers would have moved sample means.

VERBATIMS FROM OPEN-ENDED QUESTIONS

Another benefit of DBM surveys mentioned previously is the potential quality of verbatim answers to open-ended questions. Apple believes DBM is the ideal medium for gathering this kind of data from their customers for three reasons:

- No interviewer bias
- No interviewer abbreviation or paraphrasing
- More efficient for both the respondent and the researcher than pencil-and-paper

The first two points are self-evident. Disk-by-mail verbatims are more efficient because no transcription is necessary and no errors can be introduced during data entry. Of course, this is true for all data gathered disk-by-mail. Apple also believes that entering an answer on a computer may be easier for many respondents because they can edit their work.

In the last wave, respondents were asked several open-ended questions, two of which were, "Why did you buy this Macintosh," and (at the end of the survey) "What else would you like Apple to know?" In response to these two questions alone, 1,100 respondents typed in over 70,000 words.

This volume of information would be virtually impossible (at reasonable cost) to deal with if gathered through a paper-and-pencil survey. Given the unstructured nature of the questions, it is probably inappropriate to code the answers anyway, since people often addressed multiple topics in one answer.

USES OF THE INFORMATION

Apple takes two approaches to disseminating information from this rich, yet potentially overwhelming source of open-end and survey data:

1. **File-server access**. Open-end data files can be accessed by everyone at Apple. Interested employees, from entry-level marketers and engineers to the CEO, John Scully, can access the files. They can read as much or as little as they like, but it is recommended that they use a key-word search available with any word processing package. At the beginning of each file a conventional researcher's caveat about the appropriate use of qualitative information has been inserted.

2. **Research summaries**. The research staff takes a little time each week to explore the data and develop brief summaries explaining patterns or hypotheses that are developed. It provides an opportunity to show off the abilities of the research staff to speculate about forces at play among Apple's customers. The summaries are distributed through Apple's internal electronic mail (E-mail) system.

Though the open-end data are qualitative, they share some features of quantitative data. Since the sample is random and the measurement device is uniform for all respondents, it is reasonable to make certain projections to the population of recent Macintosh buyers. For example, counting the number of occurrences of the word "compatibility" gives an indication of the salience of this issue compared to other like issues. However, when there are unanticipated synonyms for a particular topic, one has to be careful about making such projections.

BEST PRACTICES FOR DISK-BY-MAIL SURVEYS

In designing and executing DBM surveys, each phase should be approached in a conscious, professional manner to minimize biases and maximize response rate. Some of these approaches include:

1. **Research sample**. Prior to finalizing DBM as the data collection methodology, it is necessary to analyze the sample for the study, ensuring that no biases are being introduced by using a disk-based survey methodology. Additionally, the following points should be considered as discussed earlier:

- Determine whether a pre-screen interview is necessary to identify the correct respondent, the respondent's market segment or quota group, and access to a PC
- Determine type of computer available (Macintosh vs. IBM-compatible)
- Pre-notify of approaching survey to increase response rate

2. **Questionnaire design**. As discussed earlier, the decision to use DBM as a data collection methodology may be driven by objectives of the research which translate to long or complex surveys. Once the decision has been made to utilize DBM, a proven approach is to:

- Develop the questionnaire on paper, as usual, to provide an easy form of communication between the client and researcher, and to create the questionnaire text which can be imported for use in the computer-aided interviewing package
- Finalize question types, question order, respondent instructions, and skip patterns before programming on disk
- Where possible, pre-test the survey on paper prior to programming, and then again on disk once it has been programmed

3. **Survey disk**. 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 for DBM surveys:

- The layout of questions should be consistent, professional, and non-distracting from the content of the questions
- Use appealing, easy-to-use software to collect the data, such as Sawtooth Software's new Ci3 System
- Include adequate instructions in the cover letter, on the diskette sleeve, or the diskette label for the respondent to start and stop the survey, if possible, and to provide accurate responses for doing so
- When appropriate, graphics on the diskette label, such as a sponsor logo, will enhance response rate
- Pre-test the survey on disk prior to fielding to ensure that all instructions are clear, questions are interpretable, and the intent of questions is clear to the respondent

4. **Disk duplication and serialization.** The best laid plans can fail in implementation. A critical component of DBM surveys is the duplication of the survey instrument, and the serialization of the disks with unique respondent numbers. To avoid potential problems during this phase, the following procedures are recommended for Ci2 (Ci2 System, by Sawtooth Software), and Ci3 surveys:

- After duplication and serialization, randomly select 5% to 7% of the survey disks and confirm:
  - The correct files are present, by checking the disk directory
  - The "numstart" file has the correct respondent number and is accurately set to deliver one or multiple survey modules
  - No detectable viruses are present on the disk

It is recommended that procedures be set up and that regular personnel be trained on quality disk duplication and serialization. When outsourcing this component of the research study, as with other phases, it is best to establish a relationship with a regular supplier whose processes can be incorporated into the planning of the project and whose quality control can be assured.

5. **Survey package.** As discussed previously, the appearance of the survey package impacts the response rate. The packaging also provides critical instructions to the respondents, communicates the incentive being offered, and protects the lawful use of the survey software. To most effectively address these issues, the following approaches are recommended:

- Create a better-looking package to get a better response
- Include instructions for operating the disk-based survey in the cover letter, on the diskette sleeve, and on the diskette label
- Communicate the benefit(s) to the respondent for participating in the study
- Decide what, if any, an appropriate incentive should be, and present it accordingly
- Include Sawtooth Software (or other appropriate) copyright on diskette label
- Use a fast method or First Class mail (stamping "First Class" on the envelope), using stamps when possible, to help draw attention to the packaging

Mail-out packages should consist of:

- Personalized cover letter on letterhead signed by the president of research firm or sponsor communicating the benefit(s) to the respondent for participating in the study
- Incentive/description of incentive
- Printed supplementary materials such as a glossary of terms (see the Appendix for an example) or concept illustrations

- Survey disks (both 3.5" and 5.25" for IBM-compatible machines)

- Postage-paid return disk mailer

6. **Fielding**. Thus far, this paper has discussed the preparation for, and execution of, the mail-out. Additionally, some overall guidelines need to be kept in mind when designing a DBM survey:

- Provide an 800 number for respondents to call toll free with questions about the diskette-based survey, since technical support is an issue unique to DBM surveys

- While responses to IntelliQuest surveys indicate that the majority of responses are returned in the first three weeks, allow adequate time in the field (5 weeks or more unless sent by a fast method such as Federal Express or USPS Priority Mail), since the return rate does not decline significantly until approximately the sixth week

7. **International.** There are many issues which are unique to international studies, in addition to those previously presented in this document:

- Questionnaires should be translated to the language of the target country, and then reverse translated by a different party to confirm that it has been correctly translated

- Questionnaires should be reviewed by someone familiar with the customs and peculiarities of the country, as well as with the product category (Sawtooth Software, 1991)

- Legal requirements should be verified regarding obtaining mailing lists, collecting certain types of information (for example, demographics), and transmitting data to companies outside the country

- Use of a local client office or international mailing house such as TNT Express Worldwide will facilitate distribution and collection of surveys in other countries, as well as assist with local customs

- When possible, provide respondents with a local number to call if they encounter problems with the survey disk

- Incentives should be appropriate and legal for each country

- Expect projects to run longer than planned

DBM surveys, as some other data collection methodologies, are impractical for some countries, but where they can be utilized they can be a benefit by providing consistent data collection in all regions.

MULTIMEDIA AND THE FUTURE OF DISK-BY-MAIL SURVEYS

Survey software and DBM survey techniques are improving every year. At the 1987 Sawtooth conference, Sawtooth set up a demonstration of a product called Interactive Video. It used a PC, a TV monitor, and a VCR to gather reaction to video stimuli. By selecting items on the PC's menu screen, the respondent could view any of a series of brief video segments in any order — a very powerful tool for measuring response to stimuli.

Unfortunately, this solution was too cumbersome and perhaps too expensive for most research applications. Sawtooth discontinued the product. Current multimedia (MM) systems can perform the same function with much less hardware and for less cost.

At present, awareness of multimedia computing is very low, and comprehension of what it is, what it's for, and what it's benefits are much lower.

In a nutshell, MM is a way to convey concepts using sound and full-motion (digitized) video alone, or as part of a presentation, or as part of a document. That is really all there is to it. Much like the original personal computers, it will be some time before people figure out how they can benefit from it.

MM will be, without doubt, a powerful tool for disk-based research.

MM technology is available now on Macintosh and some IBM-compatibles. On a Macintosh, all that is needed is modification to current disk-based survey applications. The modification is minor, and a standard Mac hardware configuration (CPU/monitor/hard disk drive) will run multimedia "movies." To convert available video from analog videotape to bits & bytes on a hard drive, you'll need to buy a $450 add-in card (for Macintosh). You don't need the card to play movies, only to convert them.

The application of MM to disk-based research is limited only by one's imagination. Here are examples of possible applications:

Advertising Research
Design monadic ad tests by randomly inserting alternative executions among a collection of competitive and non-competitive ads. The order of all ads in the "clutter reel" can be fixed or randomized for every respondent. Do this for *any* medium, TV, print, radio, billboard, or direct mail.

New Product Research
In many industries, the high cost of developing alternative prototypes prohibits testing in the market. Mock-ups can be developed for smaller scale testing, but they are often too flimsy to allow much hands-on testing by respondents. On the other hand, measuring new product ideas with concept statements or story boards is too unreliable.

A reasonable compromise could be a movie of a prototype in use by an end-user, or an animation of the product concept when even a prototype is too expensive to develop. This would be a powerful application for consumer durable industries.

Packaging and Merchandising Research

Test alternative packaging, on-shelf location, or point-of-sale materials by showing a picture of a store shelf and instructing people to point and click the product(s) they're interested in.

Even better, create a "virtual" store by videotaping a walk through the aisles and splicing in different shelf configurations and so forth for random exposure to respondents. As the respondent watches the movie, she can "stop walking" at anytime and scrutinize a product on a shelf, turn around and go back, turn left or right, or whatever. In short, she can simulate the entire shopping experience. This is because the digitized frames in a multimedia movie can be accessed "randomly" as opposed to serially, as they are on videotape.

Conclusion

Disk-based research application developers can exploit the multimedia technology available today. Compared with technology available five years ago, the solutions are easy-to-use, inexpensive, and not difficult to set up. Moreover, MM solutions run on standard, non-specialized equipment that can be used for all the other personal computing solutions a research firm needs.

If we market researchers can grasp this technology, we can reap a qualitative improvement in the power of disk-based research, an already powerful medium.

SYNOPSIS

Disk-based surveys are an additional tool which offer unique advantages in certain situations. However, they will not replace telephone or paper-by-mail methodologies for all studies. When conditions are right for using them, DBM surveys can be cost-effective and accurate.

REFERENCES

Alreck, Pamela and Robert B. Settle (1985). *The Survey Research Handbook.* Homewood, Illinois: Richard D. Irwin, Inc.

Bahner, Lesley (1987). "Long Self-Administered Questionnaires." *Sawtooth Software Conference Proceedings*, 11-21.

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.

Joselyn, Robert W. (1977). *Designing the Marketing Research Project.* New York, New York: Petrocelli/Charter.

Kress, George (1988). *Marketing Research*, 3rd Edition. Englewood Cliffs, New Jersey: Prentice Hall.

Lehman, Donald R. (1979). *Marketing Research and Analysis.* Homewood, Illinois: Richard D. Irwin, Inc.

Morrison, Richena (1988). "Disks-By-Mail," *Sawtooth Software Conference Proceedings.* 375-381.

Peterson, Robert A. (1982). *Marketing Research.* Plano, Texas: Business Publications, Inc.

Peterson, Robert A., Gerald Albaum, and Roger A. Kerin (1989). "A Note on Alternative Contact Strategies in Mail Surveys." *Journal of the Market Research Society*, Volume 31, Number 3.

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

Rossi, Peter H., James D. Wright, and Andy B. Anderson (1983). *Handbook of Survey Research.* New York, New York: Academic Press.

Sawtooth Software (1991). "International Interviewing." *Sawtooth News*, Volume 7, Number 2.

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

Key Networking Terms

APPENDIX

Example Glossary of Terms

8-BIT, 16-BIT, 32-BIT
The size of the data bus between the computer memory and the adapter card. With an 8-bit card, a maximum of 8-bits of data can be transferred at one time across the bus. If the bus clock rates are equal, a 16-bit bus transfers data twice as fast as an 8-bit bus, and a 32-bit bus transfers data twice as fast as a 16-bit bus. (see bit specifications)

10BASE-T
IEEE standard for 10 Mbps 802.3 (Ethernet) local area networks running over unshielded twisted pair rather than coaxial cable.

AUI
Autonomous Unit Interface or Attachment Unit Interface: the 15 pin D type connector used to connect the computer to an external Ethernet transceiver.

ACCESS METHODS
Techniques and rules for figuring which communications devices --e.g. computers-- will be the next to use a shared transmission medium. Access method is one of the main methods used to distinguish between LAN hardware. Examples of access methods are token passing (Arcnet, FDDI, and Token Ring) and Carrier Sense Multiple Access with Collision Detection (CSMA/CD) (Ethernet).

ADAPTER (OR NETWORK INTERFACE CARD)
A card that connects a workstation to a network. Usually it fits into one of the expansion slots inside a personal computer. It works with the network software and computer operating system to transmit and receive messages on the network.

ADAPTER STATISTICS
Information regarding adapter transmissions including packet size, number of packets received/sent, collisions, and re-transmissions.

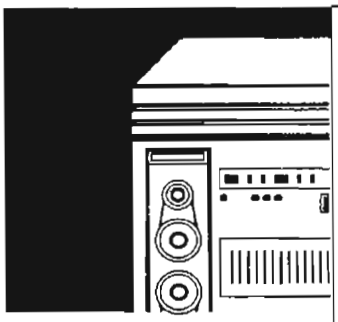
BNC
A commonly used connector for coaxial cable. The plug looks like a cylinder with two short pins on the outer edge on opposite sides. After the plug is inserted, the socket is turned, causing the pins to tighten the plug within it.

BOOT ROM (OR DISKLESS BOOTING)
A read-only memory chip (usually an option on an adapter) which allows the PC or diskless workstation to boot from the file server on the LAN.

BRIDGE
A device that connects two networks on the same type together; in contrast with a gateway, which connects two different types of networks. See router and brouter.

BROUTER
A communications device that performs functions of both a bridge and a router. Like a bridge, the brouter functions at the data link level (layer 2) and remains independent of higher protocols, but like a router, it manages multiple lines and routes messages accordingly.

@INTELLIQUEST
1250 Capital of Texas Hwy. S.
Building Two, Suite 250
Austin, Texas 78746



BUS MASTERING
A bus design that allows add-in boards to process independently of the CPU and to be able to access the computer's memory and peripherals on their own.

BUS NETWORK
All communications devices share a common path. Typically in a bus network, a "conversation" from each device is sampled quickly and interleaved using time division multiplexing. Bus networks are very highspeed -- millions of bits per second -- forms of transmission (e.g. on a local area network) and switching. They often form the major switching and transmission backbone of a modern PBX. The printed circuit cards which connect to each trunk and each line are plugged into the PBX's high-speed "backbone" --i.e. the bus network.

BUS WIDTH
Size of the computer or adapter card's data bus (8-, 16-, or 32-bit).

CLIENT
A node on a local area network. It utilizes the file server as a remote hard disk (or application server for databases) and shares remote peripherals (printers/modems, etc.) with other clients or nodes on the network.

COAXIAL CABLE
A high-capacity cable used in communications and video, commonly called coax. It contains an insulated solid or stranded wire that is surrounded by a solid or braided metallic shield, which is wrapped in an external cover. Teflon coating is optional for fire safety. Coax provides a much higher bandwidth than twisted wire pair.

CONCENTRATOR (OR HUB)
A device that joins several communications channels into a single one. A concentrator is similar to a multiplexor, except that it does not spread the signals back out again on the other end. The receiving computer performs that function. Used on local area networks to establish a star topology as with 10BASE-T Ethernet or to extend total network distances by repeating/amplifying signals in a bus topology.

DMA DATA TRANSFER METHOD
Method of data transfer between Ethernet adapter card and host PC system memory. The PC system DMA (direct memory access) chip interrupts the PC CPU to move data from the Ethernet card to PC system memory.

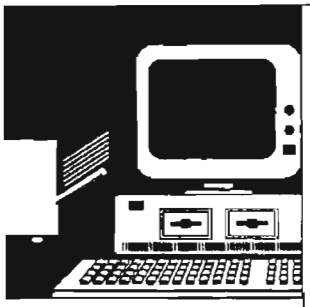
DATA THROUGHPUT
Rate at which data is transferred through the network's media. Ethernet transmission is 10 Megabits per second; FDDI specifies a 100 Mbps data transmission rate.

DISKLESS BOOTING
On a LAN, a diskless PC runs by booting DOS from the file server. It does this via a read-only memory chip on its network interface card called a remote boot ROM. Diskless PCs appeal primarily to users interested in security.

EISA
Extended Industry Standard Architecture. EISA is the independent computer industry's alternate to IBM's PS/2 Micro-Channel data bus architecture. EISA expands the original 16-bit ISA (Industry Standard Architecture) to 32-bit offering backward compatibility. EISA is useful in computing environments where multiple high performance peripherals are operating in parallel. The intelligent bus master can share the burden on the main CPU by performing direct data transfers into and out of memory.

ETHERNET
A local area network developed by Xerox, Digital and Intel that interconnects personal computers via coaxial or twisted pair cable. It uses the CSMA/CD access method and transmits at 10 megabits per second. Over coax wire, Ethernet uses a bus topology that can connect up to 1,024 personal computers and workstations within each main branch. Ethernet has evolved into the IEEE 802.3 standard which was recently extended to include operation over unshielded twisted pair wire (10BASE-T) in a star topology.

@INTELLIQUEST
1250 Capital of Texas Hwy. S.
Building Two, Suite 250
Austin, Texas 78746



FDDI
Fiber Distributed Data Interface. FDDI is an emerging ANSI standard for a 100 Mbps fiber optic LAN. It uses a "counter-rotating" token ring topology. It is compatible with the standards for the physical layer of the OSI model.

FILE SERVER
A computer in a local area network that stores the programs and data files shared by the users on the network. Also called a network server, it acts like a remote disk drive. If the file server is dedicated to database operations, it is called a database server.

GATEWAY
A computer that connects two different types of communications networks together. It performs the protocol conversion from one network to the other. For example, a gateway could connect a personal computer LAN to a centralized mainframe network. This is in contrast with a bridge, which connects similar networks together.

HUB (OR CONCENTRATOR)
A device that joins several communications channels into a single one. A concentrator is similar to a multiplexor, except that it does not spread the signals back out again on the other end. The receiving computer performs that function. Used on local area networks to establish a star topology as with 10BASE-T Ethernet or to extend total network distances by repeating/amplifying signals in a bus topology.

ISA
(Industry Standard Architecture) The 8-bit, (PC, XT) and 16-bit (AT) buses in IBM's first personal computer series. EISA is a 32-bit extension of ISA. Contrast with Micro Channel.

MEDIA
Media is the conduit or link that carries transmissions. Transport media include copper wire, radio waves and fiber.

MICRO CHANNEL
A 32-bit bus used in high-end models of IBM's PS/2 series. It is designed for multiprocessing, which allows two or more CPU's to work in parallel within the computer at the same time. Micro Channel boards are not interchangeable with PC bus boards.

NETWORK INTERFACE CARD (OR ADAPTER)
A card that connects a workstation to a network. Usually it fits into one of the expansion slots inside a personal computer. It works with the network software and computer operating system to transmit and receive messages on the network.

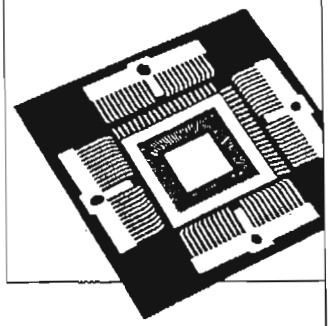
NETWORK OPERATING SYSTEM (OR NOS)
The software side of a LAN. The program that controls the operation of a network. It allows users to communicate and share files and peripherals. It provides the user interface to the LAN, and communicates with the LAN hardware or network interface card. Most NOS's have drivers to support a variety of network interface cards.

ON-BOARD CPU
Adapters sometimes integrate a CPU into their own architecture independent of the PC host CPU. This extra CPU can offload protocol processing from the host CPU thus increasing overall performance.

PROGRAMMED I/O DATA TRANSFER METHOD
Method of data transfer between Ethernet adapter and host PC system memory. The host PC CPU performs input and output to an I/O port and loops until the Ethernet packet has been received or sent.

RJ-45
The connector specified for the ends of the twisted pair wire in a 10BASE-T network.

@INTELLIQUEST
1250 Capital of Texas Hwy. S.
Building Two, Suite 250
Austin, Texas 78746



REPEATER
A device that amplifies or regenerates the data signal in order to extend the distance of the tranmission. It is available for both analog and digital signals. Repeaters are used extensively in long distance transmission systems to keep the signals from losing their strength. They are used in local area networks to extend normal distance limitations.

ROUTER
Roughly like a bridge, but more protocol-dependent. A router can usually link only LANs with the identical protocol --two Starlan LANs, two NetWare LANs, etc. Bridges are more protocol-independent, and can link more dissimilar LANs, though they can't link entirely dissimilar networks. That requires a gateway. Like bridges, routers restrict a LAN local traffic to itself, only passing data on to the bridged, or routed LAN when that data is specifically intended for it. This is in contrast to a repeater, which indiscriminately passes all data along, regardless of its destination.

SHARED MEMORY
A data transfer method used by Ethernet adapters to move data to/from the wire to the receiving/sending machine. This method employs buffer RAM on the adapter card which is assigned available address space by the host machine CPU. As such the adapter memory appears to the PC CPU to be an extension of the PC system memory.

STAR NETWORK
A communications network in which all terminals are connected to a central computer or central hub. Examples include IBM's Token Ring, AT&T's Starlan, and 10BASE-T local area networks.

TSR
Terminate and stay resident program running on the PC. Typically a small program (e.g. electronic mail receipt notification, calculator program, etc.) not requiring a lot of system memory such that other major programs can still run while the TSR is "resident."

TOKEN RING NETWORK
A local area network from IBM that uses a shielded twisted pair cable and the token passing access method transmitting at 4 or 16 Mbits per second. It uses a star topology in which all computers connect to a central wiring hub, but passes tokens to each of up to 255 stations in a sequential, ringlike sequence. Token Ring conforms to the IEEE 802.5 standard.

TOPOLOGY
In a communications network, the pattern of interconnection between nodes; for example, a bus, ring or star configuration.

TWISTED PAIR
A pair of small insulated wires that are commonly used in telephone cables. The wires are twisted around each other to minimize interference from other wires in the cable. Twisted pair wires have limited bandwidths compared to coaxial cable or optical fiber.

@INTELLIQUEST
1250 Capital of Texas Hwy. S.
Building Two, Suite 250
Austin, Texas 78746

IMPROVING RESPONSE RATES IN DISK-BY-MAIL SURVEYS

Arthur Saltzman
California State University, San Bernardino

The factors which influence respondents to complete and return disk-by-mail questionnaires are evaluated. A respondent decision tree is developed which identifies the survey process from the respondent's perspective, beginning with the receipt of the survey package and concluding with the final disposition of the disk by a respondent. The effect on response rates of the decisions made at each step in the process is reviewed. This is followed by a discussion of how the response rate is affected by many factors such as the novelty of the disk-by-mail technique, mailing variables, prenotification, length of questionnaire, incentives, and the respondent's affinity for the sponsor and the topic. Insights from published sources on conventional paper-and-pencil surveys, case studies of disk-by-mail surveys, and controlled experiments with this technique are used in the analysis.

INTRODUCTION

Response rates have always been a major problem for mail surveys. They range from under one percent for randomly selected samples of consumers to well over 50% when incentives are used with a quallfied sample. Because there is such a wide variation in response rates, the researcher is always faced with a great deal of uncertainty when having to specify how many outgoing questionnaires will yield the desired sample size. For this reason the survey research industry has proliferated research on factors which affect response rates. Early efforts focused on relatively straightforward issues such as incentives and questionnaire length. In trying to squeeze the last few respondents from a sample there have been more esoteric areas investigated such as whether the respondents are right or left handed (Cornall and McManus, 1992) and whether premiums such as tobacco pouches, lottery tickets or turkeys will be most effective (Linsky, 1975; Kanuk and Berenson, 1975).

Because each disk-by-mail (DBM) survey package is inherently more expensive than the conventional paper-and-pencil technique, we would expect that there would be substantial research available on the factors which influence DBM response rates. For several reasons this is not the case. First, a relatively small number of DBM surveys have been conducted. Since the introduction of this technology in the mid 1980's, there have probably been fewer than 500 surveys conducted using this technique. This is an estimate based on my interviews with researchers who have used the technique (See Appendix). While there are undoubtedly some who have not been counted in this summary, they are probably relatively infrequent users. As of June 1992 there were no organizations which had performed over 100 DBM studies. Four had conducted between 20 and 100. Five organizations were identified who had done 5 to 20 using the technique, and there were fewer than ten companies found who had done between 1 and 4 surveys using the DBM procedure.

Another reason for the lack of published reports on response rates is that DBM produces relatively high response rates without much fine tuning of the technique. When we can easily achieve a 30-60% response rate, there is little incentive to experiment. Also, clients are reluctant to allow experiments to be included in projects they are funding because they accurately perceive that more complex research designs are more likely to have problems during the implementation phases.

Nevertheless, several controlled experiments have been conducted. But the data from these projects have not been widely disseminated because the studies have been done for specific clients, and except for a few cases (Pilon and Craig, 1988; Higgins *et al.*, 1987; Goldstein, 1987; Zandan and Frost, 1989) the results have remained unpublished. To supplement these published reports I conducted a series of telephone interviews with other DBM practitioners who were willing to share their data on response rate experiments. Thus, the information presented in this paper includes previously published results, data from DBM researchers, and the results of my own research with the DBM technique.

I will discuss why the DBM technique is fundamentally different from the traditional paper-and-pencil mode. Then I will review each step of the survey process and analyze the impact of each step on response rates.

FUNDAMENTAL DIFFERENCES: PC ACCESS, NOVELTY, AND FOOT-IN-THE-DOOR

The evidence is clear. For mail surveys, the disk-based technique results in significantly higher response rates than the paper-and-pencil technique. But there are inherent limitations because respondents must have access to a PC. Thus, the two most popular sampling frames used by DBM researchers are subscriber lists from computer magazines and lists of MIS managers. There were no studies found which used DBM with a study of the general population.

Several DBM practitioners mentioned that the novelty of the DBM technique accounts for much of its ability to generate higher response rates. The argument is that this novelty factor will induce many to respond who would not complete and return a paper-and-pencil questionnaire. There are specific junctures in the process when this novelty issue manifests. The first is when the respondents receive the package and they are informed by a label or by the weight and feel of the envelope that there is a disk inside. This is an initial "foot-in-the-door" (Furse *et al.*, 1981) which helps insure that the package will be opened and its contents perused.

The novelty is sustained as the disk is inserted into the computer and the respondent types answers to get to the next question. Having put in the effort to get through the questionnaire, the respondent is likely to place it in the self-addressed stamped envelope that is always included in the original packet, and drop it in the mail.

RESPONDENT DECISION TREE

To facilitate the review of the factors which affect DBM response rates, the respondent decision tree is presented in Figure 1. It outlines the major steps taken by a respondent after a DBM survey package has been mailed and received. The discussion of the factors which influence response rates follows this outline.

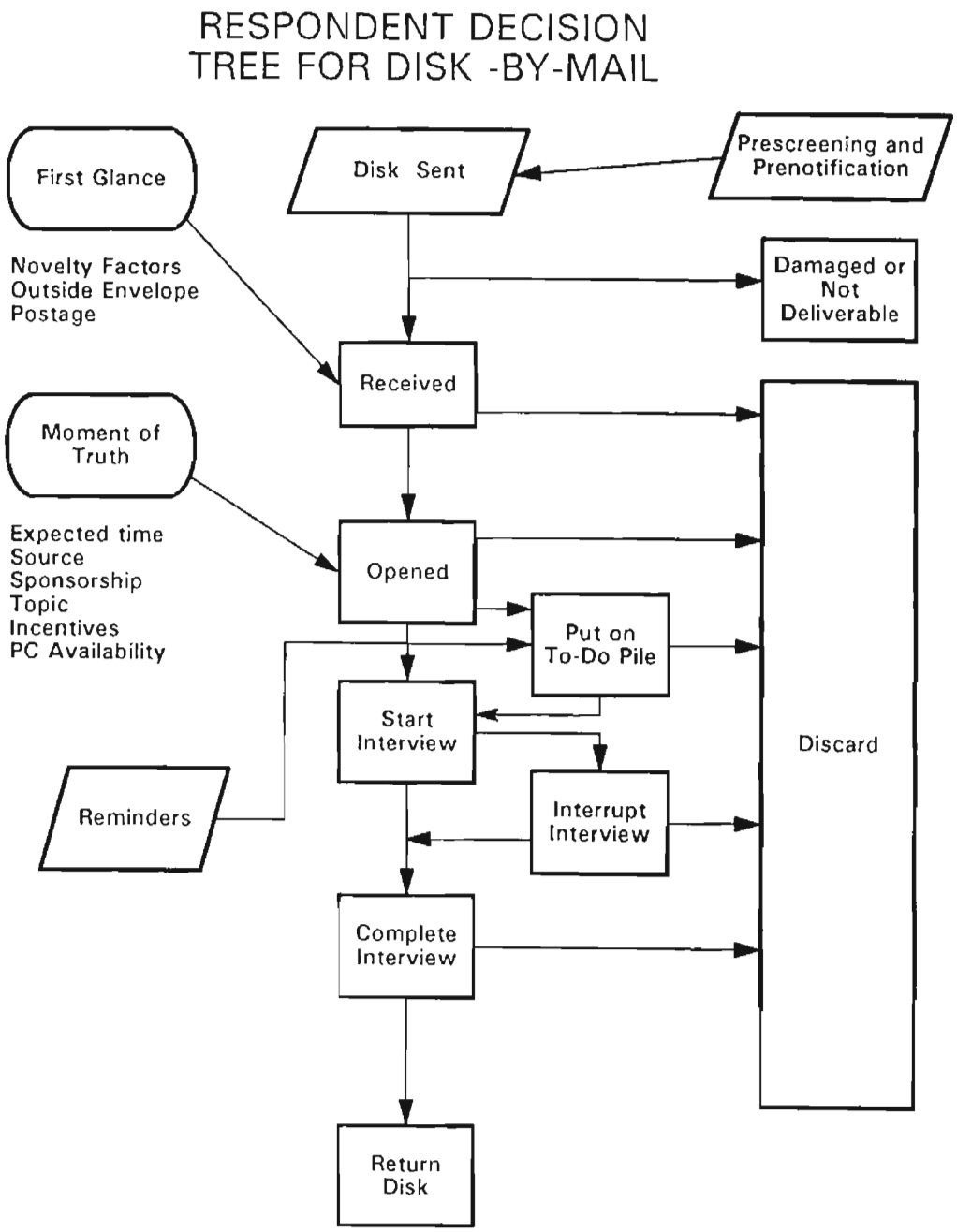
Prenotifying and Prequalifying Respondents

Conventional wisdom is that notifying persons that a survey is being sent to them will substantially enhance the likelihood that they will return the questionnaire (Jobber, 1986; Linsky; Kanuk and Berenson). DBM practitioners have done this in several cases and believe that it enhances the overall response rate. For example, Populus, Inc. reported one project in which disks were sent to a

client's employees. Management then sent E-mail messages to those in the sample to encourage them to respond. A 60% response rate was achieved in this study but since there was no control group we cannot estimate how much the prenotification helped.

Potential respondents have also been prequalified. This is especially important if there is question about whether they will have access to the computer necessary to answer the questionnaire. While I will refer to a study in which respondents were prequalified by telephone, no experimental data are available to indicate how much of an improvement can be expected from prequalifying procedures in DBM surveys.

Figure 1



Surviving the Outbound Trip

From the time the disks are mailed to the time they are physically in the possession of the intended respondents there are many opportunities for failure. However, the consensus is that the postal service is a relatively minor problem for disk-by-mail. Everyone reports some small number of mangled disks, but in only one instance was the loss greater than 1%. This occurred when a DBM practitioner, to achieve some postage savings, used a 6" x 6" mailer which was constructed of lighter card stock. Ten percent of the disks were damaged. However, others have indicated that they successfully used lighter weight mailers and have not experienced significant damage to the disks.

My own experience is that mailing 5 1/4" disks in a standard 6"x9" mailer with a "Magnetic Media" label is very reliable. I have mailed questionnaires to and from Canada and received disks from as far away as England. There were no more disk failures with respondents from other countries than with those from the United States. With the Canadian survey, there was the added logistical problem of securing Canadian stamps to be used on the return mailer. But I encountered no problems with U.S. or Canadian customs.

THE FIRST GLANCE

The first critical interaction occurs when the respondent receives the survey package. This is the first manifestation of the novelty aspect of DBM. Also relevant to the respondent at this point are several aspects of the outside envelope.

Outside Envelope and Postage Characteristics

Do these factors make a difference? The literature on generic mail surveys suggests that at least two aspects of the envelope are important: personalized addresses and teasers (Dommeyer, 1991; Carpenter, 1974/75; Khale and Sales, 1978; Neider and Sugrue, 1983; Peterson, 1975). Few of us will complete a questionnaire which is addressed to the "Occupant" and we are also not predisposed to answer one which is sent to "MIS Manager." Many authors have substantiated the importance of whether the address on the envelope is handwritten, typed, or on a computer-generated adhesive mailing label (Neider and Sugrue; Duncan, 1979; Kanuk and Berenson). Dommeyer *et al.* found that there was a substantial increase in response rates from a teaser on the envelope which referred to a monetary benefit. Little and Pressley (1980) found a barely significant increase due to using a larger envelope sizes. This result favors the DBM technique since the size of the survey disk means that the envelope will be larger than the standard #10 commercial size. Some DBM researchers use a snug 6" x 9" outside envelope while most prefer a 9" x 11" size.

In no instance was bulk postage used to mail DBM surveys, which suggests that DBM researchers universally believe in using some form of First Class postage on the outside envelope. However, there are results with generic surveys suggesting that hand-affixed First Class postage is superior to metered mail in the same amount (Jobber; Diamantopoulos *et al.*, 1991; Vocino, 1977). Most DBM suppliers seem to favor the meter technique because it can be more automated than pasting on stamps.

The postage issue for the return trip is somewhat different and has been thoroughly investigated for the paper-and-pencil technique. Business Reply versus hand-affixed stamps, commemorative versus regular stamps, and other postage experiments are all represented in the literature (Armstrong, 1990; Harris and Guffey, 1978; Wolfe and Treiman, 1979; Brook, 1978; Vocino).

The DBM practitioners used either First Class Business Reply or postage stamps in all cases, except for two instances when Federal Express overnight service was used because of extreme time pressures. One experiment was reported by Answers Research, Inc. in which Business Reply mail was compared to First Class postage. There was no difference in response rates.

THE MOMENT OF TRUTH

While the respondent makes some judgments about a survey by looking at the outside envelope, the critical determination occurs immediately after the contents of the envelope are examined.

Expected Completion Time
Recipients of a conventional survey tend to scan the pages to estimate the amount of time required to complete it. This is then balanced against the perceived benefits of answering it. But when the questionnaire is on disk, there is no size surrogate to indicate the length. Thus, many researchers administering DBM surveys include a statement in the cover letter on the expected time to complete the survey.

There are two recent experiments which look at how these statements affect response rates. My own 1991 experiment was part of a DBM survey of PC users. The sample was divided into three groups. The letter enclosed in the survey packet with the disk was the same except that the first group was told that it would take 10 to 20 minutes to complete the questionnaire; for the second group the estimate was 20 minutes. The third group was given no estimate of the time it would take to complete the survey. A similar experiment by IntelliQuest, Inc. was done during their 1992 study for a hardware manufacturer. The results of each are given in Table 1.

Table 1

EFFECT OF ESTIMATED TIME TO COMPLETE QUESTIONNAIRE ON RESPONSE RATES

ESTIMATES IN 1991 USER GROUP SURVEY BY SALTZMAN	RESPONSE RATE
10 - 20 MINUTES	62%
20 MINUTES	50%
NO ESTIMATE GIVEN	63%
ESTIMATES IN 1992 HARDWARE SURVEY BY INTELLIQUEST	**RESPONSE RATES**
15 MINUTES	44%
20 MINUTES	36%
NO ESTIMATE GIVEN	38%

In both experiments the lowest response rates occurred when respondents were told that the questionnaire would take 20 minutes to complete. When the estimate was either 15 minutes or 10 to 20 minutes, there were significantly more questionnaires returned. However, the two studies gave conflicting results about the response rates when there was no estimate given. For the user group study no time estimate was as good as the 10 to 20 minute estimate. In the hardware survey the absence of an estimate was almost as detrimental to the response rate as the 20 minute estimate.

The difference in effect may be due to the source of the cover letter and the appeals used in each study. The user group study included a letter from the group president who encouraged participation, while the corresponding letter for the hardware survey came from the marketing research firm conducting the study. Perhaps the appeal from the user group president created a higher level of trust, removing respondent apprehension about the length of the questionnaire. The overall impact on response rates of these appeals will be reviewed in the following section.

Affinity for Source

"Source" usually refers to the person or organization identified as requesting the data. Several studies have investigated the differences among university, government, and commercial sponsorship of conventional mail surveys, with mixed results on which will elicit the highest response rates (Houston and Nevin, 1977; Little and Pressley; Peterson; Jones and Linda, 1978). None of these studies explore the relationship between the source and the respondent, which I believe is a critical issue. A partial explanation of this lack of data is that when a marketing research firm is conducting research for a commercial enterprise, the respondent is seldom told who is funding the research. It is assumed that knowledge of the funding source might bias the responses. Thus, I could find no published reports concerning the effect of the respondent-sponsor relationship on response rates .

Another practical reason for the paucity of data on this issue is the difficulty of manipulating these factors in a response rate experiment. Response rate experiments are almost never the major function of a survey project. They are conducted in situations in which a researcher manages to convince the sponsor that valuable experimental data can be collected without compromising the main objectives of the survey. There are few circumstances in which the perceived sponsor could be varied in order to conduct a controlled experiment. The same reasoning would explain why there are no published studies on the effect on response rates of the affinity of the respondent toward the subject matter of the study.

To circumvent these problems, I turned to a case study approach and contacted the organizations (See Appendix) who had conducted the largest number of DBM surveys. We had extensive conversations about the most recent disk-by-mail studies. These interviews suggested that the respondent's relationship to both the sponsor and the topic did indeed have an impact on the response rate. This information is presented in Table 2. Although these case studies cannot be considered a random sample, they are representative of recent DBM research.

The data in the table strongly suggests that the respondents' willingness to complete the questionnaires is influenced by their perception of the project's sponsor.

The highest reported response rate (95%) is from a study done for a software publisher in which the identity of the software firm was disclosed. The topic of the questionnaire was a shareware product

for which registration and payment of a registration fee is voluntary. Shareware programs are freely distributed on disks and downloaded from bulletin boards at no charge. Only after users decide to continue using the program are they expected to pay a registration fee to the author. Since the sample was drawn from those who had registered their copy, the affinity of the respondents for the author of the software explains the extremely high response rate.

The other studies, however, also support the contention that the relationship of the respondent to the sponsor has a definite impact on response rate. While each of the studies was conducted by a marketing research organization, those in which the sponsors' names were disclosed had the highest response rates. Either the cover letter came from the project sponsor, or the cover letter from the researcher identified the sponsor.

The table shows one exception to this finding, where the sponsor was not identified but the response rate was a high 78%. But in this survey respondents were prequalified by telephone: only those who owned the product and had a PC available for answering the DBM survey were included. And, most important, only those who agreed to participate received the questionnaire packet. Thus, the reported 78% response rate does not include the respondents who did not pass the telephone screening and therefore is not comparable to the other case studies.

Table 2

REPRESENTATIVE DISK-BY-MAIL SURVEY CHARACTERISTICS AND RESPONSE RATES

SURVEY TOPIC	SPONSOR INDICATED	TYPE SPONSOR	RESPONDENT TYPE	INCENTIVE	LENGTH GIVEN IN LETTER	PRE-SCRNG	RSP RTE (%)
PC product features	Yes	Software publisher	Product user	No	Yes/20 min	No	95
PC product features	No	Hardware manufacturer	Product user	No	Yes/5 min	Yes *	78
Corporate image and products	Yes(for group) /No (for manftr)	PC User group/ hardware manufacturer	User group members	Yes 386/33 PC by raffle	Yes/15-20	No	69
PC product features	Yes	Computer mag. subscribers	Computer Mag.	No	No	No	68
Demographics & PC use	Yes	PC User Groups	User group members	Yes Software by raffle	Yes/20 min	No	63
Program eval.	Yes	University	Exec. MBA grad.	Report	Yes/25 min	No	62
EDP Standards	Yes	EDP Industry managers Group	MIS/EDP	$1	Yes/15-20 min	Yes **	55
Office equipment	No	Office Equip. manftr.	Equip purchasers	$1 + small lottery prize	Yes/20-30 min	Yes **	52
Not given	No	Not given	Sales leads	$1 + large lottery prize	No	No	52
Industry stats	Yes	Software Trade Association	CFO's & CEO's	Report	Yes/35-40 min	No	50
Program evaluation	Yes	University	Evening MBA grads	Report	Yes/25 min	No	46
Subscriber study (& ACA)	Yes	Mainframe computer mag.	Subscribers	$2 bill	Yes/20 min	No	40
Office equipment	No	Office Equip. manufacturer	Home based & small businesses	$2 bill	Yes/15-20 min	No	40
PC product features (& ACA)	No	Hardware manufacturer	MIS directors	$2 bill	Yes/15-20	No	36
FAX features	No	Hardware manufacturer	Computer Mag. subscribers	No	Yes/10 min	No	23

* Phoned to determine if they owned product and would participate in survey.
** Pre-notified by phone that survey was coming.

Affinity toward the sponsoring organization can also explain why PC user groups were so responsive. In each case the president of the group wrote a letter on the group's letterhead that pointed out how the group would benefit from the survey, and requested that the members complete the questionnaire. The affinity dimension is also relevant to the 55% and 50% achieved by the EDP Industry Group and the Software Trade Association. In both cases the respondents were associated with the sponsoring group or closely identified with the group's objectives.

Respondents Interest in the Survey Topic

One could anticipate that the respondents' interest in the topic of the survey would influence the response rates. But again there was no published research on this issue. This is not surprising, for a controlled experiment in which only the topic varied would require a single sponsor who is interested in collecting data on two different subjects. Several researchers I interviewed suggested that affinity toward the topic was an important factor in response rates. That is why most studies try to draw their sample from a population that is likely to be involved in the subject matter of the study. The lowest response rate in Table 2 illustrates this point. The subject was fax machines and the consultant used lists of computer magazine subscribers for the sample. The respondents' of interest in the topic was evident from the low response rate of 23%, and also from the open-ended comments from many who did respond to the DBM survey.

Incentives

Many researchers have written about how a variety of incentives impact the response rates of traditional mail surveys (Furse *et al.*; Jobber; Linsky; Kanuk and Berenson). Both monetary and non-monetary rewards that are included with the package induce more respondents to complete and return a questionnaire. The prepaid incentives seem to have a stronger effect than the promise of a reward for each person who responds. This is true for the general population and is also the case for surveys of industrial populations (Jobber).

A key question is whether the amount a researcher spends on incentives will be justified by the increase in quantity and quality of the responses. This is especially important with DBM surveys because the unit cost of sending a questionnaire package is significantly higher than sending a paper-and-pencil version. The cases reported in Table 2 indicate that many researchers use incentives. However, one of the leading DBM practitioners said that the disk-based technique by itself achieves such high response rates that he did not believe the incentives have impact. An experiment was conducted by including a $1 bill with 100 of the questionnaires, but these had no higher response rate than those with no incentive. However, since this market researcher's clients believe in incentives, one is always included with the questionnaires. A similar DBM experiment reported on by Morrison (1988) also concluded that a $1 pre-paid incentive had no effect. Marketing Metrics, Inc. reported an experiment with incentives in a DBM study where half of the sample were promised $5 for responding and half were promised $10. This doubling of the incentive brought the response rate from 30% to 45% and suggests that the promise of a substantial incentive (that is, something above $5) will have a major impact on the number of interviews which are completed and returned.

DBM intrinsically says to respondents that you care enough about their opinions to send them an expensive package. All DBM researchers attempt to project a high quality image with the survey. Some use bond paper for the enclosed instructions and personally sign each letter. But Trade-Off Marketing Services, Inc. conducted an experiment which demonstrated that those who received a letter with a handwritten signature were no more likely to respond to a DBM survey than those who were sent a letter with a printed signature. This experiment also found that colored disks elicited a 6% to 8% higher response rate than conventional black disks.

Reminders and Speed of Response

Often a conventional mail survey will be followed up by a reminder, especially when there is a relatively small initial response. The reminder can be sent to all recipients of the original mailing, or if the researcher can identify and keep track of the returned responses, the reminders can be sent only to those who have not returned the questionnaire after a specified period of time. Prior research

on the subject suggests that considerable increases in response rates are possible by using one or two waves of follow-up reminders (Duncan, 1979; Jobber).

Although Goldstein suggests using reminders with DBM surveys, only one of the DBM practitioners reported having actually used them to increase response rates. The primary explanation for this is the relatively high response rates which are achieved without reminders. Several com-mented that their clients want results immediately after the disks have been returned and are not willing to wait for another cycle of mailings of reminders. They also indicated that respondents seem to return DBM more rapidly than paper-and-pencil surveys. Because of the novelty factor, a respondent is much more likely to immediately complete the survey. The disk is more compelling and less likely to be put on the to-do pile. The consensus among DBM researchers was that DBM questionnaires are returned a week sooner than paper-and-pencil questionnaires, enabling the data analysis to begin more quickly. In addition, the analysis of data is inherently faster with DBM because the data entry task is eliminated for all but open-end responses. Thus, while DBM cannot compete with the rapid 1-3 day turnaround time which is possible with telephone or personal intercept survey techniques, DBM studies have yielded results to clients in 2-3 weeks, which is at least a week faster than the usual time required for conventional paper-and-pencil mail surveys.

Interrupted Interviews

Problems with re-starting interrupted disk-based interviews were mentioned by several researchers. Depending on the length of the questionnaire, there may be a significant number of respondents who would like to pause in the middle of an interview, turn off their computers or use them for some other task, and then resume the interview at a later time. There are various techniques for handling this situation, such as including instructions in the cover letter and the use of batch files, but almost all who use the DBM technique think it is essential to provide a toll-free number for respondents to call about this and other problems that may arise. In my own experience with DBM, we received calls from about 2% of the respondents and over 90% of them were to ask how to restart an interrupted interview. The remainder were either about a damaged disk, or requests for a 3 1/2 inch disk since we had only been including the 5 1/4 inch size. Some researchers include disks of both sizes to avoid this problem.

CONCLUSION

There are fundamental differences between the response rates achieved with DBM and conventional paper-and-pencil mail surveys. While paper-and-pencil surveys are typically returned by 1% to 50% of those who receive one in the mail, the corresponding response rates for DBM is 25% to 70%. A major part of this difference is because of the novelty of the DBM technique. But this paper presents data which indicate that several other factors can improve DBM response rates. Substantial prepaid monetary incentives can improve the response, as can aspects of the survey package. DBM respondents are also sensitive to the perceived length of a questionnaire. Case studies suggests that respondents are more likely to complete a survey when they are told who the sponsor is. The response rates are even higher when respondents have an affinity for the survey sponsors and are interested in the subject of the survey.

While there is a large body of literature on response rates for paper-and-pencil surveys, there have been relatively few controlled experiments performed on what affects DBM response rates. Since we cannot always expect results to be transferrable, there is a need for researchers and sponsors to do more experimenting with variables which can affect disk-by-mail response rates.

APPENDIX

Marketing Research Organizations Interviewed

Answers Research, Inc.; Solana Beach, CA — Albert Fitzgerald
Duke University; Durham, NC — Joel Huber
IntelliQuest, Inc.; Austin, TX — Karlan Witt
MACS, Inc.; Seattle, WA — Rebecca Elmore-Yalch
Marketing Metrics, Inc.; Paramus, NJ — Terry Vavra
Morrison & Morrison, Ltd; Louisville, KY — Richena Morrison
Populus, Inc.; Boise, ID — Lesley Bahner
Trade-Off Marketing Services, Inc.; Encino, CA — Harris Goldstein

REFERENCES

Armstrong, J. S. (1990). "Class of Mail Does Affect Response Rates to Mailed Questionnaires — Evidence from Meta-Analysis." *Journal of the Market Research Society,* 32 (July), 469-71.

Brook, L. L. (1978). "The Effect of Different Postage Combinations on Response Levels and Speed of Reply." *Journal of the Market Research Society,* 20, 238-44.

Carpenter, E. H. (1974/75). "Personalizing Mail Surveys — a Replication and Reassessment." *Public Opinion Quarterly,* 38, 614-20.

Childers, Terry, William M. Pride, and O. C. Ferrell (1980). "A Reassessment of the Effects of Appeals on Responses to Mail Surveys." *Journal of Marketing Research,* 17 (August 1980), 365-70.

Cornell, Elizabeth and I. C. McManus (1992). "Differential Survey Response Rates in Right- and Left-Handers." *British Journal of Psychology,* 83, 39-43.

Diamantopoulos, A., B. B. Schlegelmilch, and L. Webb (1991). "Factors Affecting Industrial Mail Response Rates." *Industrial Marketing Management,* 20 (November), 327-39.

Dommeyer, Curt J., Doris Elganayan, and Cliff Umans (1991). "Increasing Mail Survey Responses with an Envelope Teaser." *Journal of the Market Research Society,* 33, 137-140.

Duncan, W. (1979). "Mail Questionnaires in Survey Research: A Review of Response Inducement Techniques." *Journal of Management,* 5, 39-55.

Furse, David H., David W. Stewart, and David L. Rados (1981). "Effects of Foot-in-the-Door, Cash Incentives, and Followups on Survey Response." *Journal of Marketing Research,* 18 (November) 473-8.

Goldstein, Harris (1987). "Computer Surveys by Mail." *Sawtooth Software Conference Proceedings,* 55-59.

Harris, James R. and Hugh J. Guffey, Jr. (1978). "Questionnaire Returns: Stamps Versus Business Reply Envelopes Revisited." *Journal of Marketing Research,* 15 (May 1978), 290-3.

Higgins, C. A., T. P. Dimnik, and H. P. Greenwood (1987). "The DISKQ Survey Method." *Journal of the Market Research Society,* 29 (1987).

Houston, M. J. and J. R. Nevin (1977). "The Effects of Source and Appeal on Mail Survey Response Patterns." *Journal of Marketing Research,* 14, 374-8.

Jobber, D. (1986). "Improving Response in Industrial Mail Surveys." *Industrial Marketing Management,* 15, 183-95.

Jones, W. H. and G. Linda (1978). "Multiple Criteria Effects in a Mail Survey Experiment." *Journal of Marketing Research,* 15 (1978), 280-4.

Kahle, L. R. and B. D. Sales (1978). "Personalization of the Outside Envelope in Mail Surveys." *Public Opinion Quarterly,* 42, 547-50.

Kanuk, Leslie and Conrad Berenson (1975). "Mail Survey and Response Rates: A Literature Review." *Journal of Marketing Research,* 12 (November), 440-53.

Linsky, Arnold (1975). "Stimulating Responses to Mailed Questionnaires: A Review." *Public Opinion Quarterly,* 39 (Spring), 82-101.

Little, T. E. and M. M. Pressley (1980). "A Multi-factor Experiment on the Generalizability of Direct Mail Advertising Response Techniques to Mail Survey Design." *Journal of the Academy of Marketing Science,* 8, 390-404.

Morrison, Richena (1988). "Disk-By-Mail." *Sawtooth Software Conference Proceedings,* 375-81.

Neider, L. L. and P. K. Sugrue (1983). "Addressing Procedures as a Mail Survey Response Inducement Technique." *Journal of the Academy of Marketing Science,* 11, 455-60.

Peterson, R. A. (1975). "An Experimental Investigation of Mail Survey Responses." *Journal of Business Research,* 3, 199-210.

Pilon, Thomas L., and Norris C. Craig (1988). "Disk-By-Mail: A New Survey Modality." *Sawtooth Software Conference Proceedings,* 387-96.

Vocino, T. (1977). "Three Variables in Stimulating Responses to Mailed Questionnaires." *Journal of Marketing.* 41, 76-7.

Wolfe, Arthur and Beatrice Treiman (1979). "Postage Types and Response Rates on Mail Surveys." *Journal of Advertising Research,* 19 (February), 43-8.

Zandan, Peter, and Lucy Frost (1989). "Customer Satisfaction Research Using Disk-By-Mail." *Sawtooth Software Conference Proceedings, Volume 1* (June), 5-17.

Comment on Saltzman

Steve Bernstein
Apple Computer

Saltzman's paper is a solid attempt to uncover the elements driving response rates in DBM surveys.

- The decision tree he presented well describes the decision process.

- The next step would be to ascribe weights to the various actions the researcher can take according to their impact on response rate.

- In addition, "expected value" response rates could be calculated based on combinations of actions on different sample types to guide researchers in designing DBM surveys.

Unfortunately, he is handicapped by a scarcity of data available on the subject. Although the data to support this probably won't be available for quite some time, this sort of analysis would be a real contribution to market research practitioners.

Now, regarding some of the specifics in the paper...

Saltzman discussed the possible impact of the following items on DBM response rates:

- novelty
- prenotification & prequalification
- personal, hand-lettered addresses
- bulk rate versus First Class postage; hand-affixed stamps versus metered mail
- stated survey length
- disclosed sponsorship and affinity for the sponsor
- respondents' interest in the survey topic
- incentives
- reminders
- colored versus black diskettes

Of the items on this list, novelty, stated survey length, and the use of colored diskettes are unique to DBM research, while the rest apply to all mail studies.

Novelty probably overwhelms the other actions one might take. Most of us assume that novelty is the most important factor explaining increased response rates among DBM surveys. An important extension to this research, therefore, would be to learn how many DBM surveys one can take before the novelty wears off. Can the novelty wear off in the course of one long (boring) questionnaire, such that a second DBM survey would be welcomed with no more respondent enthusiasm than a paper-and-pencil survey? Are there principles beyond "keep questionnaire length to a minimum" to which we should adhere to maintain novelty? How can novelty be increased?

Regarding the effect of stated survey length, again Saltzman is hampered by sparse data. His findings here are mixed — there doesn't appear to be a clear pattern suggesting whether practitioners should state the length or not. Many reasons may explain the conflict in the findings, but I believe he at least needs to control for actual survey length. One hypothesis is that a significant under-estimation for an interview violates the respondent's trust and could result in more interrupted interviews than no estimate, or even, perhaps, a long interview accurately estimated.

Similarly, when comparing response rates across studies, Saltzman should account for the length of time in field for each. Practitioners often trade time for response rate when circumstances call for it.

Finally, adding provocative open-end questions at the beginning and end of the questionnaire may induce the respondent to invest emotionally in the survey, increasing the chances of returning a completed questionnaire. This hypothesis is not unique to DBM research, except that respondents may (correctly) believe that a typed "electronic" answer is easier for the researcher to read and use than a handwritten one. In the grand scheme, this effect may turn out to be minor, but probably no more so than some of the other effects Saltzman is trying to track.

CALL FOR PARTICIPATION: A DATABASE ON THE EFFECTIVENESS OF DISK-BASED SURVEYS

Arthur Saltzman, California State University
Lesley Bahner, POPULUS, Inc.
John Fiedler, POPULUS, Inc.
Joel Huber, Duke University
Karlan Witt, IntelliQuest, Inc.

During the conference session titled "Insights Into Computer Interviewing," both Karlan Witt of IntelliQuest, Inc. and Arthur Saltzman of California State University presented the results of many disk-by-mail survey projects. They used these results to draw conclusions about the best practices in disk-by-mail: those techniques and methods which results in the greatest response rates and the highest quality data.

During the discussion period that followed, several persons in the audience recommended that there would be benefits to the research community if a mechanism were developed to allow those who conduct disk-by-mail surveys to share their experiences. This sharing would lead to a database which described previous results and would provide any researcher with access to information on what has worked best in disk-by-mail surveys.

An *ad hoc* group consisting of Lesley Bahner and John Fiedler of POPULUS, Inc., Joel Huber of Duke University, Arthur Saltzman of California State University, and Karlan Witt of IntelliQuest, Inc. met during the conference to discuss this sharing concept and develop a plan to implement it.

We first decided to expand the scope of the original concept so that we would include projects using computer-assisted personal interviews (CAPI), as well as disk-by-mail (DBM) surveys. Next we developed the following three activities to implement this concept:

1. Include two standard questions in future surveys.
   We suggest that all researchers who conduct disk-based projects should start including the following two questions at the end of each of their questionnaires.

   a) Have you ever answered a survey using a computer before?

      1 Yes
      2 No

   b) Based on your experience in answering this computerized survey, would you be more or less willing to answer another survey using a computer?

      5 Much more willing
      4 Somewhat more willing
      3 About the same
      2 Somewhat less willing
      1 Much less willing

2. Collect descriptions of disk-based surveys.
   A questionnaire is being developed to gather information from researchers about their projects. For each project, information about the following will be requested:

   Type of survey (CAPI or DBM)
   Type of software used and developer (for example, Ci3, MacSurvey, APM, or ACA)
   Pre-screening
   Pre-notification
   Outbound mail method — service type and postage
   Teasers on envelope
   Personalized address
   Personalized cover letter
   Inbound mail method — service type and postage
   Survey topic
   Description of respondents
   PC availability
   Interest in topic
   Affinity for source/sponsor
   Disk size, type
   Anonymity/Confidentiality Guaranteed
   Sponsor indicated
   Type sponsor
   Incentives
   Estimates of survey length in cover letter
   Actual time (distribution)
   Reminders
      telephone call
      post card
      with new disk
   Restart option available
   Instructions given
      800 number
      how to restart
   Response rate
      # mailed; # returned undelivered;
      # returned complete and usable; # returned incomplete
   Results from the two questions described above (see point 1)
   Date of survey
   Country
   Language used
   Experiments conducted
   Name of research organization
   Estimated response rate (in-going)

The data from each project will be entered into a database which will be available to researchers upon request.

Although we will ask each contributor to identify his or her organization, such information will not be included in the database.

3. Encourage controlled experiments.
   There have been relatively few experiments conducted to investigate the effects of the different aspects and techniques in disk-based research projects which are listed above. We encourage all researchers and clients who use either CAPI or DBM to include appropriate methodological experiments in future projects and then to share their results by including them in the database.

A review of these activities was presented at the conference and there was broad support for this plan. Both POPULUS, Inc. and IntelliQuest, Inc. have agreed to participate in this effort and incorporate the two standard questions in their future projects.

The Department of Marketing of California State University, San Bernardino will be the repository of the database.

To request a copy of the questionnaire or database, contact:

Professor Arthur Saltzman
Department of Marketing
California State University
5500 University Parkway
San Bernardino, CA 92407

Telephone: 714-546-8614
Fax Number: 714-546-4607

SURVEY NON-RESPONSE AND BIAS AS A FUNCTION OF PAPER, DISK AND PHONE FORMATS

Elizabeth Smith, Ph.D.
Ruth Behringer
Aid Association for Lutherans

INTRODUCTION

The automation of data entry through the use of diskette mail surveys offers significant savings in data entry time for fixed response surveys. Diskette surveys offer an additional advantage of allowing the automated entry of open-ended responses which can be coded, or sorted and edited for printing. These advantages need to be balanced against potential differences in the effects of survey format on the rate at which respondents will return surveys, on the tendency of some respondents rather than others to return surveys, and the tendency of respondents to answer questions differently, depending on survey format.

In this paper, we report on an experiment in which sales agents received mail, telephone and disk surveys. Respondents were randomly assigned to receive mail, telephone or disk surveys. The entire field sales staff of Aid Association for Lutherans (AAL) was surveyed. The following questions will be addressed:

1) Do some survey formats produce higher response rates than others?
2) Which type of survey format do respondents say they prefer? Does the survey format affect preference?
3) Do respondents who respond to different formats differ in terms of their productivity and length of service?
4) Does the tendency of respondents to agree or disagree in global terms reflect the format in which the survey is presented?
5) Are respondents more likely to answer in a consistent manner (selecting only high or low responses) in some survey formats than in others?
6) From an administrative standpoint, when should disk versus mail or phone surveys be best used?

DESCRIPTION OF THE EXPERIMENTS

The study was undertaken because AAL desired to monitor sales agents' satisfaction with service at the service team level, using a survey of 18 questions. At AAL, groups of agencies are provided underwriting, certificate change and claims services by teams at the Home Office, so that service is more personalized. Because each service team serves between 80 and 150 sales agents, in 4-8 agencies, the entire population of 2000 agents had to be surveyed to achieve statistically reliable results. We wanted to be able to compare the advantages and disadvantages of doing the survey in different ways, because it would be done on a semi-annual basis. We also wanted to study the cost-structure of the survey.

The survey questions are stated in affirmative form: "My service team offers practical and objective suggestions." The responses are coded from 1 for "Strongly Agree" to 5 for "Strongly Disagree". They are presented in matrix form on the paper survey, as shown in Table 1.

Table 1

Matrix Survey Form

SA - Strongly Agree
A - Agree
N - Neither Agree nor Disagree
D - Disagree
SD - Strongly Disagree

My service team.....

Offers practical and objective suggestions.	SA A N D SD
Was easily accessible	SA A N D SD

We carried out the experiment in the summer of 1991. We randomly assigned 655 of the agents to receive the standardized mail survey, 549 to receive telephone calls from a telephone research facility, and 744 to receive diskette surveys. We sent different quantities because we knew the response rates would vary by method. We sent follow-up letters reminding the field agents to return the surveys, three weeks after the initial surveys were mailed. The diskette surveys were created using Ci2 software (Ci2 System by Sawtooth Software). The questions were presented screen-by-screen rather than in matrix form (see Table 2 below). A one-page instruction sheet accompanied the mailed diskette surveys. The move to diskette-based surveys was motivated by the fact that all field agents at AAL have portable computers.

Table 2

Ci2 SURVEY SCREEN

My service team offers practical and objective suggestions.

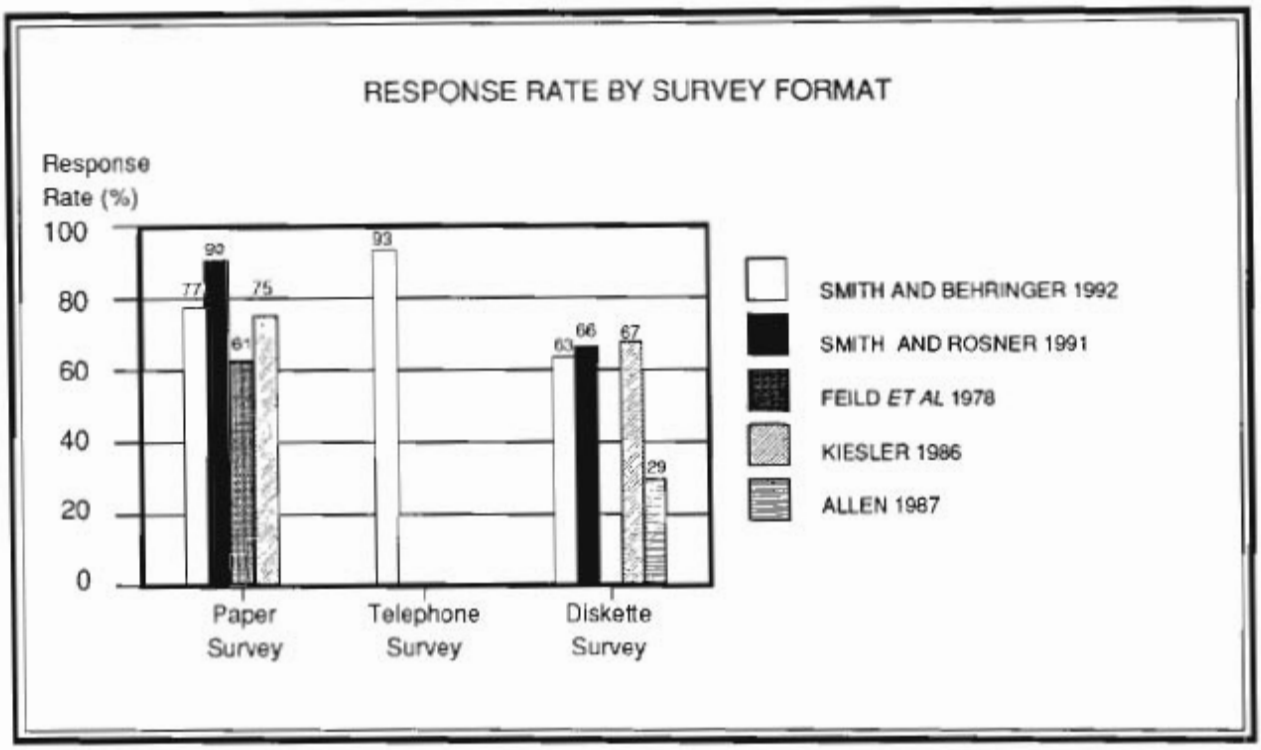
1 STRONGLY AGREE
2 AGREE
3 NEITHER AGREE NOR DISAGREE
4 DISAGREE
5 STRONGLY DISAGREE

Please type the number corresponding to your answer.

RESPONSE RATE

The highest response rate was that for telephone surveys: 93% of respondents were reached by telephone. Those refusing stated their lack of experience with Home Office Service as a reason for not participating in the survey. The lowest response rate was to the diskette survey, which was 63% in this study and 66% in an earlier study. This response rate was fairly typical of what we have experienced at AAL. The paper response rate at 77% was not as high as in previous surveys, though still above the average for surveys done with field agents in most companies. With Ci3 our response rates have been in the 75-80% range, however.

The differences in response rates correspond to the magnitudes of in differences found in Feild *et al.* (1978), Allen (1987), and by Keisler and Sproull (1986), although the Keisler and Sproull study did not involve mailed diskettes, but rather direct terminal linkages.



Why did these differences occur? In the case of telephone contact, the burden is on the interviewer to reach the respondent. The lack of anonymity on the part of the respondent also makes the respondent less likely to refuse a company study, even when done by an outside vendor. The diskette study, on the other hand, requires the most effort on the part of the respondent, since in addition to answering the questions, the respondent must read the instructions, start up the computer, be familiar enough with DOS to go to the A drive, and start up the survey. This adds about 5 minutes to the time required for the survey. In addition, respondents are not clear about how long the survey will actually take before they start the survey. A disk survey is not as easy to start and stop as a paper survey, which may be a significant factor for field agent respondents who are subject to numerous interruptions and telephone calls.

RESPONDENT PREFERENCES

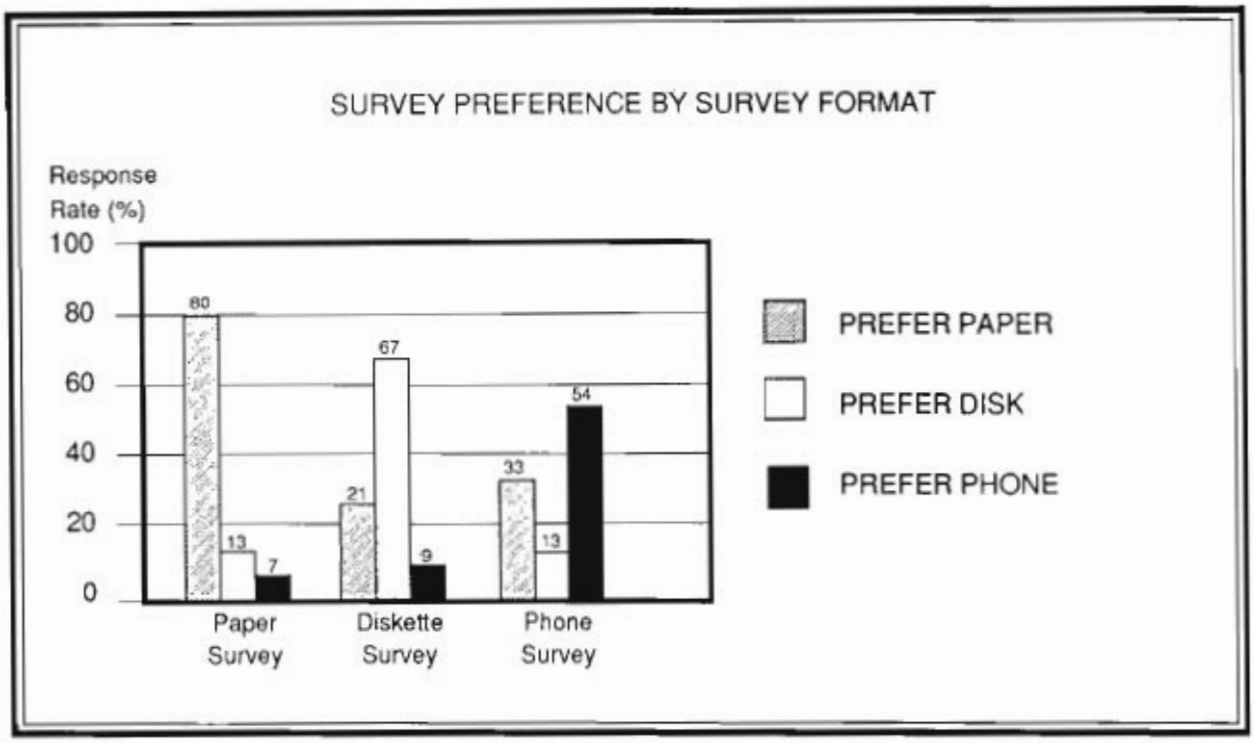
Respondent preferences for survey format were heavily influenced by the format of the survey. When respondents to the paper survey were asked what format they preferred, 80% indicated a paper survey, 13% diskette and 7% telephone. The respondents who received the diskette survey manifested a distinct preference for diskette surveys (67%) as compared to paper (24%) and telephone (9%). Finally, the telephone survey respondents indicated a preference for the telephone survey (54%), followed by paper (33%) and diskette (13%). These are large differences that may indicate two things: a higher overall comfort level with paper surveys, but perhaps also a social desirability effect: that is, a tendency to agree with whatever process is already going on. Overall, respondents preferred the paper survey.

SELF-REPORTED CHARACTERISTICS OF RESPONDENTS

1991 SALES CREDITS	PAPER	DISK	TOTAL
LOW	27%	22%	24%
LOW-MEDIUM	24%	27%	25%
MEDIUM	23%	22%	23%
MEDIUM HIGH	11%	16%	13%
HIGH	16%	13%	15%
TOTAL	100%	100%	100%
TOTAL N	483	472	955

SELF-REPORTED CHARACTERISTICS OF RESPONDENTS

We keep track of which respondents return surveys (the surveys themselves were anonymous) through the use of labeled postcards. When the respondent returns the survey, whether paper or diskette, he or she is also instructed to return the accompanying postcard separately, so that a follow-up reminder would not be sent. This let us track which respondents did and did not return surveys. An analysis of production and years of experience data revealed no statistically significant differences in the characteristics of respondents to the paper and diskette surveys. Using diskettes therefore did not produce bias in the types of respondents who answered the survey.

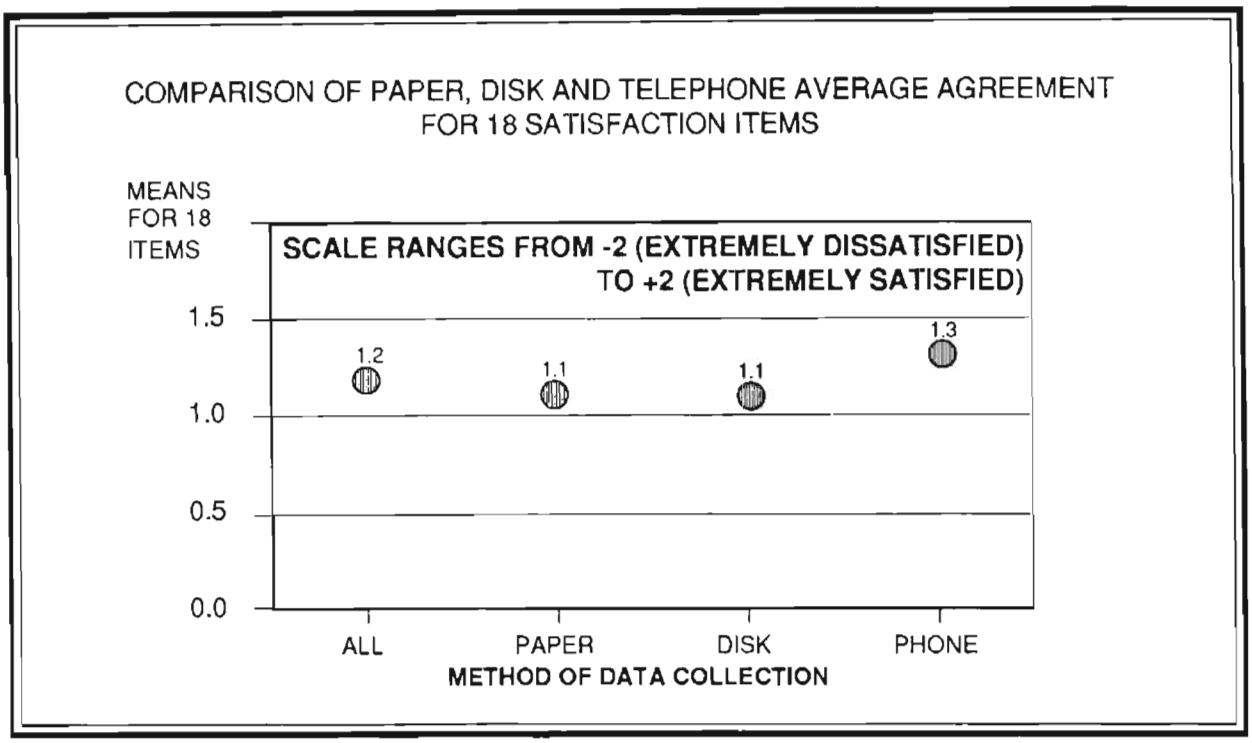


ITEM BIAS

We were concerned about whether respondents would respond differentially to the different survey formats, so that we could adjust the data for future baseline comparisons. We computed the means of the 18 items in the survey of each survey format to see if there were differences in response patterns.

An analysis of variance reveals that the telephone survey respondents had the most favorable responses, even when we adjusted for length of service (which is also related to favorableness of response). The mean response for the telephone survey was 1.3 on a scale from -2 (strongly disagree) to +2 (strongly agree). The participants in the paper and disk survey gave nearly identical mean satisfaction ratings of 1.1. The overall mean was 1.2.

Since the differences in attitude by format were nearly as great as differences among groups of agents served by different teams, it was of concern to us that we decide which ratings were more valid. We decided that the paper/disk results were probably more valid because there was anonymity for the respondents. Asking about agents' attitudes toward home office service is a somewhat sensitive form of questioning, since an agent's livelihood depends on this service.

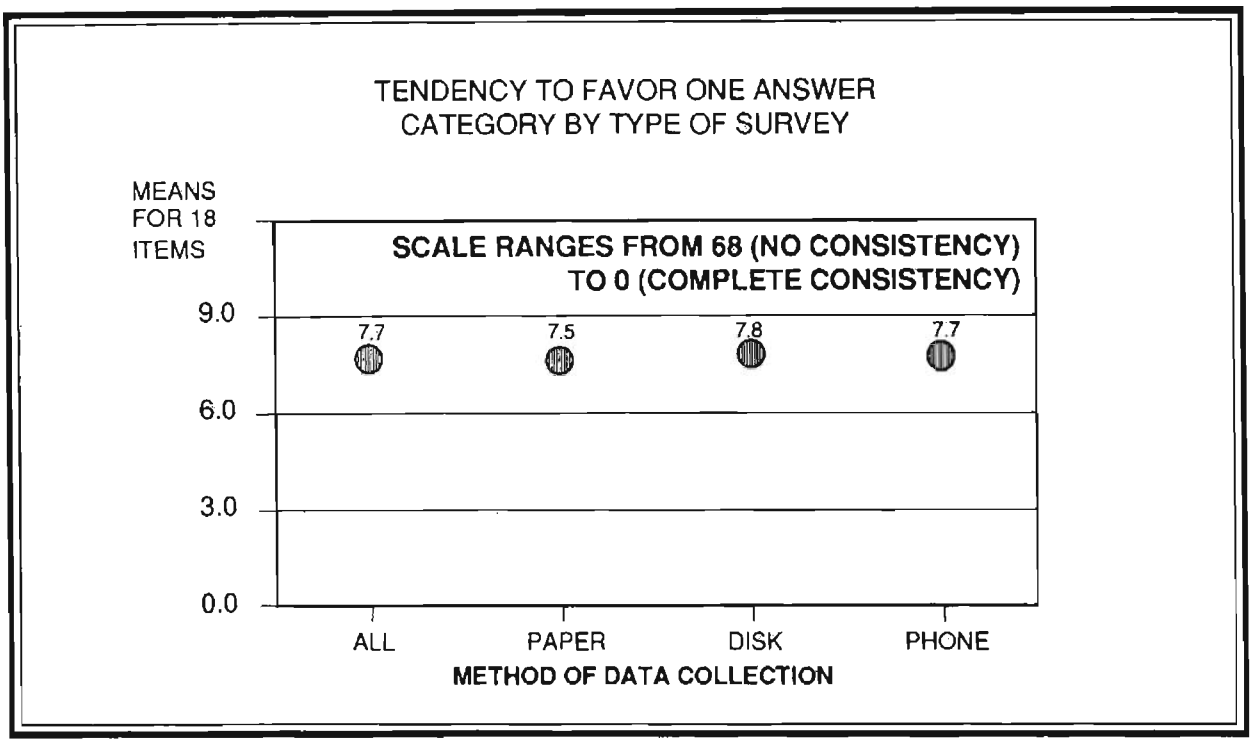


Keisler and Sproull found that responses on their electronic survey were less socially desirable than those on the paper survey. In an earlier study (Smith and Rosner, 1991), we thought it might be possible that respondents would provide more positive responses on the paper than on the diskette survey, as a result of this previous finding. This study confirmed the results of our previous study: that there were no substantial differences between paper and diskette formats in terms of item bias.

Another element of item bias that interested us was the tendency of respondents to stay with one answer category. We felt that this would be the highest in paper surveys, less in diskette surveys and least in telephone surveys. To measure this we took the differences between successive items on the survey and added them up. The smallest difference would be 0, if a respondent circled the same number twice. The largest difference would be 4, if a respondent circled a 1 and then a 5 (or vice versa). The largest sum would then be 68 (17 differences between 18 items times 4, the maximum difference). The overall sum was 7.7, which is quite low and reflects the strong convergence of opinion among field agents. The differences among formats are not statistically significant.

OTHER ADMINISTRATIVE CONSIDERATIONS

Since our response rates are fairly high — even with diskette surveys — we continue to use this format. We have found that the self-administered diskette survey is most useful when respondents all have computers, the survey is longer than 5 pages, has many open-ended responses, and there are 200 or more respondents.



Since the development time for a researcher to use the software involves about as much effort as learning a variety of other software (such as Lotus or some statistical software), it is not suitable for the casual user — a researcher who does fewer than 4 surveys a year would probably not be using his or her time effectively.

The cost factor of the survey depends on what resources are available. When there is relatively little clerical support for data entry and typing of open-ended questions and rapid data-processing is required, diskette surveys have a clear advantage. An advantage of the Ci2 survey software that we had not originally anticipated was the ability to process open-ended text responses. Since the files produced are in ascii format, they can be transmitted to mainframe files for sorting, editing and printing. The Ci2 open-ended answer files are originally in respondent order, not in question order. DOS is too restrictive an environment, allowing sorting for only about thirty respondents at time. By transferring the files to a mainframe environment, we can sort and edit the responses of several hundred respondents at a time. Management often requests to see the actual text of open-ended responses, to assist them as they strive to improve field satisfaction with existing service. This text-processing capability has made it economical to process even as a few as a 100 respondents using Ci2, as it saves days of typing.

CONCLUSIONS

Each type of survey appears to have advantages, depending upon the purpose required. For less sensitive questions in very short surveys of small numbers of respondents where fast turnaround is required, telephone surveys can provide an optimal format. For sensitive questions which can be influenced by social desirability, anonymous paper or diskette surveys are more valid.

Paper surveys are most cost-effective for shorter surveys (less than 5 pages) with mostly multiple choice items with under 200 respondents, largely because of the programming time required for a Ci2 (or Ci3 System by Sawtooth Software) survey, and the more careful pre-testing and quality control required with diskette surveys. For longer surveys, the extra respondent effort required to load the diskette is negligible compared to the effort of filling out the survey. Paper surveys are also much cheaper to duplicate and mail, particularly when they are short.

The optimal choice of survey format depends on the sensitivity of the questions, the length of the survey, the number of respondents, and the costs of mailing and clerical assistance.

LIST OF REFERENCES

Allen, David F. (1987). "Computers versus Scanners: An Experiment in Nontraditional Forms of Survey Administration." *Journal of College Student Personnel.* V28. N3. 266-73. 1987.

Feild, Hubert S., William H. Holley, and Achilles A. Armenakis (1978). "Computerized Answer Sheets: What Effects on Response to a Mail Survey?" *Educational and Psychological Measurement*, 38.

Keisler, Sara and Lee Sproull (1986). "Response Effects in Electronic Surveys." *Public Opinion Quarterly.* Vol 50: 402-413.

Ory, Jon C. and John P. Poggio (1980). "Response-Mode Variation Effects on Affective Measures." *Educational and Psychological Measurement.* V41. N3. 625-634. Fall.

Smith, Elizabeth and Lisa Rosner (1991). "Survey Nonresponse and Bias as a Function of Mark Sense and CATI Formats." unpublished paper presented at the American Statistical Assocation.

USING COMPUTER INTERVIEWING TO DEVELOP PERSONALIZED SCALES

Ira Goodman
J.M.R. Marketing Services, Inc.

INTRODUCTION

Good morning. It's so nice here in Sun Valley, Idaho. The plant life is beautiful. The mountains are majestic. The weather is seasonably warm. I mean it is warm to me. Some of you from the southern portions of the country might actually consider this weather cool. Would any of you call it hot? What about cold? I find this whole matter of weather so confusing. It's no wonder that the National Weather Service has difficulty forecasting the weather when there is such confusion just talking about it.

In a "Peanuts" cartoon I saw recently, Lucy asks Charlie Brown, "Kind of chilly out today, eh Charlie Brown?" And he replies, "It's more than just chilly...it's cold outside..." He reflects on this statement and adds, "In fact, it's more than just **cold**...it's **mitten** cold!!"

Even Lucy and Charlie Brown seem to have difficulty communicating about the weather. Lucy thinks it's cool outside while Charlie Brown calls it cold. "Mitten cold!" Why can't they agree on an appropriate label for the weather they are both experiencing at the same time in the same place? Are difficulties in communicating limited to discussions of the weather? Could these problems somehow relate to issues involved in scaling responses in survey research? I think so, and you will see how as this paper progresses.

BACKGROUND

Why do we need another type of scaling approach? We already have hedonic scales, semantic differentials, word scales, number scales, picture scales, anchored scales, nominal, ordinal, interval and ratio scales. Some of you may feel that these scales satisfy your informational needs. But others of you may have had some nagging questions, like I did for many years, about the quality of the data obtained with these scales.

Let's imagine that you are designing a questionnaire for measuring attitudes about several competing products. In setting up the questionnaire, you randomize the order in which the products are evaluated. You also randomize the order in which the attributes are presented. We have all been taught to randomize in this way to minimize order bias. When the data come back, we usually check to see if there was any order bias in the evaluations of competing products. Often we find that the order bias is there despite the randomization process.

Order bias is just one of many types of biases that creep into our data despite our best efforts to minimize bias.

Another form of bias is a halo effect. For example, we might test a product with a new improved color system. The rating of the product on the color dimension is improved over its current formulation. In addition, the new color formulation is rated more favorably for its taste, sweetness and texture which were not improved. The carry-over effect of the color on the taste, sweetness and texture attributes represents the halo effect. Why should this occur?

Another example of data bias is when respondents develop specific response patterns. Those who primarily use the positive end of the scale are "Yea-Sayers." Those who give predominantly negative responses are "Nay-Sayers."

All of these biases are normally treated as relatively minor sources of irritation in the research process. In fact, they are often simply noted and the analysis continues with the data unchanged. Sometimes we adjust the data in order to try to overcome the error. This might be done through weighting or by statistically normalizing the data by converting to "Z" scores.

There are, however, far more troublesome outcomes from the use of current rating techniques. It is seen in how much error there is in predicting the success of products introduced to the marketplace. We have all heard that most products taken into the marketplace tend to fail even after extensive attitudinal and in-market research. This was most dramatically demonstrated a few years ago when a major manufacturer sued a research supplier for ostensibly predicting greater sales potential for a new product than it actually achieved in the marketplace.

Look at the distribution of scores on a typical rating scale for popular products like candy, desserts, soda and other thirst quenching beverages and snacks. Have you ever noticed how positive the responses are to almost every concept, product and commercial tested for these products? Very often these items receive 50% or better top box scores and minimal bottom box responses. If we take these scores at face value, they imply that the items receiving such favorable scores should perform extremely well in the marketplace. Yet they often fail when introduced into the marketplace. Doesn't this raise a question about the validity of these types of measurement procedures?

Perhaps the error is not the fault of any individual or company. It may be due to the rating tools we use as an industry.

UNDERLYING ASSUMPTIONS OF TYPICAL RATING SCALES

A monadic rating scale is a measurement device. It might be considered analogous to a ruler that measures distance, or a thermometer that measures temperature. A monadic rating scale is intended to measure how favorable a person's attitude is toward an object.

These mental and physical measurement devices are superficially analogous. However, the analogy does not hold much beneath the surface.

Measurements done with physical scales are consistent across many situations and their meanings are easily understood. For example, a person in New York City can cut a piece of wood one foot long and another person in California can do the same thing. The two pieces of wood, if placed side by side, would be basically the same.

This is possible because there is a clear standard for measuring distances. The National Institute of Standards and Technology in Washington, D.C. maintains standards for units such as the inch and pound. Therefore, all measuring devices are built to match these standards. This is what makes it possible for many different people to have a similar understanding of what an "inch" means. It is also why people in many different parts of the world can measure objects to the same length specifications.

Mental measurements suffer from a lack of such standards. Think about it. When I say that the coffee that I had this morning was "very good." Can all of you agree on what I mean by "very good"? There is no standard in Washington, D.C. or anywhere else telling people when to use the descriptive phrase "very good" to express their reactions to objects.

What then is the principle underlying the construction of attitudinal scales? I believe that the predominant model goes back to some ideas of the philosopher Plato. Plato suggested that there are ideals for forms. For example, he looked around and saw many different kinds of tables. He asked why so many differently shaped objects are called tables. He reasoned that they were all approximations of an ideal for a table. He believed that this ideal exists on the ideational level and not at the sensory level. In other words, we can think of the perfect table but never experience it through our senses.

I believe that we assume that typical monadic rating scales measure how well objects are thought to approximate the ideal. It is assumed that an object rated "poor" is very distant from the ideal while one rated "excellent" is very near it.

AN ALTERNATIVE MODEL

I would like to offer an alternative framework in which to understand this rating process. Let's derive it from a real life experience involving the weather!

Imagine that you are watching the evening news and it's time for the local weather report. The meteorologist says that it will be cold tomorrow. Now, let's see how well we understand what the meteorologist is saying. I need your assistance. I am going to ask you to choose among three temperatures: 60 degrees, 40 degrees and 20 degrees. Raise your hand to indicate the temperature to which the meteorologist is referring. Do you think that when the meteorologist says it is cold the temperature is 60 degrees? Or do you believe it is 40 degrees? Or do you think it is 20 degrees? Or, did you decide to not answer?

The meteorologist continues the report by indicating that tomorrow will be the last day of winter. Do you think the temperature is 60 degrees? 40 degrees? 20 degrees? Didn't answer?

Finally, the meteorologist switches from the local weather to the national scene and indicates that it is snowing in New York. But, the meteorologist is happy to report that there in Sarasota, Florida, from where the show is broadcast, it will be bright and sunny.

Now, what do you believe was the cold temperature in Sarasota, Florida to which the meteorologist was referring? Was it 60 degrees? 40 degrees? or 20 degrees? Or, did you not answer this time?

I believe there was a shift in your responses. You probably did not select 60 degrees initially. But, by the end you probably chose 60 degrees.

What made you change your answers? Could it have been the additional information that I gave you about the time of year and the location which helped you fix an appropriate temperature to the label of cold?

This example parallels what I believe happens with attitude ratings. That is, the word cold is like the ratings "excellent," "good," "fair" or "poor" or any number of other points on attitude scales whether they use words, numbers or pictures. You can see that the word cold does not have an absolutely, clearly defined meaning. Its meaning is determined by the context in which it is used. Similarly, the rating "excellent" does not have an absolute meaning. Its meaning is also relative to the context in which it is used.

In our weather example, the meteorologist establishes the context and provides the meaning for the evaluative term "cold." Who provides the contextual meaning for the rating "excellent"? It is the person participating in the study. It is each of the 1,000 or so respondents participating in the study. That is, each respondent provides contextual meaning while answering each rating question. The context for each rating comes from the experiences of each respondent. Therefore, the context in which each rating term is used will be as unique as each individual's experiences. In fact, it is possible that one individual may change the context several times during the course of an interview. I believe that these shifting contexts are the cause of the "yea-" and "nay-say" responses as well as the order and halo biases that I mentioned earlier.

I am suggesting that people do not rate objects by comparing them to some concept of an ideal which they carry in their minds. Instead, they evaluate items versus similar things which they find more or less appealing.

I read a standard example of this thought process in the Wall Street Journal a couple of years ago and I would like to share it with you. The paper reported on a test concept by McDonald's Corporation: A 1950's style diner which the food company was testing in Hartsville, Tennessee. The reporter indicated that the residents in Hartsville "don't particularly like the hamburgers" served at the diner. The reason for this was that "everybody thought they'd be like McDonalds', but they're not." On the other hand, it was reported that the french fries fared better since they were exactly like the ones available at the nearest McDonald's.

This example highlights how the consumer does not have an abstract, ideal hamburger in mind. The respondent has a set of experiences. It is this set of experiences that are the basis for making judgments. In other words, respondents who answer research questionnaires may not be thinking in absolute, idealistic terms as suggested by Plato's original thesis and as implied by current interval type scales. Their thoughts may instead be grounded in relativistic comparisons. There are theories in physics and philosophy or religion which support this relativistic view of reality.

The Hindu religion of India has ancient scriptures called the Veda. It declares that the physical world, which we as researchers aim to measure, operates under one fundamental law of maya. This law states that the world is relative and dualistic. That is, there are always a pair of forces that are equal and opposite. This is exemplified by Newton's Law of Motion which states that "To every action there is always an equal and contrary reaction; the mutual actions of any two bodies are always equal and

oppositely directed." We see these principles in nature. Electricity, for example, is a phenomenon of repulsion and attraction; its electrons and protons are electrical opposites. The entire phenomenal world is under the inexorable sway of polarity; no law of physics, chemistry, or any other science is ever found free from inherent opposite or contrasted principles.

Einstein's Theory of Relativity proved that the velocity of light, 186,300 miles per second, is the only constant of a universe in a state of flux. All human standards of time and space depend on the sole "absolute" of light velocity. That is, time and space are not abstract and eternal but relative and finite. They derive their conditional measurement-validities only in reference to the yardstick of light velocity.

We may now have a basis for moving from the Platonic view of an absolute, idealistic standard to a theory and practices that stress the relativistic and dualistic nature of reality.

It is this type of thinking which led to the development of the "Relative Interval Scale" for use in evaluating attitudes in survey research.

THE RELATIVE INTERVAL SCALE

The foregoing statements suggest the principles necessary for measuring consumer attitudes. They include:

> Each respondent needs to make evaluations relative to a meaningful frame of reference
>
> Each respondent needs to make evaluations relative to items which represent positive and negative dualisms within that respondent's frame of reference.

As this suggests, measurement through the use of the "Relative Interval Scale" is personalized for each respondent. This makes attitude measurement a more complex task to administer than is the case for current scaling procedures. Computerized interviewing makes it possible to easily and efficiently administer this personalized interview procedure.

An example might help clarify the "Relative Interval Scale" procedure. Imagine that we would like to evaluate a new car concept.

We would begin by showing a respondent some current models.

We would ask a variety of questions to determine which of these cars are familiar to the respondent and which have the strongest and weakest appeal.

Next we would have the respondent evaluate the new car versus the one that is least appealing. This would be done by asking a simple question like: "Which of these two cars would you most want to buy?"

The respondent would then evaluate the new car versus the most appealing model in the same manner.

We would recombine the answers to these questions into a five point "Relative Interval Scale" as follows:

RELATIVE INTERVAL SCALE

Top Box	= New Car Liked Over Most Appealing Car
Second Box	= New Car Liked Same as Most Appealing Car
Middle Box	= New Car Liked Less Than Most Appealing Car But More Than Least Appealing Car
Fourth Box	= New Car Liked Same as Least Appealing Car
Bottom Box	= New Car Liked Less Than Least Appealing Car

The "Relative Interval Scale" can be used with all the attributes and rating dimensions addressed with the current scaling procedures. We can ask about things such as purchase request and overall appeal.

VALIDATION OF THE RELATIVE INTERVAL SCALE

I will present some data demonstrating the improved quality of the data obtained from the "Relative Interval Scale" versus standard scales among both adults and children.

We find that current scaling procedures tend to overstate the appeal of items and understate the amount of turn-off. This is reflected in actual market sales which tend to be less favorable than rating scales suggest they would be.

For example, a major toy manufacturer conducted studies using its standard rating methodology as well as the "Relative Interval Scale." The company's data suggested that a new toy would perform well above average for the category. The "Relative Interval Scale" suggested that the toy would have just average sales for the category. The actual sales were only average for the category as projected by the "Relative Interval Scale."

Currently used monadic scales tend to produce strong top box scores and minimal numbers in the bottom box. By comparison, the "Relative Interval Scale" produces lower top box scores and can produce greater bottom box numbers. The differences in projected sales noted earlier could be a function of basic differences in data patterns.

Here are some sample data which reflect these patterns:

DATA DISTRIBUTIONS

	CONCEPTS		PRODUCT	
	Monadic Rating	Relative Interval	Monadic Rating	Relative Interval
	%	%	%	%
Top Box	36	17	66	23
Bottom Box	3	20	1	12

The sequential monadic design is often used in market research in order to obtain ratings of two products or concepts from the same respondent. As noted earlier, attitude rating scales which are typically used today yield results that contain order biases. That is, the order in which the product was evaluated, first or second, influences the ratings. This bias generally does not occur with the "Relative Interval Scale."

The following example contains data from a recently conducted sequential monadic product test using the "Relative Interval Scale" and a typical monadic scale. It indicates that the monadic scale yielded significant second position order biases for two products. There were no significant order biases with the "Relative Interval Scale."

ORDER BIAS
= Top Two Boxes =

	MONADIC SCALE		RELATIVE INT.	
	1st	2nd	1st	2nd
	%	%	%	%
Product A	47	66**	33	35
Product B	54	54	23	37
Product C	35	55**	26	26
Product D	78	80	35	41

** = Significantly different at 90% confidence level

Children's questionnaires tend to produce biases not usually seen in adult data. One such bias is an age skew. That is, the results obtained from typical kid scales tend to be more positive among younger than older children. This bias is virtually eliminated with the "Relative Interval Scale," as you can see in this slide which presents the results of three products evaluated by a typical monadic scale and the "Relative Interval Scale."

The monadic scale resulted in two of the three products being rated significantly better by younger children. The other product had non-significant results that also favored the younger children. The "Relative Interval Scale" results indicated that one product was significantly favored by younger children. The other two products tended to be favored by the older children.

PRODUCT EVALUATION AMONG CHILDREN
= Top Box =

	MONADIC		RELATIVE INT.	
	6-8	9-12	6-8	9-12
	%	%	%	%
Product A	71	62**	21	25
Product B	61	47**	39	15**
Product C	65	61	28	33

** = Significantly different at the 90% confidence level.

MARKETING SUCCESSES

I would like to offer you two examples of the effectiveness of marketing efforts based on results from the "Relative Interval Scale." The first involves a toy that we researched in 1990, "Go Go the Walking Dog" by Hasbro-Bradley. The product rated 28% in the top box on purchase request, which is significantly higher than our norms for the category. The product was recommended for development.

We have found that there is a relationship between the strength of purchase request and actual in-market performance. A product which is rated significantly above the norms for a category tends to have strong sales in the marketplace. "Go Go" was the hottest selling toy in the category and continues to be a strong performer.

The second example has to do with an advertising campaign. We had been working with our client for about 5 years. Over the course of time, we identified a variety of opportunities which could enhance the effectiveness of the advertising. About nine months ago, the client company revamped its advertising in accordance with our recommendations.

Prior to revamping their advertising, the client's commercials evoked a top box purchase request score of 14% while the category leaders got comparable scores of 21-24%. During this time, the client was experiencing declining sales and loss of share to its key competitors.

The new ad campaign based on hypotheses suggested by our research resulted in a top box purchase request score of 22%. This score was significantly higher than the client's previous campaign and it was comparable to its key competitors.

The good news is that the client is reporting a substantial increase in sales, which may be traced to the new advertising program.

These examples suggest that the "Relative Interval Scale" has demonstrated success in screening products which perform strongly in the marketplace and in identifying commercials which motivate sales of products.

APPLICATIONS

The "Relative Interval Scale" has been used with adults and children, consumers and business people. It overcomes various data biases including those found in studies with children and across cultures in international studies. It has been used in concept, product, tracking, advertising copy, commercial and television program testing.

It has been used successfully in conjunction with various multivariate statistical approaches including multiple regression and correspondence analysis.

We have applied the relativistic philosophy to other issues and developed new approaches for assessing product diagnostics and measuring commercial effectiveness. For example, we have found that questions such as: "Is the product too sweet, just right or not sweet enough?" can produce misleading information which our relativistic approach can overcome.

We have also developed a comprehensive approach to analyzing commercials which gives advertising agencies and marketers clear directions for making more effective commercials. Among some of the key information made possible through the use of the "Relative Interval Scale" in commercial testing is: identification of the attributes which should be stressed in a commercial in order to maximize purchase of the product; analysis of the attitudes evoked by each frame of a commercial; identification of the focal message of a commercial.

All of these advances in scalar interviewing and data interpretation have been made possible through the integration of ancient Hindu philosophies and modern computerized interviewing procedures.

Finally, imagine two conversations. In one, the speaker says, "Have a good day." In the other, someone says, "May today be better than yesterday." The first phrase represents an absolute manner of speaking while the second is a relativistic approach. Both are intended to offer positive thoughts to the listener. Think about which communicates the hopes and intentions of the speaker more directly. Which do you feel is clearly more positive in nature? Those of you who sense that the relative approach communicates more effectively, please consider its potential for measuring attitudes more accurately than the current absolute, interval scale procedures.

Comment on Goodman

Robert V. Miller
MarketVision Research, Inc.

According to Gary Maranell, a well respected sociologist, measurement can be defined as "the assignment of numbers . . . to objects according to rules." Basically, what Goodman is providing are some new rules to follow in the development of scales, or measurements. In practical day-to-day marketing research I am surprised at how little attention we give this subject. We all think we are measuring what we think we are measuring, but sometimes we are not.

I am very much in agreement with Goodman in his "relativistic view of reality." It is based upon theories of symbolic interaction where a symbol stands for something called a "referent," which is the object or item the symbol represents. Man has done a remarkable job of creating a symbolic system that accounts for the vastness of these referents. For instance, if I use the symbol "red," all of you would be able to differentiate from among a selection of the primary colors. However, should I introduce to you a variety of shades of "red," there is greater difficulty agreeing to which referent, or shade of red, I am alluding. Now, let's make the task even more difficult. To what do I refer when I use the symbol "love?" Or "democracy?" Or "spiritual?" Or to take it to a product level . . . "reliable," "safe," "easy to use?"

The basic thesis of Goodman's paper is that we should provide the respondent a clearly defined referent from which to assess attitudes. I think this is a great idea so long as the respondent understands the referent and we are able to obtain accurate and repeatable measurements from either end point of the scale.

RELIABILITY AND VALIDITY

I am sure that each of you understands the importance of reliability (the ability of the scale to measure something consistently) and validity (the ability of the scale to measure what it proports to measure). In my view, validity is a much more important issue than reliability, and is the most difficult to assess. In most cases we simply depend upon face validity (it looks like it measures what it is supposed to measure) and more sophisticated approaches have often been avoided largely because they require more time and effort. Goodman argues that his Relative Interval Scale has greater predictive validity than typical monadic or interval rating scales. Although I would like to believe that is true, I would like for him to provide us with more evidence and to discuss validity issues in greater detail.

GOODMAN'S ALTERNATIVE MODEL—THE RELATIVE INTERVAL SCALE

As I understand Goodman's concept, it based upon the theory of "benchmarking." That is, a standard is provided by which to evaluate an object or a concept (for example, a perfect orange is shown to a rater so that he or she can evaluate all other oranges.

However, Goodman goes further. He is suggesting that raters "evaluate items versus similar things which they find more or less appealing" (emphasis mine). What does this mean?

First, it means that the respondent will "make evaluations relative to a meaningful frame of reference." Second, it means that the respondent will "make evaluations relative to items which represent positive and negative dualisms with the person's frame of reference." If I leave the concept of "positive and negative dualisms" out of this for the moment, this means that the rating scale must become personalized for each respondent, which is done by using an interactive computer system.

In theory, I believe that Goodman presents a very convincing case. But I am unclear in how the individual benchmark or frame of reference items are derived. However, with this system the respondent now has a definitive "referent" on either end of the interval scale versus the "theoretical construct" that now exists in the minds of respondents when evaluating products and services.

HOW WELL DOES IT WORK?

Goodman refers to an interesting case study in the toy industry in which the relative interval scale predicted only "average sales" for the category versus a more inflated prediction using a standard interval scale. However, I would like more insights on why the Relative Interval Scale produces lower top box scores and greater bottom box scores. Also, other variables could be influencing these results and in future reporting on this work I would recommend that Goodman discuss some of these potential intervening variables and how they were controlled, as well as discuss more fully order bias and age bias.

The marketing successes that Goodman presents are impressive and very exciting. It would be useful if he would provide more information on testing procedures and protocols as well as address more explicitly reproducibility and validity issues. Overall, this is a very good paper that provides a new way of thinking about scaling.

MULTI-LINGUAL, MULTI-CULTURAL INTERVIEWING

Catherine M. Coffey
Freeman, Sullivan & Co.

BACKGROUND

Interviewing in Spanish is Becoming Routine in California

Because of the ethnic makeup of the California population, public announcements are frequently provided in both English and Spanish. Everything from utility bill inserts to bank machine instructions are now printed in both languages. At Freeman, Sullivan & Co. (FSC), most of our major clients now require that mail survey materials and interviews be conducted in Spanish as well as in English. By using the Ci2 CATI system, FSC has found that programming CATI in Spanish and conducting the actual telephone interviews in Spanish is easy.

The method FSC uses incorporates the initial development of the survey instrument and a pre-test of the survey instrument in English. The survey instrument is then programmed as a CATI questionnaire and tested again to make sure that it is operational in the CATI environment.

Then the questionnaire text file is saved in ASCII format (a universal format readable by most word processors). The file is loaded into a word processor and sent to a translation service. The service translates the text into Spanish and prepares a word processing file — usually by overtyping the English version. In highly technical survey designs used in scientific research (for example, epidemiological studies) the survey instrument is backtranslated by staff. These translations occur at either a translation service or in-house by staff in the survey laboratory. In ordinary market research, backtranslation is not done (usually because the clients do not wish to pay for it).

It helps to have Spanish speaking staff who can review, evaluate and work with the translated materials (and the translating services if need be) to ensure that high quality translations are furnished. Figure 1 shows a typical screen in English and Figure 2 shows the same screen in Spanish. The most important part of this process is to be sure that the text lines up. For consistency, interviewers must see the same question layouts in both language versions. If the layouts change between language versions, interviewers lose track of the "feel" of the interview and become more easily fatigued. For those frames with multiple response choices, or questions in which previous answers are restored, the identical layout is essential.

Once the translated file is returned to the survey lab, the translated frame file is loaded into Ci2 CATI and tested. It is then incorporated into the English version and accessed through the software's toggle capability. Spanish speaking households are assigned a Spanish language code by English-only interviewers. These cases are then directed to Spanish speaking interviewers who log in using a specific number. These cases are then called by Spanish speaking interviewers who toggle to the Spanish language screens at an initial point in the interview.

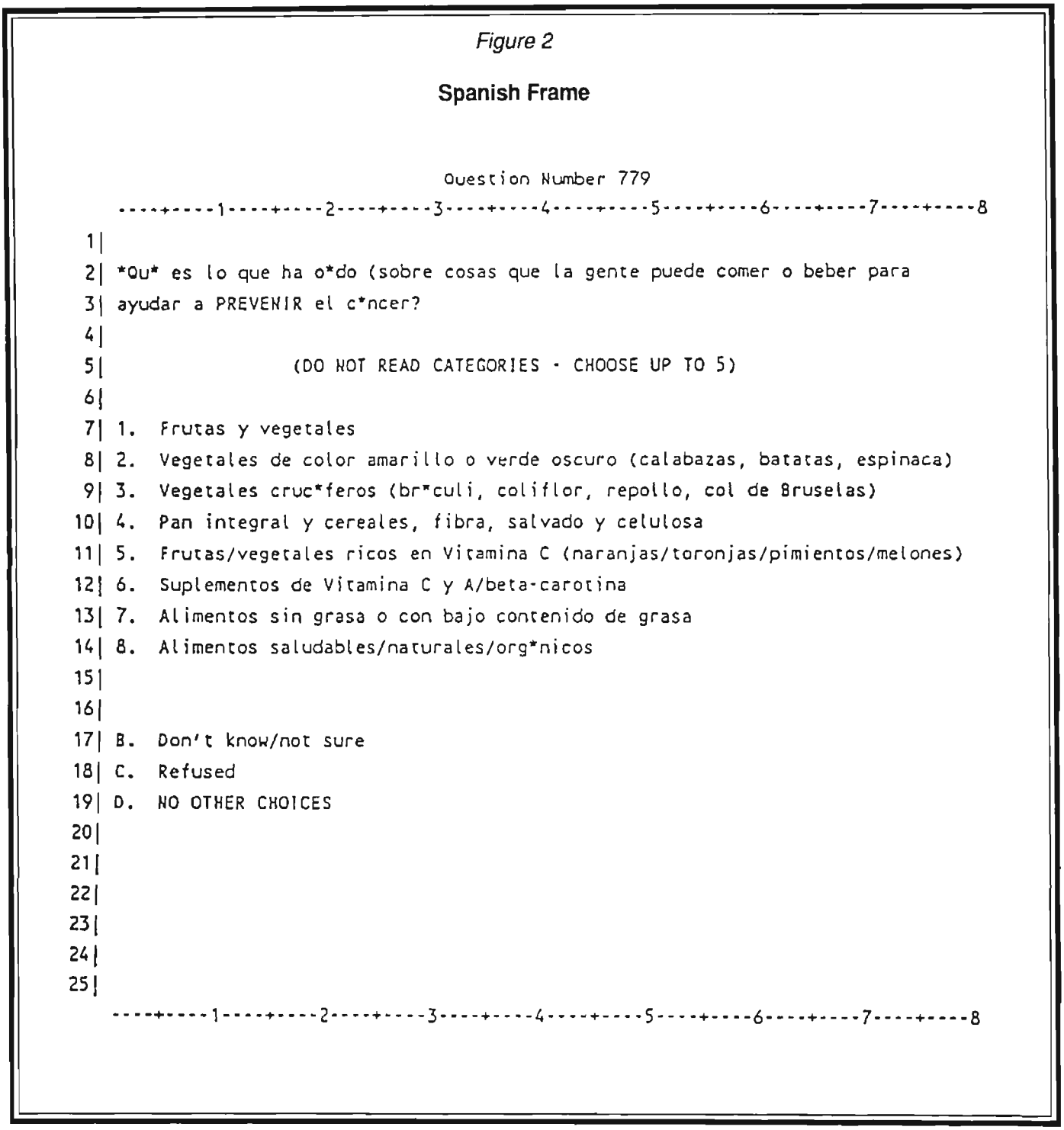
Figure 1

English Text

```
                                Question Number 779
    ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
  1|
  2| What things have you heard of (that people could eat or drink that might help
  3| PREVENT cancer)?
  4|
  5|                (DO NOT READ CATEGORIES - CHOOSE UP TO 5)
  6|
  7| 1.  Fruits and vegetables
  8| 2.  Deep yellow/dark green vegetables (squash, yams, carrots, spinach)
  9| 3.  Cruciferous vegetables (broccoli, cauliflower, cabbage, brussels sprouts)
 10| 4.  Whole grain breads and cereals, fiber, bran, and roughage
 11| 5.  Vitamin C fruit/vegetables (oranges, grapefruit, peppers, cantaloupe)
 12| 6.  Vitamin C and A/beta-carotine supplements
 13| 7.  Low/no-fat foods
 14| 8.  Healthy/natural/organic foods
 15|
 16|
 17| 8.  Don't know/not sure
 18| C.  Refused
 19| O.  NO OTHER CHOICES
 20|
 21|
 22|
 23|
 24|
 25|
    ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
```

Figure 2

Spanish Frame



If a Spanish speaking respondent is initially contacted by a bilingual interviewer, the interview is conducted on the spot, again by using the toggle option. In either case, the output data are identical — in the final data set, the respondents are identified as Spanish speaking by a coded response supplied by the system.

Demand is Growing for Survey Work in Other Languages

FSC does a lot of interviewing in California, talking to the general public as well as to targeted sub-populations. The state has almost 30 million residents, and over 43% of them are minority ethnic groups who use languages other than English.

One of our practice areas is health and nutrition. In recent years, we found that clients in this area were particularly interested in ethnic groups that speak English as a second language (if at all), and we began to investigate the possibility of using CATI for displaying the survey questionnaire in languages other than English and Spanish.

FSC Started with Vietnamese

FSC's first opportunity to use languages other than English and Spanish was to conduct a survey entirely in Vietnamese for the University of California at San Francisco, Department of Epidemiology. The interview was a modified version of Behavioral Risk Factors Survey (BRFS) carried out in many states in support of general health surveillance activities underway at the Centers for Disease Control. Although the project was successful, the solution to the management of the screen displays was a definite "kludge."

The written Vietnamese language, like English, uses Roman characters, but uses many more diacritical marks than English — few of which are found in the full ASCII character set. Figure 3 shows an example of a survey question in Vietnamese.

The method used to create the Vietnamese BRFS survey instrument was as follows: First, a method to display Vietnamese using a standard word processor was developed. (Since the client agreed to provide the translated survey materials, this was an asset in developing this stage of the project.) The approach was to double-space the materials and fill the extra space with word processing characters that approximated the Vietnamese diacritical marks. Figure 4 shows an example of a screen created this way. The commas and apostrophes represent the characters used in Vietnamese.

Next, the questionnaire was programmed using the Vietnamese language screens. The entire study was performed in Vietnamese, so there were no English screens used. The system worked normally, but the interviewers needed extra training time to become familiar with the questionnaire, and to learn the method of coding for the displays.

Figure 3

Vietnamese Survey Question

5. _____ có thường bỏ thêm nước cốt dừa vào thức ăn không? Nói một cách khác, bao nhiêu lần một ngày, một tuần, một tháng?
(VN)

READ IF NECESSARY

Có thể nói:
1. luôn luôn
2. gần như luôn luôn
3. thỉnh thoảng
4. rất ít
5. không bao giờ
8. không biết/không chắc
9. từ chối

6. Những câu hỏi sau đây là về thức ăn mà _____ thường dùng. Tất cả thức ăn mà _____ ăn kể cả ở nhà lẫn ở ngoài.

_____ ăn thức ăn nấu sẵn như là đồ hộp, thức ăn khô như khô mực, tôm khô, cá khô hay mì gói bao nhiêu lần một ngày, một tuần, một tháng?

PLEASE READ:

Có thể nói:
1. mỗi ngày
2. mỗi tuần
3. mỗi tháng
4. mỗi năm
5. không bao giờ
8. không biết/không chắc
9. từ chối

7. Lúc ăn, _____ có thường bỏ thêm muối, nước mắm, xì dầu, Magi, hoặc nắm vào thức ăn không ?

READ IF NECESSARY

Có thể nói:
1. luôn luôn
2. gần như luôn luôn
3. thỉnh thoảng
4. rất ít
5. không bao giờ
8. không biết/không chắc
9. từ chối

Certain aspects of this project go beyond just translating from English to Vietnamese, and are important whenever individuals from another culture are interviewed. Experiences with this project worth noting include:

- Many respondents were not able to report their ages. "Routine" demographics were a challenge because, in this case, immigrants from Vietnam did not know their dates of birth — birthdates are not an element of their culture.

- Questions about diet required an understanding of Vietnamese eating habits and how they have been adapted in America. It was necessary to know about Vietnamese food items, the way they are eaten and what might be substituted for these items in local groceries.

- Routine health procedures required relatively graphic descriptions. It was necessary to know the respondent's medical exam history, including pap smears and prostate exams. Often the respondents did not know these terms, although they certainly remembered the procedures when described.

Of course, the interviewer's ability to make cultural connections was an essential element in conducting the interview successfully.

It was possible, for this project, to recruit a small group of interviewers who previously worked with us to translate and perform interviews in Vietnamese for other health related surveys. Vietnamese is now offered as an additional language capability in the lab. For surveys that are performed in English and Spanish, as well as in Vietnamese, the same procedure for the Vietnamese translations is used as described earlier for the Spanish translations.

Chinese was Next, But it was a More Difficult Transition

The next opportunity was a study in the Chinese language — another modified BRFS survey — that focused on smoking. These are the reasons why this was a challenge:

- The standard set of Chinese language characters for computers includes over 13,000 characters — versus the 256 characters used by standard English language computer displays;

- Chinese language characters are not a part of CATI's standard display — so a new method to display them had to be developed; and

- Chinese is not a single language. To most Westerners "Chinese" generally means Mandarin — the primary language of both Taiwan and Singapore. However, Cantonese is spoken by people from Hong Kong and Canton, as well as by older ethnic Chinese peoples (meaning all Chinese peoples who derive from the same cultural, racial and linguistic traditions). Significant numbers of Cantonese speakers live in the Bay Area, therefore it was necessary to interview in both languages to complete the project.

The remainder of this paper describes the development of the first-ever survey in the Chinese language using a CATI.

Figure 4

Vietnamese Frame

```
                                Question Number 123
    ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
  1|
  2|          /                 _    ,           _    /               ,
  3|    Khi nau an, ____ co thUIng bo bot ngot vao thUc an khong? Co the noi:
  4|                                    ½    ½
  5|
  6| [PLEASE READ]     1  luon luon
  7|                       _
  8|                   2  gan nhU luon luon
  9|                        ,      ,
 10|                   3  thinh thoang
 11|                        /
 12|                   4  rat it
 13|                                _
 14|                   5  khong bao gil
 15|                              /
 16|                   8  khong biet/khong chac
 17|                       _    /
 18|                   9  tU choi
 19|
 20|
 21|
 22|
 23|
 24|
 25|
    ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8



          1  CLR
          2  GET    3
          3  RNG    9
          4  REJ  706  707
```

METHODOLOGICAL APPROACH

Description of the Study

Researchers at University of California at San Francisco know that the rate of lung cancer in Chinese males is higher than in Japanese and Filipino males, and that Chinese males have the highest rate of buccal cavity and pharyngeal cancer among all ethnic groups. In addition, the rate of lung cancer for Chinese females exceeds all other ethnic groups. Despite these disease incidence statistics, very little is known about the prevalence of smoking and passive smoking among ethnic Chinese in the U.S.

FSC's task was to assess the attitudes, practices and behavior related to smoking among Chinese-Americans in the San Francisco Bay Area. Objectives of the study were:

- to measure the prevalence of smoking and passive smoking among Chinese adults;
- to investigate factors related to the initiation and cessation of smoking behavior;
- to investigate attitudes, knowledge and practices related to smoking;
- to investigate the awareness of anti-smoking information through mass media and the availability of smoking cessation services; and
- to investigate the relationship between the extent of acculturation and the factors related to smoking.

Interviewers were to enumerate all Chinese people over 18 in a household and, with the appropriate respondent, conduct a 20-minute interview that included:

- current smoking status;
- recent smoking history;
- other tobacco use;
- passive smoking;
- health and social behavior;
- attitudes and knowledge about smoking;
- media exposure;
- pregnancy history; and
- education, income and acculturation

The goal was to complete 1400 interviews, 600 women and 800 men, of Chinese descent in the study area. We agreed to interview in English, Cantonese or Mandarin, based on the respondent's preference, and to do so using Chinese display characters in a CATI survey.

Sampling Approach

The sample design for the survey was a modified Waksberg Random Digit Dialing (RDD) sample. Chinese surnames are relatively unique in their spelling (for example, Huang, Yee and He), and few in number (fewer than 100). Consequently, it is easy to identify Chinese surnames from lists. Because ethnic Chinese are numerically rare, it is tempting to use listed sample frames. But listed frames have unknown and potentially large biases. For this reason, the research team settled on RDD sampling.

In the Bay Area, there are two well established Chinese communities — the Chinatowns of Oakland and San Francisco. These are places where recently arrived immigrants first live; people who have

strong ties to the Chinese culture remain. It is estimated that about 30% of the Chinese population in the Bay Area resides in these two locations. The remaining 70% of the Chinese population is dispersed in a thin layer throughout the nine Bay Area Counties — a population of about 4 million people.

The research team believed that ethnic Chinese people living in the dispersed layer would be more acculturated or "Americanized" in their behavior, including smoking habits, while ethnic Chinese in the Chinatowns would be less acculturated (that is, more likely to engage in smoking and other health risks that contribute to high cancer rates).

The sample was therefore stratified to measure any differences between these populations as precisely as possible. In addition, stratification provided a method of improving the efficiency of the contacts required by RDD sampling.

The sample design is considered a modified Waksberg because FSC developed the sample stratification scheme and sampling procedures using principles similar to those proposed by Waksberg in the 1970's. Waksberg's basic sample design is a two-stage procedure which primarily involves a random sampling group of 5-digit Primary Sampling Units (PSU). This sample is contacted and the incidence of "targeted" households is noted. Then the PSU's for which "targeted" households were found are included in a second random sampling, which is used for the main survey sample.

The first stage of a Waksberg sample is expensive and is useful only for targeting. FSC modified Waksberg's first stage by substituting a statistical analysis of the exchanges in the target area — characterizing the incidence of households in the exchanges with Chinese surnames. The exchanges were then assigned to sample strata based on the incidence of listed Chinese surnames and telephone numbers that were randomly drawn from those exchanges.

The sampling plan specified three strata: High, medium and low density Chinese residents. Each stratum contained telephone exchanges based on the incidence of listed Chinese surnames in each exchange:

- High — exchanges for which more than 30% of listed surnames were Chinese;
- Medium — 10% to 30% of listed surnames were Chinese; and
- Low — 2% to 10% of listed surnames were Chinese.

Exchanges with an incidence of Chinese surnames of less than 2% were excluded from the sample frame. Within each stratum, quotas for males and females were established. A higher number of males was specified because there was an expectation that fewer women would be smokers, and the information that could be provided by smokers was important to the study.

Results

FSC's efforts to finish this project included 5 months of interviewing that resulted in:

- 44,615 total attempts;
- contact with 17,656 non-Chinese households;
- contact with 1,971 confirmed Chinese households; and
- 1,511 completed interviews.

Analysis of the study is now underway. Results of the study should be available this fall.

TECHNICAL APPROACH TO DEVELOPMENT OF CHINESE CATI

The Problem

FSC wanted to display Chinese characters on CATI screens. English-language computer systems use the ASCII character display system shown in Figure 5. Even the characters in the extended character set (ASCII characters over 128), do not much resemble the Chinese characters that needed to be displayed. An example, the Big 5 — a standard Chinese character set — is shown in Figure 6.

A number of technical alternatives were considered in attempting to identify a CATI system that would display Chinese characters. Alternatives considered included:

- Running a TSR ('terminate and stay resident') or other program piggybacked with CATI that could display text in a window on the screen;

- Using two machines — one CATI PC with an English survey in process, and one Macintosh showing Chinese character displays for each station;

- Rewriting CATI for Windows (or perhaps UNIX);

- Writing a simplified CATI system in Foxbase; or

- Modifying or extending the ASCII character set to include Chinese characters.

All of the above approaches are possible solutions. However, for reasons which should be obvious, FSC wanted to modify our existing CATI system as little as possible to accommodate Chinese interviewing. As it turned out, it was possible. The next section describes the solution.

Figure 5

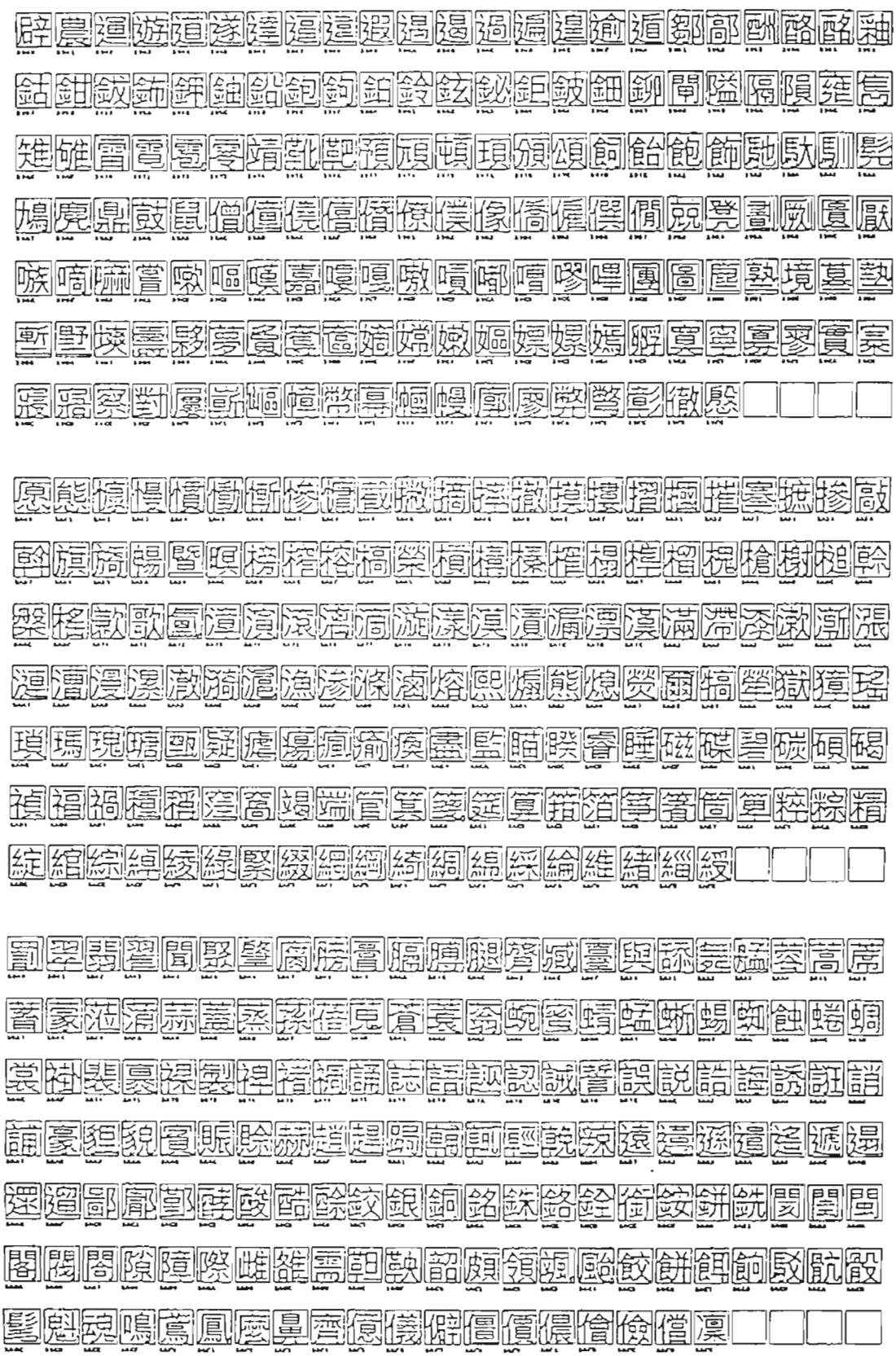
ASCII Character Set

0	00		16	10	►	32	20		48	30	0	64	40	@	80	50	P	96	60	`	112	70	p
1	01	☺	17	11	◄	33	21	!	49	31	1	65	41	A	81	51	Q	97	61	a	113	71	q
2	02	☻	18	12	↕	34	22	"	50	32	2	66	42	B	82	52	R	98	62	b	114	72	r
3	03	♥	19	13	‼	35	23	#	51	33	3	67	43	C	83	53	S	99	63	c	115	73	s
4	04	♦	20	14	¶	36	24	$	52	34	4	68	44	D	84	54	T	100	64	d	116	74	t
5	05	♣	21	15	§	37	25	%	53	35	5	69	45	E	85	55	U	101	65	e	117	75	u
6	06	♠	22	16	▬	38	26	&	54	36	6	70	46	F	86	56	V	102	66	f	118	76	v
7	07	•	23	17	↨	39	27	'	55	37	7	71	47	G	87	57	W	103	67	g	119	77	w
8	08	◘	24	18	↑	40	28	(	56	38	8	72	48	H	88	58	X	104	68	h	120	78	x
9	09	○	25	19	↓	41	29	)	57	39	9	73	49	I	89	59	Y	105	69	i	121	79	y
10	0A	◙	26	1A	→	42	2A	*	58	3A	:	74	4A	J	90	5A	Z	106	6A	j	122	7A	z
11	0B	♂	27	1B	←	43	2B	+	59	3B	;	75	4B	K	91	5B	[	107	6B	k	123	7B	{
12	0C	♀	28	1C	∟	44	2C	,	60	3C	<	76	4C	L	92	5C	\	108	6C	l	124	7C	¦
13	0D	♪	29	1D	↔	45	2D	-	61	3D	=	77	4D	M	93	5D	]	109	6D	m	125	7D	}
14	0E	♫	30	1E	▲	46	2E	.	62	3E	>	78	4E	N	94	5E	^	110	6E	n	126	7E	~
15	0F	☼	31	1F	▼	47	2F	/	63	3F	?	79	4F	O	95	5F	_	111	6F	o	127	7F	⌂

128	80	Ç	144	90	É	160	A0	á	176	B0	▒	192	C0	└	208	D0	╨	224	E0	α	240	F0	≡
129	81	ü	145	91	æ	161	A1	í	177	B1	▒	193	C1	┴	209	D1	╤	225	E1	β	241	F1	±
130	82	é	146	92	Æ	162	A2	ó	178	B2	▒	194	C2	┬	210	D2	╥	226	E2	Γ	242	F2	≥
131	83	â	147	93	ô	163	A3	ú	179	B3	│	195	C3	├	211	D3	╙	227	E3	π	243	F3	≤
132	84	ä	148	94	ö	164	A4	ñ	180	B4	┤	196	C4	─	212	D4	╘	228	E4	Σ	244	F4	⌠
133	85	à	149	95	ò	165	A5	Ñ	181	B5	╡	197	C5	┼	213	D5	╒	229	E5	σ	245	F5	⌡
134	86	å	150	96	û	166	A6	ª	182	B6	╢	198	C6	╞	214	D6	╓	230	E6	µ	246	F6	÷
135	87	ç	151	97	ù	167	A7	º	183	B7	╖	199	C7	╟	215	D7	╫	231	E7	τ	247	F7	≈
136	88	ê	152	98	ÿ	168	A8	¿	184	B8	╕	200	C8	╚	216	D8	╪	232	E8	Φ	248	F8	°
137	89	ë	153	99	Ö	169	A9	⌐	185	B9	╣	201	C9	╔	217	D9	┘	233	E9	Θ	249	F9	·
138	8A	è	154	9A	Ü	170	AA	¬	186	BA	║	202	CA	╩	218	DA	┌	234	EA	Ω	250	FA	·
139	8B	ï	155	9B	¢	171	AB	½	187	BB	╗	203	CB	╦	219	DB	■	235	EB	δ	251	FB	√
140	8C	î	156	9C	£	172	AC	¼	188	BC	╝	204	CC	╠	220	DC	▄	236	EC	∞	252	FC	n
141	8D	ì	157	9D	¥	173	AD	¡	189	BD	╜	205	CD	═	221	DD	▌	237	ED	φ	253	FD	²
142	8E	Ä	158	9E	₧	174	AE	«	190	BE	╛	206	CE	╬	222	DE	▐	238	EE	ε	254	FE	■
143	8F	Å	159	9F	ƒ	175	AF	»	191	BF	┐	207	CF	╧	223	DF	▀	239	EF	∩	255	FF	

Figure 6

Subset of the Big 5 Chinese Character Set



The Solution

FSC's method of displaying Chinese characters in the CATI environment uses a combination of hardware and software fixes. This solution was adapted from techniques used by Chinese word processing systems for DOS machines.

First, the board that controls the screen displays and the keyboard interface — the video display board — had to be replaced. The replacement board generally functions in a standard English language mode, but also gives the machine the capability of expanding its display to include Chinese characters. (Our vendor is James Caldwell at Pacific Rim Connections, Inc., 3030 Atwater Drive, Burlingame, CA 94010, (415) 697-0911.)

Second, the Chinese character font set had to be installed in the system. This file is used by the graphic display board to produce the screen displays and printed characters. The same character set enables Chinese language word processing.

When the system is changed in this way, the board overrides the ASCII display limitations. ASCII, the conventional English language display, is a 256 character set. Each character is stored as one byte, of eight bits. To display the number of Chinese characters required, the number of bytes used to store each character in the character set must be expanded.

To do this, the board operates as a 2-byte system. Two bytes gives the machine 16 bits to work with and expands the available number of characters to 65,000, which is more than enough room for the 13,000 characters used in the standard Chinese word processing font set. The first 128 characters of standard ASCII — the ones used by most English language systems, including CATI — remain available.

Third, the CATI station had to be upgraded. FSC's standard interviewing station is a 286/16 with 2 megs of memory. Each Chinese language CATI station was changed to a 386/25 SX system with 4 megs of memory. This configuration supports the board's operating requirements and runs a speedy interviewing session in English or Chinese. The cost of upgrading was about $1,200 per station, including the new graphic display board.

OPERATIONAL APPROACH

Preparations

FSC used a Chinese translation process similar to that used for Spanish or Vietnamese translations. The translators for this project (part of the UCSF research team) hand wrote the questionnaire in Chinese with the appropriate Mandarin and Cantonese phrases. An example from the survey instrument is shown in Figure 7. The handwritten documents were input into a Chinese word processing system and delivered to FSC as a flat file. FSC staff then hand edited the file into Ci2's screen format. The system was tested in CATI and reviewed and corrected by the translation team. (FSC has successfully used WordPerfect with a Chinese font set for minor edits on Chinese frame files in-house. The process of inputting the questionnaire was done with a fully operational Chinese word processing system at Pacific Rim Connections.) An example of a screen prepared in this manner is shown in Figure 8.

Figure 7

Chinese Survey Question

F# 139 ----------

When you joined a paid stop smoking course/support group, what method/methods did you use to quit smoking?

0係 你參加收費嘅戒煙班，你用乜嘢方法去戒煙呀？

(DO NOT READ CATEGORIES, ENTER THOSE THAT APPLY)

F# 140 ----------

When you went to a health professional, what method/methods did he/she suggest to you?

0係 你去搵醫學界人士幫助你戒烟個陣時，佢地教你用也嘢方法去戒煙呀？

(DO NOT READ CATEGORIES, ENTER THOSE THAT APPLY)

Figure 8

Chinese Frame

F# 140 ———

係你去搵醫學界人士
嚟幫助你戒煙嘅陣時，
佢地教你用乜嘢方法去戒煙呀？
(DO NOT READ CATEGORIES, ENTER THOSE THAT APPLY)

F# 141 ———

係你上一次戒煙嘅陣時，你食佐幾耐尼古丁香口膠呢？

1 NO. OF DAYS:

To operate a Chinese CATI station, a disk is prepared with the Chinese questionnaire. This disk is used to boot the station and load the Chinese language drivers during the boot process.

Interviewing

During this study FSC interviewers used an RDD method to contact respondents. Many dialings were required to locate each respondent, because most of the households contacted during the study were not members of the target group. FSC interviewed in two stages:

- English-only and bilingual English/Cantonese-speaking interviewers who had been trained for the survey made initial calls using the RDD sample. Less than 3% of calls resulted in a household with at least one person of Chinese descent.

- For eligible households with English-speaking respondents, the interviewer enumerated all Chinese household residents and (if possible) conducted the interview with the enumerated individual.

- For eligible households with Chinese-speaking respondents, the English-only speaking interviewers coded the sample for Chinese language. If the household spoke Cantonese and contact was made by a bilingual English/Cantonese-speaker, the enumeration information was collected to select the respondent using the English screens. The interviewer then set up a time on a subsequent day to call back and conduct the interview using the Chinese instrument.

- The coded sample was transferred to a study set up for Chinese language interviews that operated as a separate study on CATI. Chinese-speaking interviewers, mostly bilingual Mandarin/Cantonese, drew sample from this study while using Chinese interviewing stations.

- The Chinese-speaking interviewer called the household and determined the language to be used, using the toggle option to move from Mandarin to Cantonese. A second toggle option was available to the interviewer after the enumeration process. The interview was conducted in the respondent's preferred language.

- All other CATI procedures remained in normal operation.

CONCLUSIONS AND FURTHER WORK

Conclusions

FSC successfully completed over 1,000 interviews using the Chinese language interviewing system previously described. Our conclusions regarding multi-lingual interviewing include the following:

- Sawtooth's Ci2 CATI system can fully support multi-lingual operations, including Asian languages. English, Spanish, Vietnamese, Cantonese and Mandarin screens all operate successfully and correctly under routine interviewing conditions at FSC.

- The procedures described here could be expanded to include other language character sets such as Korean or Japanese by using similar hardware and software upgrades.

- There are substantial costs involved: The CATI station upgrade is over $1,000 per station, and the recruitment of high quality bilingual interviewers can be a difficult and lengthy process, in addition to the higher hourly rate for those interviewers. However, once the first project is in place, installation of Chinese-language versions of other studies is as routine as doing Spanish. Initial training costs are also reduced with an established pool of Chinese-language interviewers.

- To manage and train a successful interviewing team, organizations that wish to perform interviews in other languages must have staff members who are fluent in those languages. This is essential for quality control and cost containment.

- Contact procedures change during multi-lingual interviewing. FSC interviewers encountered a number of languages while making initial contacts. Our procedures include a "channelling" process whereby contacts are passed through the system to an appropriate interviewer. This is an easy process for studies with only one additional language with a toggle option built into the survey. However, where more than one additional language is involved, those language cases may have to be reassigned to a different study that is dedicated to that particular language.

Further Investigations

FSC has successfully conducted an RDD telephone survey of the ethnic Chinese population in the Bay Area — an enormously difficult and expensive undertaking. The question remains: Was the RDD survey effort really worth the money, or could a listed sample have provided equally valid results? That is the next step of our investigation. A proposal to the National Cancer Institute has been submitted to carry out a replication of the Chinese study using a listed sample frame and dual-frame sampling methods.

This approach to sampling will cost less than one-third of the cost of an RDD study and may yield a more representative sample of the population (because the dual frame sample design can be used to measure and control non-response bias). It is our hope that we will be able to present the results of that work in the future.

USE OF PEN-BASED COMPUTERS IN AUTOMOTIVE PRODUCT RESEARCH

Daniel F. MacRae
Chrysler Corporation

This paper presents the history of the development of the use of pen-based computers in automotive product research at Chrysler Corporation. Solutions to the problems encountered with these devices, the benefits Chrysler has realized with their use, and their potential will be discussed.

BACKGROUND ON PRODUCT CLINICS

First, let me furnish some background on what is entailed in a typical Chrysler product research clinic. At several points during the development cycle of new products, the Product Research Department takes the prototype product to consumers to measure their acceptance of the styling, features, image, and positioning in the marketplace of the new product.

These clinics are expensive and complicated exercises. Depending on the scope of the project, we will ship up to 20 different vehicles to a convention center that offers adequate floor space (50,000+ square feet). Prior to the weekend of the research we will have recruited from between 600 to 1,200 consumers from a targeted list of new vehicle owners and intenders. Depending on the area of the country, consumers are paid an incentive fee ranging from $50 to $75 to participate in the two to two and a half hours it typically takes to complete the questionnaire. Product clinics are always held on weekends and, if necessary, weeknights to accommodate more respondents' schedules. Normal clinics take from two to four days and/or nights to complete.

The Product Research Department at Chrysler is not large enough, nor would it be efficient, to handle all of the different tasks needed to conduct a product clinic. For example, the mechanics of recruiting, questionnaire production, data processing, vehicle rentals, hiring of guides and other personnel are all handled by outside research suppliers. The more critical and sensitive aspects of the research project are tightly controlled and specified by Chrysler. In the course of a typical research clinic, the Product Research Department will:

- **Specify the sample.** We decide exactly who we want to go through our clinic. If we are taking out a new pickup truck to research, for example, we probably won't want to talk to consumers that never have owned or have never been interested in a truck. We know that the absolute key to conducting solid research that yields actionable results is to make sure we ask the right people the right questions.

- **Decide on methodology and questionnaire content.** Some firms use research companies to determine what methodology to use and determine the flow and content of the questionnaire. At Chrysler we strictly control these aspects of the study.

- **Specify and conduct the analysis.** We tell our research suppliers, in detail, what we want from them in terms of tabulations and other output from a clinic. Our internal staff analyzes and prepares the reports, recommendations, and presentations from the clinic results.

Until we started to use computers, we employed traditional paper questionnaires to collect data in our product clinics. The use of paper questionnaires limited us in numerous ways:

- **Development of questionnaires required a long lead time.** Because every research project is to some degree unique, we spent a lot of time prior to a clinic developing the questionnaire. This involved rewriting pages, cutting and pasting pages together from previous exercises, and sending the questionnaire being developed — by fax or overnight air courier — back and forth between Chrysler and the research supplier until we were satisfied with its format and content. After the questionnaire was finally typeset and proofread, additional time was required to print, collate, and bind the final version. The questionnaires were then boxed and shipped to the clinic site. Considering that an average clinic questionnaire ran between 150 and 200 pages in length, the magnitude of producing up to 1,500 of these documents was no simple task.

- **Late changes were difficult to incorporate.** If a mistake in the questionnaire was discovered after it was printed or a late change became necessary, correcting the questionnaires became a major, if not impossible, task. More than once, everyone working at a clinic could be found madly photocopying, pulling staples, and re-collating questionnaires late into the night.

- **Difficult skip patterns and sophisticated methodologies were difficult or impossible to employ.** Only very simple skip patterns can be executed using paper questionnaires, and even simple patterns are bound to confuse some respondents. This limited our ability to design different study flows that would depend on different responses to questions within the study itself. In addition, more sophisticated techniques such as conjoint analysis were found to be impractical for use in our clinic format.

- **Quick turnaround was limited.** Depending on the urgency of obtaining results from a product clinic, we had two alternatives for processing questionnaires. If we enjoyed the luxury of having a week or two between the time the clinic ended and the time we needed our tabulations, we would have all the questionnaires shipped back to our research supplier's home office and have them keypunched and verified using the supplier's personnel and equipment. If we were under tighter time constraints, data entry personnel and equipment were employed at the clinic site and data entry was performed while the clinic was in progress. On-site data entry allowed us to obtain tabs a day or two after the clinic closed, but the incremental costs incurred to accomplish this fast turnaround were considerable.

- **Errors were made by respondents filling out the questionnaires.** No matter how explicit and straightforward we made our directions, and no matter how well we formatted our questionnaires, a number of respondents were certain to make errors while filling out the instruments. The vast majority of the respondents were conscientious in their efforts to follow instructions and clearly fill out their questionnaires. Very few intentionally skipped sections or haphazardly rated vehicles, but errors did occur. For example, respondents might check two boxes when they were only supposed to check one, or they got a little sloppy and made it hard for the data entry person to tell if the check mark was in box one or box two.

- **Data entry errors were made.** Even with data entry verification, some data entry errors were made. This was especially true with deciphering respondents' handwritten verbatims and numeric characters.

DEVELOPMENT OF GRIDPAD COMPUTERS FOR USE IN RESEARCH

With the advent of the lightweight pen-based Gridpad computers, the Business Planning and Corporate Research Department recognized the potential of these devices to revolutionize the process of questionnaire development, data collection, and data processing for product research.

Using computers to collect research data was not a new idea, but the Gridpad machines were the first computers small enough and versatile enough to be carried around by a respondent participating in a product clinic. In addition, the Gridpads are pen-based systems that let respondents answer by touching the screen with the pen; no keyboards are involved. And, because the machines also can record an image of handwriting, verbatim comments can easily be captured.

Starting in 1989, the Business Planning and Corporate Research Department, with the strong support of the Marketing Systems Department, pursued the idea of using these new computers as research tools.

OBSTACLES ENCOUNTERED IN DEVELOPMENT

A primary obstacle in adapting the Gridpad machines for use in research was the fact that existing interviewing software would not run on these pen-based, non-keyboard types of machines. David Pietrowski, who then worked for Research Data Analysis, Inc., also recognized the potential for this new technology in research clinics. In early 1990, he began to develop software and interfaces to allow the Gridpad computers to be used as portable interviewing devices. In December 1990, Pietrowski founded Advanced Data Research (ADR), which focuses entirely on the development of pen-based computer applications in research.

A major limiting factor to implementing surveys on the Gridpad machines was the need for custom programming in languages such as "C" for each survey. ADR solved this problem by creating a set of programming tools that could be used to easily design pen-based surveys.

One of these tools, SidePad (patent pending), is designed to allow popular software such as Ci2, Ci3, and ACA (computer interviewing and conjoint analysis software from Sawtooth Software) to be used on the Gridpad. SidePad is a TSR ("terminate and stay resident") software program. When it is run, it loads itself into memory and returns control to DOS. When another program such as ACA is run, SidePad "steals" time from the application to check for the pen touching the screen. Pen touches are converted into simulated keystrokes, and the application is unaware that a keyboard is not connected to the computer. The major advantage to this approach is the ability to write an application that will work on a variety of both keyboard-based and pen-based computers. This can be important for companies looking for a convenient migration path to pen-based systems.

Chrysler's approach to collecting data with GridPads has been to capture answers to closed-ended questions through the use of check-boxes. Check-boxes are not only familiar to anyone who has filled out questionnaires before, but also, check-boxes lend themselves extremely well to pointing devices like the pen.

To collect open-ended responses where one wants to capture the verbatim answers to questions in the respondents' own words, Chrysler has avoided the use of handwriting recognition. The ability of the GridPad computers to read handwriting is currently not sufficiently developed to use with untrained respondents. We collect these types of data by using "electronic ink," or bit-mapped images of the respondents' printing or writing. These images can then be printed on paper or loaded to a diskette for subsequent display on a computer monitor.

Over the past year and a half, ADR has assisted Chrysler and other companies in computerizing their questionnaires and adapting them to the GridPad. For Chrysler product clinics, ADR routinely supplies 150 GridPads.

In addition to software development, there were two other obstacles to overcome before the use of GridPads in product clinics became feasible.

The first was the weight of the Grids. At a little over four pounds, the computers were too heavy to carry through a product clinic for two hours. The solution to this problem was the development of a simple strapping system which carried the weight of the machines on the shoulders and allowed freedom of the arms and hands.

To overcome the second obstacle, which involved the security and safe transportation of the machines, wheeled and lockable cases were designed that afforded not only protection for the devices, but also allowed for easy charging of the batteries.

IMPLEMENTATION

In early 1991, ADR had access to a number of GridPads, the carrying straps for the machines were on hand, and the storage/shipping cases were ready. By March, this new process was ready to be tested.

The initial test was conducted at a product clinic in San Diego, using 50 machines. Most respondents used the traditional series of paper questionnaires, and a sub-sample used the Gridpads. Acceptance of this new tool by the respondents was greater than anticipated, and no significant differences in the data collected by paper or by Gridpad were found. (In fact, the data collected by the Gridpads contained fewer errors than the data collected by the paper questionnaires.) The results of this test were so positive that all Chrysler product clinics since then have employed only the Gridpads; no paper questionnaires.

BENEFITS ACCRUED THROUGH THE USE OF GRIDPAD SURVEYS

The use of these machines has benefitted Chrysler in a number of ways:

- We have stored all of the questionnaires we have used on the Gridpads in our PC's. Now instead of physically cutting, pasting, and rewriting on paper, we are able to assemble our questionnaires from the on-disk libraries containing batteries of questions.

- During questionnaire development, we no longer fax or use Federal Express to send working copies back and forth between our offices and those of our suppliers. We simply connect to the supplier's computer via modem and transmit the current versions of the questionnaires electronically. This allows for much faster questionnaire development.

- In addition to only collecting the respondents' answers to questions, the software also is designed to record how long it takes each respondent to answer a single question, a page of questions, or a whole battery of questions. In the questionnaire development phase, we now can compute accurate estimates as to the length of time respondents will need to devote to the exercise. This lets us design studies that make the best use our respondents' time.

- Because printing, collating, and assembling of paper questionnaires is eliminated, we are now ready to execute a project within a matter of minutes after the final questionnaire has been approved.

- We now regularly use complicated skip patterns that we always wanted to employ but couldn't when we were using paper questionnaires.

- Two conjoint, or tradeoff, studies have been successfully conducted using ACA and other packages. As mentioned above, conjoint studies were almost impossible to execute using paper.

- The data we collect with the Gridpads are error free, compared to the data collected in a paper questionnaire. This is because respondents cannot skip required fields, enter numbers outside the range of the scale presented, check more than one box, or not follow the flow of skip patterns.

- Data entry errors have been eliminated because this step in data collection is no longer necessary. Data entry now consists of a simple downloading of the information from the Gridpad RAM cards to a PC. This procedure takes but a few seconds.

- Eliminating data entry allows extremely fast turnaround. Tabulations are now required within hours of the close of a clinic, and, when advantageous, interim tabulations can be produced as the clinic is in progress. With paper, this would have only been possible by employing a large pool of data entry personnel on-site.

- Late changes and corrections can now be made with a fraction of the effort and cost associated with late modifications of a paper questionnaire.

- On check-out, the eligibility of a respondent for follow-up focus groups can be immediately determined.

- We are now very successfully using detailed line drawings of vehicles displayed on the Gridpad screen to record ratings, likes, dislikes, and images with more precision and accuracy than has been previously possible. Collection of data in this graphic format has also allowed us to present the findings in a visual manner rather than just in tables and charts. The VUCOM company has developed, and is enhancing, software to easily produce sophisticated graphic presentations at reasonable cost.

- Many costs have been eliminated. All the money we used to spend on typesetting, printing, collating, binding, data entry, paper and even pencils is now saved.

- As our principal suppliers become more adept in programming and using the Gridpads, they inform us that their overhead costs are coming down, and these savings are being passed on to us in the form of lower costs.

In summary, the considerable effort and money that Chrysler, ADR, and our suppliers have expended in pioneering the use of pen-based computers in research has been well worth it. We now enjoy capabilities and have a flexibility that we could only have imagined just a few years ago.

Ci3: INTRODUCTION AND EVOLUTION

Richard M. Johnson
Sawtooth Software

THE CHANGING ENVIRONMENT

Early in 1992, Sawtooth Software introduced the Ci3 System for Computer Interviewing. Like its predecessor Ci2, Ci3 is a software system for authoring and administering questionnaires using PCs.

Those of us who started Sawtooth Software were already familiar with PC-based interviewing. I still have vivid memories of one of the first computer interviewing studies. We swallowed hard and purchased a dozen Apple II computers, which arrived a day before they had to be sent to Atlanta, where the field work was to be done. Half of the computers were dead on arrival. By combining parts of the various machines, we got nine of them to function. Having learned that they didn't travel well, we transported them from Chicago to Atlanta on a soft mattress on the floor of a van.

Those early machines had severe reliability problems, not the least of which was their sensitivity to static electricity. Respondents sometimes scuffed their feet on the carpet, and that caused real problems. When an interviewer would call in complaining of machine problems, our first response was to say "Have the respondent take off her shoes." That often seemed to work.

In those early days each interview had to be programmed "from scratch" in BASIC or a similar language. The limited capabilities of the first computers, together with the lack of an authoring system, meant that computer interviewing was difficult, expensive, and not to be undertaken lightly.

However, respondents enjoyed being interviewed by computer, and the author had more control over the interview than with paper-and-pencil or interviewer-administered interviews. Also, questionnaires could be more complex, and tailored for each respondent. We were confident that the market research industry would move toward PC-based interviewing as PCs increased in capability and decreased in cost, so we decided to develop a system to compose and administer computer interviews.

Development of Ci2 began in the early 1980's. In the decade that followed there were dramatic improvements in the capabilities of small computers.

The computers that we used in our first efforts had only 64K bytes of memory, external disk drives, small monochrome displays, and cost about $3,000. With the introduction of the IBM PC, several limitations were removed. The IBM PC had 256K of memory, four times as much as the early Apple II. Color monitors were also available, though more costly than monochrome. It appeared that the IBM PC and its clones would be adopted by the business world, and we decided to develop our system for the IBM platform.

We knew there was a "chicken and egg" problem with PC interviewing. Researchers wouldn't be able to field computer interviews unless field agencies had PCs, but field agencies had little incentive to acquire computers for interviewing unless first assured of a volume of business.

The cost of computers for interviewing remained a problem; but there were rumors of an exciting new development that might solve the affordability problem: the PC Junior! The PC Junior was expected to cost less than the regular PC and run most of the same software, although it would have only 128K of memory. Of course, we couldn't know that the PC Junior would fail in the market and eventually be withdrawn by IBM. We welcomed the appearance of a relatively low-cost interviewing machine, and we designed Ci2 to run in just 128K of memory.

Today a PC costs much less and can do far more. It's interesting to compare the early IBM PC with the machines we use today. For example, this talk was written using a '486 machine. Although such machines cost about what the IBM PC did ten years ago, mine has about 60 times as much memory as the early IBM PCs, and operates about 70 times as fast. At the lower end, a machine capable of interviewing costs hundreds of dollars today — rather than thousands — and it has much more memory, runs far faster, and has a vastly superior color screen. Perhaps most important, computers are incomparably more reliable today. Ten years ago, when there was a problem in the field it was almost always a hardware problem. Today, hardware problems are rare.

Best of all, hardware is still improving at an astonishing rate, with performance per dollar sometimes doubling within a single year. It's hard to imagine where this technology will lead, but it's certain that computers will become increasingly important for interviewing.

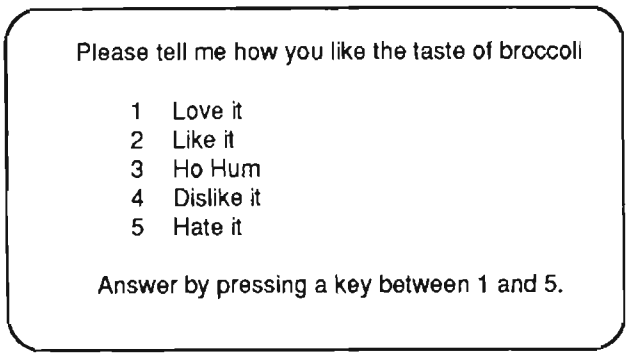
Today I'd like to tell you about several ways that PC-based interviewing has improved in the past decade. I'll describe several challenges in computer interviewing, the solution for each that we adopted in Ci2, and then how Ci3 handles the same problem.

ENTERING QUESTIONNAIRE MATERIAL

Any questionnaire development system has to let the author enter two kinds of information: the text that the respondent will see on the screen, and the questionnaire logic or instructions that govern the flow of the interview.

Text: It would not be unusual for a long questionnaire to have more than 128K characters of text all by itself, not counting logic instructions or anything else. For example, the computer screen normally has 25 lines of 80 characters, or 2000 characters per screen. Just 64 such screens would have filled the entire 128K memory of the PC Junior, with no room for other essentials such as logic instructions, the interviewing program, or data provided by the respondent!

We realized we could save a lot of space if we avoided duplicating the same text several places in the questionnaire. For example, suppose one question was:



and other questions asked the same thing about cauliflower, spinach, and asparagus. We needed to reuse the same text, rather than clutter up memory with many copies of the same screen.

Our solution was to have a file of "Frames." Frames were pictures of screens that might be used at any point in the interview. A question's instructions could tell the computer first to "GET" a particular frame, and then later to display whatever text was unique to that question.

That solved part of the problem, but not all of it. There's nothing special nowadays about creating a page of text using a computer, since so many of use word processing software; but in the early 80's that wasn't the case. We also had to provide a special text editor for composing frames.

The concept of frames did lead to efficiencies, but was also inconvenient. It was confusing to have one series of logic instructions and another series of frames; and to change a questionnaire, you might have had to make changes in both places.

Logic: Ci2 used a "language" consisting of instructions such as GET, RNG, and SKP. Each Ci2 instruction had a three -character name, often followed by numbers. For example:

GET 13 meant to "get the 13th frame, and display it on the screen.

RNG 9 meant "collect a single-digit answer, typed by the respondent in the range of 1 to 9."

SKP 5 99 meant "if the answer to this question is 5, then skip to question number 99."

To put things in perspective, the programming language that served as our model for Ci2 was "Assembly Language." Ci2 provided many instructions, and together they gave the author considerable control over fine points of the interview. However, with the technology available in the early 80's, we had to make the author responsible for many details. A patient author can make Ci2 do almost anything commonly desired in an interview, but patience is a critical attribute for a Ci2 author. Compared to Ci2, Ci3 is a higher level language with more powerful instructions, and you can do more things with a single instruction.

People are more sophisticated today; many of us have word processors or text editors with which we're familiar. With Ci3 you can use your own word processor to enter text and logic instructions. There aren't any frames; text and instructions are entered at the same time, and stored in the

same place. You can still reuse text if you want, but you just "borrow" it from another question. Ci3 questions are written in different sections, denoted by these symbols:

Q: means a new question starts here.

T: means the following information is text, to be displayed on the screen when the question is shown.

I: means the following lines are logic instructions.

The broccoli question we saw earlier could be written in Ci3 like this:

```
Q: Broccoli
T:

                    How do you like the taste of broccoli?

                        1   Love it
                        2   Like it
                        3   Ho Hum
                        4   Dislike it
                        5   Hate it

                    Answer by pressing a key between 1 and 5.
I:
KEY 1 - 5
```

The material under the T: is the text, just as it should appear on the screen. The single instruction, KEY 1-5, says to accept an answer consisting of a keystroke, in the range of 1 to 5.

Ci2 referred to questions by number, both for what we thought then was simplicity, and to preserve memory. If you refer to questions by name, the computer needs to use memory to store a table of those names, and it has to reference that table whenever the question name is used. Since computers have more memory now, Ci3 refers to questions by name.

Referring to questions by name rather than number has two benefits. With question numbers, if you want to specify a skip to a question about Widgets, then you have to remember the number of the Widgets question. That number will probably change if you move questions around in making changes to the questionnaire. By contrast, if you refer to questions by name, then the names never change. In Ci3, to skip to the Widgets question you write:

```
SKIPTO Widgets
```

And, you never have to renumber questions; you can change the order of questions just by using your word processor's Copy or Move features to rearrange them.

MULTIPART QUESTIONS, LISTS, AND REPETITION

Many paper-and-pencil questionnaires collect several answers per question. For example, a single question might ask about familiarity with each of several brands, or collect ratings of a brand on many attributes.

When we were developing Ci2, the need to conserve computer memory was paramount. It was more economical to permit just one answer for each question than to maintain a more flexible data structure that could keep track of multiple answers. For that reason, Ci2 permitted just one answer per question; if you wanted to have a 30-part question, you had to leave the next 29 questions blank to receive those answers.

However, with more computer memory available today, as well as more sophisticated programming techniques, Ci3 handles multipart questions differently.

Ci3 lets you collect as many answers as you want in each question, and for this it employs "Lists." For example, suppose we want to know how many meals an individual eats in restaurants each day of the typical week. With Ci2 you would include a separate question about each of the seven days in the week.

But Ci3 has a FOR...ENDFOR capability that does things automatically for each member of a list. With Ci3 you can define a list named "Days" whose members are "Sunday," "Monday," and so on, and the question can be asked automatically about each day.

To find out how many restaurant meals are typically eaten each day of the week, we could use a questionnaire like this:

```
LIST Days
   Sunday
   Monday
   Tuesday
   Wednesday
   Thursday
   Friday
   Saturday
ENDLIST

Q: Meals
T:

How many meals do you typically eat in restaurants on
this day of the week?

I:
FOR Days
   SHOW LISTEXT 10 20
   NUM 0 3 10 30
ENDFOR
```

First, a list is defined that contains the seven days of the week.

The question named "Meals" asks how many meals the respondent typically eats in restaurants on that day of the week.

LISTEXT is Ci3's shorthand for "the member of the list currently being asked about," and first contains the word "Sunday," then "Monday," and so on.

The SHOW instruction tells Ci3 to display that member of the list on the screen in line 10, starting in column 20.

The NUM instruction tells Ci3 to accept a numeric answer between 0 and 3 in a field on line 10 starting in column 30.

For each day of the week in turn, text is shown identifying that day, and a numeric answer is obtained between 0 to 3.

You can use predefined lists, as in this example, that are specified before the interview begins. However, you can also build constructed lists during the interview. For example, suppose you started with a list of brands.

> With a single question, you could ask awareness for every brand on the predefined list.
>
> Then you could construct a reduced list of just those brands of which the respondent was aware, and with a second question ask about usage of those brands on the reduced list.
>
> Finally, you could make another list consisting of just those brands the respondent has used, and with one more question ask about preferences among those brands.

We've seen how single questions can be asked automatically for every member of a list. But often one wants to ask a group of questions for each member of a list. For example, suppose for each family member you want to ask Age, Sex, Height, and Weight. You don't know how many members there will be in a particular family. In Ci2 you'd have to write as many individual questions as needed to obtain all the answers for the largest possible family.

Ci3 handles this automatically with a ROSTER instruction. For this example we'd write:

```
ROSTER Family Age Weight
```

This means: "For every member of the list "Family," ask Age and succeeding questions, up to and including Weight. This will happen automatically, and the answers will be recorded and identified by family member. Most important, you only have to write those questions once.

VARIABLES AND CONDITIONAL LOGIC

Often in a questionnaire it's necessary to do arithmetic. For example in a constant sum question you must make sure that a respondent's answers have the right total.

Arithmetic computation was possible with Ci2, where you could use an unoccupied "answer position" as a scratch pad, and with a series of ADD instructions you could accumulate a total. However, arithmetic was a little confusing in Ci2 because you had to add or subtract things referred to by question numbers rather than operating on named variables.

Ci3 provides "real" variables. For example, you might invent a variable called TotMeals. Suppose the question about restaurant meals eaten on each day of the week is called Meals. All you have to write is:

TotMeals = TOTAL Meals

In Ci3, question names can be used to refer to their answers. TOTAL is a Ci3 instruction that adds up the answers to a multipart question. With this instruction, Ci3 adds up all the answers to the Meals question and puts the answer in the variable TotMeals. That total can be displayed on the screen in subsequent questions, used in other arithmetic expressions, and saved in the output file.

Ci3's arithmetic capabilities are much richer than I can indicate here. You can do addition, subtraction, multiplication, and division with variables that you name, or using answers themselves. Suppose you were doing a study about vegetable gardening, and you already had questions named "Length" and "Width" in which a gardener told you the size of a garden. You could define a variable named "Area" and with the instruction:

```
Area = Length * Width
```

you could compute the area of the garden, to be displayed in a later question.

Further, in a multipart question such as Meals, you can refer to a particular answer with "subscript-like" notation. For example, Meals.3 refers to the answer to the Meals question for the third member of its list (which would be Tuesday).

Ci2 let you branch to one question or another based on previous answers, but didn't permit conditional action within a question. If you wanted to show one screen or another based an answer in Ci2, it had to be done by branching to different questions.

Ci3 has much richer capabilities. It has an IF instruction that permits different actions depending on previous information. For example, the IF instruction can be used to decide which of several choices of text should be shown on the screen for a particular question:

```
IF (Age > 17) SHOW "For whom will you vote?"

IF (Age < 18) SHOW "If you were 18, for whom would you vote?"
```

You can also branch based on values of variables. This logic skips to different questions based on total restaurant meals:

```
IF (TotMeals  =  0)   SKIPTO ThankYou
IF (TotMeals  <= 3)   SKIPTO LightUse
IF (TotMiles  >  3)   SKIPTO HeavyUse
```

IF statements can be used before a question is asked to determine whether a respondent is qualified for that question, during a multipart question to take immediate action if some specific response is given (such as an inconsistent answer), or after the answers are given. An example of the latter would be:

```
IF (TotMeals = 0)
    BEEP
    SHOW "Don't you eat any meals in restaurants?"   10  10
    SHOW "Please reconsider and answer again."       12  10
    PAUSE 2
    REASK
ENDIF
```

In this example, if the total number of restaurant meals is zero, then the computer beeps, the respondent is asked to reconsider, and after a two-second pause the question is asked again.

BACKING UP

The issues raised by backing up to change answers can get very complex. If an earlier answer is changed, then the subsequent pattern of branching in the questionnaire may be different, and perhaps questions that have been asked no longer should have been. One way to avoid retaining unwanted answers is to erase an answer whenever the respondent backs up over that question.

Ci2 had an "x-back" capability that let the respondent or interviewer backup to review or change an answer, and it erased the answers that were backed over. Ci2 also had a nondestructive capability called "zback," but if an answer was changed, then other answers that were no longer valid could be retained in the data file.

Many Ci2 users, particularly those using Ci2 CATI, have been frustrated by Ci2's somewhat limited capabilities for reviewing or changing answers. Ci3 has benefited from that experience, and provides more sophisticated capabilities.

Ci3 offers two ways to go back. You can go back as far as you like, one question at a time, and your previous answers to the questions are displayed. Alternatively, an interviewer can see a menu of the questions that have been answered, and select the question to be reviewed.

If an answer is changed, then the question from which the backup was initiated may no longer be on the new path through the questionnaire.

Ci3 handles this by categorizing answers that have been given, but which have been backed over, as being in "limbo." They are not forgotten, but they are no longer considered valid. When resuming forward movement in the questionnaire, there are two options:

> Come forward one question at a time, asking the interviewer or respondent to verify that each previous answer is still correct. Upon verification, each answer is removed from the "limbo"

category and reclassified as valid. This continues until a question is encountered that has not been answered previously.

Jump to the first point on the (possibly new) logical path through the questionnaire where there is no previous answer, without verifying answers in between.

Answers to questions no longer valid are not lost, but are retained in "limbo." Later, if another answer is changed, some of those answers may again become valid, in which case the questions need not be reasked. Any answers still in limbo at the completion of the interview are discarded when the data are prepared for analysis.

The author can control the amount of flexibility that the interviewer or respondent has, by permitting or blocking access to any of these functions.

MISCELLANEOUS ISSUES

The Ci3 Shell: In recent years, standards have developed for PC software user interfaces. There were no such standards when Ci2 was developed, and Ci2 has seemed increasingly "old fashioned" with each passing year. Within Sawtooth Software we're not all of the same opinion about this. I understand that most people like graphic interfaces, mice, exploding windows, and other similar modern inventions.

One of the readers of an earlier version of this talk suggested that for comic relief I should mention at this point that I myself, still consider the straight "C prompt" to be the best interface! I admit to having some attachment to C>, but Ci3 presents the author with a "look and feel" much more in the mainstream of current-day software. Ci3 menus have menu bars at the tops of the screens, pull-down menus for second-level options, and dialog boxes. It is much easier for the Ci3 user to navigate among various parts of the system than for the Ci2 user.

Multiple Studies: Ci2 was written before hard disks were common. Since it was "obvious" that people would use a different floppy diskette for each study, we made no provision for keeping information from several studies in the same directory. Ci3 manages multiple studies; many studies can coexist in the same directory, and Ci3 will never become confused about which is the current study. This system makes it easier to switch your attention from one study to another.

Text Overlays: Sometimes it's useful to overlay text on the screen temporarily, or to devote an area of the screen to some special application. Ci2 had limited capabilities of this kind; in Ci3 they're much more fully developed.

Ci3 lets the author compose HELP text for the respondent or interviewer. If a designated key is pressed during the interview, a "window" opens containing appropriate text. The size and placement of the window are chosen by the author. The HELP text can be unique for a question, or the same text can be used for questions of the same type.

Ci3 lets the interviewer record notes that become part of the data file. When a designated key is pressed, a NOTE window appears in which text can be entered. For example the interviewer might write "Resp confused" or "Respondent quit — doorbell rang."

Ci3's OPENEND and OTHER (SPECIFY) instructions also use "windows." Their sizes are chosen by the author. Open-ended answers can be up to 800 characters long, and wordwrap is automatic.

Mouse Support: In the early days of Ci2 we were all concerned that some respondents might be intimidated by computers and might be reluctant to use the keyboard. Fortunately, that concern was not justified, since most respondents actually prefer being interviewed by computer when given the choice among methods.

However, with Ci3 it's possible to conduct an entire interview without ever touching the keyboard. Respondents can use a mouse to pick items from lists or to indicate magnitudes with SELECT or ANALOG questions.

Key Assignments: In Ci2 the "X" and "Z" keys had special meanings: they were used to back up, and there was no way to change them. For example, you couldn't have the ESCAPE key mean "back up."

Ci3 lets you choose the keys you like for such functions. For example, if you want to assign the F1 key to mean "back up" you would write (in the "pre-question" part of the questionnaire):

```
BACKUP F1
```

Incidentally, in Ci3 the default key assignment for backing up is the ESCAPE key, which conforms to current standards.

Defaults: In many questionnaires there is much similarity from screen to screen. For example, many questionnaires use white text on a blue background. Ci3 gives you white characters on a blue background by "default." If you'd like to use another color consistently, Ci3 lets you set that color at the beginning of the questionnaire; once specified, it will be used automatically until you specify other colors.

Likewise, each "question type" has default values which you can change. You may decide, for example, that open-ended questions should always be answered in a window of particular size, location, and color. Once specified, that will continue automatically and you need not continually restate those specifications.

Recording Times: With Ci2 the length of the interview was always recorded in minutes, and you had the option of recording the time for each answer in seconds. With Ci3 you have the options of recording times in tenths or hundredths.

Coding: Although Ci2 produced a separate data file of open-ended answers, the early versions of Ci2 didn't provide any help with coding them. At the request of our users, we later produced a Ci2 Coder. However, it was done as an "add on," and it did not have all the capabilities we would have liked. By contrast, the Ci3 Coder was designed as an integral part of the system. It has many additional features and is more convenient to use. Best of all, it comes as a built-in part of Ci3, at no additional cost.

Database Access: Ci2 let you move data from an interview into or out of another file, but there were restrictions on how that could be done. With Ci3 we have removed many of those restrictions. You can read or write information from or to any position of any ASCII file. For example, Ci3 can be used to scan long lists of makes and models of cars, or to update large data files of respondent information.

CAPACITY AND HARDWARE REQUIREMENTS

Ci3 requires more memory than Ci2. The questionnaire program itself occupies approximately 300K of memory, about five times as much as the questionnaire program for Ci2. Fortunately, 640K has become the standard for PCs, and most machines purchased these days have even more.

Also, because Ci3 does so much more than Ci2, it places aheavier burden on the computer. We recommend that the interviewing computer have at least a '286 processor, and the machine for constructing interviews have at least a '386. Fortunately, these faster processors, like larger amounts of memory, have become widespread in the last few years.

Ci3 makes better use of memory in the interviewing computer, and can handle a questionnaire with about three times as much text as the largest possible Ci2 interview.

THE FUTURE

This account wouldn't be complete without some comment on where the field of computer interviewing seems to be going.

The most important trend, by far, is the rapid improvement of computer hardware. One important aspect is that of increasing standardization. Not long ago, "IBM Compatible" and "Macintosh" machines represented incompatible alternatives. However, this is changing, and it's likely that soon it won't matter which kind of computer one has.

Of course, computers are rapidly becoming faster, smaller, and less expensive. I believe they will soon be as common as telephones and televisions — indeed, those three instruments may merge into the same appliance. This will be a real boon for computer interviewing, because it will mean that computer-administered surveys will be common among consumers, just as they have become in business-to-business surveys, where PCs are common today.

Another trend is that software is getting easier to use. Although Ci3 is easier to use than Ci2 in some ways, there are still many ways that ease-of-use can be improved. That has become one of our highest priorities, and we expect to see dramatic improvements in the future.

SUMMARY

Ci3's advanced features are the result of three factors:

> In the last ten years computers have improved dramatically in both size and speed. Ten years ago we had to work hard to make things fit. That's less of a problem today, and now we can pay attention making things easier for the author.
>
> Ci2 was our first try at a computer interviewing system. It has grown and been enhanced through several generations, and has become the most widely used interviewing software in the world. Indeed, it may still be the best choice for some applications. However, our experience with Ci2 has taught us a lot, and those lessons are reflected in Ci3.
>
> We have had the benefit of a great many helpful user comments. Ci2 has been used by many different kinds of people. It seems as though nearly all of them have had suggestions about software improvements! Ci3 embodies many of those suggestions. It's fair to say that every significant improvement in Ci3 has resulted, directly or indirectly, from user suggestions.

Although Ci3 represents a substantial advance over previous software, we already have a "wish list" of additional features to be added. Ci3 will continue to evolve, just as Ci2 did, propelled by the desires of its users. For those of you who have provided guidance in the past, many thanks — and keep those cards and letters coming!

LESS IS MORE: TWO- AND THREE-DIMENSIONAL GRAPHICS FOR DATA DISPLAY

Leland Wilkinson
SYSTAT, Inc. and Northwestern University

INTRODUCTION

Whether in technical publications or for business presentations, marketing displays seem to take two forms: the garish or the inscrutable. The icon of the garish is the 3-D pie chart. The ideal of the inscrutable is the banner and stub table. There are cures for these afflictions, but not without price. Good design and clear presentation do not impress people. Garishness and inscrutability do. The displays in this paper are not "power graphics." But they communicate clearly.

We will look first at displays of the distribution of a single variable. Then we will examine two variable and multi-variable displays. This paper is not an exhaustive survey. Nor is it systematic. The topics chosen have been overlooked in more general discussions of graphic presentation. And the general theme is that whenever possible, display the raw data.

SINGLE VARIABLE GRAPHS

Figure 1 contains the most common single variable display: the histogram. The data are life expectancies in 17 countries for males and females, compiled by the World Health Organization.

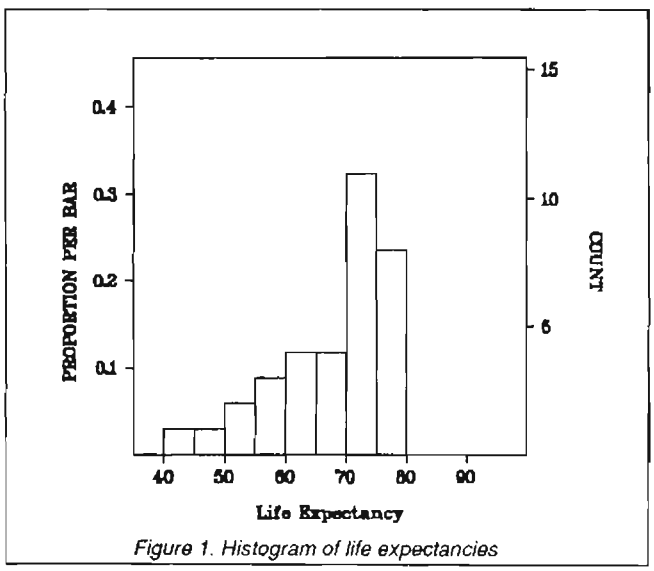


Figure 1. Histogram of life expectancies

An advantage of the histogram is that data within the categories created by the bars can be counted. It is, in effect, a graphical tabulation. Its close relative, the bar chart, is a tabulation in which the bars are discrete, or separated. Bar charts are relatively easy to construct; the categories are already intrinsic to the data. Histograms are difficult. We must decide on the number of categories before constructing them. Figure 2 shows how the shapes of histograms can be seemingly arbitrarily manipulated by choosing different bar widths and sliding the base scale on the same data. There are guidelines for making intelligent choices for the bar widths (or, concomitantly, the number of bars), but not for their location on the base scale (Sturges, 1926; Doane, 1976; Scott, 1979). What seems obvious in elementary statistics books ("pick about 15 bars and fill them") is not. Viewers who may be aware that the number of bars can affect the shape of a histogram often don't realize that the location of the cutpoints can affect it more. More generally, statistics package users don't always understand that categorizing quantitative variables such as age can affect statistical conclusions. Deciding to make

the lowest category boundary 32 instead of 40 can change the distribution of the data in the categories even when the category widths are held constant.

An antidote is to tabulate the raw data. Instead of choosing cutpoints on a scale, we can take the most significant decimal digits of the data and display them together with the next digit. Figure 3 shows this display, called the stem-and-leaf diagram. It was invented by Tukey (1977) as a form of tally which could be done with paper and pencil. Actually, unlike many statistical and graphical procedures, programming the stem-and-leaf on a computer is more difficult than doing it by hand. Good stem-and-leaf programs make intelligent decisions about picking the digits to make the display compact. Notice that we can now see the raw data. The leftmost digit of each number appears to the left of the display. The next digit (regardless of how many trailing digits there are) is to the right. At the top, for example, there is one value (44). The next line shows another value (46). The third line shows two values (51, 54). By counting the "leaves" (digits on the right) we can tell how many values there are for each "stem" (digits on the left). There are 34 leaves in all, the total count in our sample.

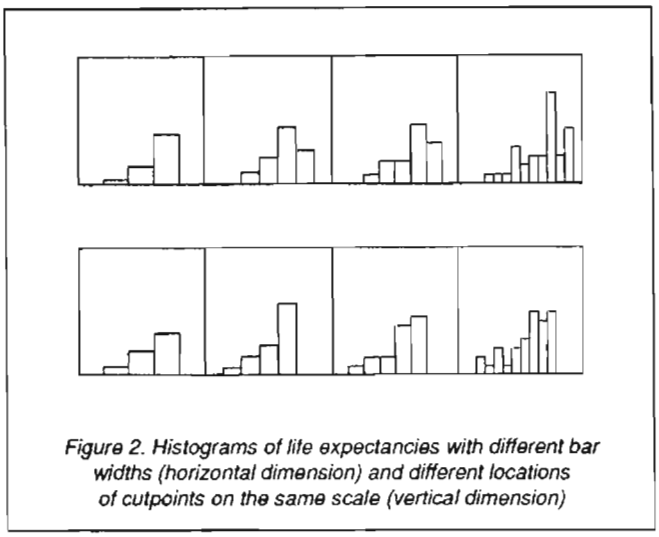


Figure 2. Histograms of life expectancies with different bar widths (horizontal dimension) and different locations of cutpoints on the same scale (vertical dimension)

Digits look crude; histogram bars look somehow more mathematical and formal. But the histogram bars are nothing more than tallies. We can make the digits little squares and then the histogram and stem-and-leaf diagram would look the same. For large samples, the digits in the stem-and-leaf diagram can be reduced in size. When counting becomes difficult, we can add a count scale. The essential point is that the stem-and-leaf diagram is immune to the scale shift problem that plagues the histogram. We will see later that by not categorizing quantitative variables, we can prevent similar problems in other displays.

```
4 4
4 6
5 14
5 557
6 1124
6 5679
7 00001122234
7 57888899
```

Figure 3. Stem-and-leaf diagram of life expectancy data

There is another way to display the density of a batch of data when we are less concerned with counting. Figure 4 illustrates this display: the kernel density (Silverman, 1986). The kernel density is, like the stem-and-leaf diagram, immune to the scale shift problem. It is also not susceptible to the bar width problem because it has no bars. The shape of the smooth can be influenced by the choice of a smoothing window

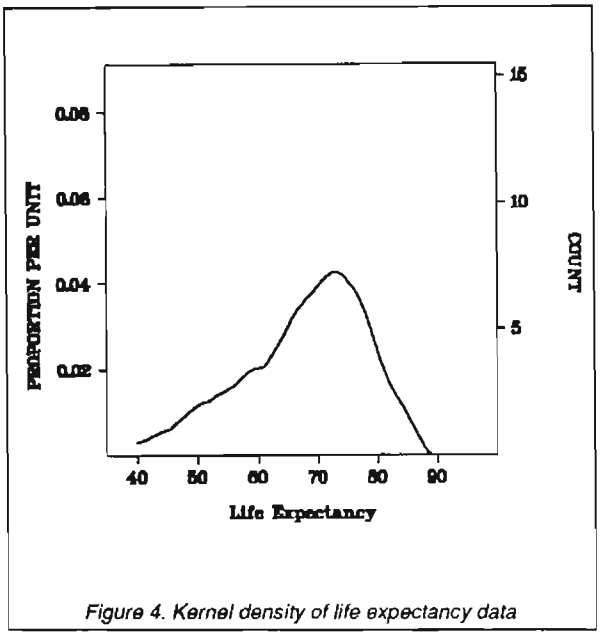


Figure 4. Kernel density of life expectancy data

width, however. Changing this width can make the kernel density look more or less smooth. Like the histogram, however, there are statistical guidelines for choosing a width.

The drawback of the kernel estimator, however, is that the data values are concealed. We cannot count values or see their location, as with the stem-and-leaf diagram. Consequently, the kernel estimator should be a supplement to other density displays and not a replacement.

There is another display which allows us to see both the raw data and the smooth: a dot-box plot. Tukey introduced the box or schematic plot along with the stem-and-leaf diagram. Its advantage is that the fractiles of the data, particularly the median and quartiles, can be seen. Its disadvantage is that it conceals the shape of the distribution. Bimodal and unimodal distributions can have the same box plot.

The dot-box plot solves this problem by displaying the box and data against the same scale. Figure 5 shows this plot for the life expectancy data. Notice that the dot values are symmetrically distributed about the center line. This type of dot plot (without the box) has been popular in the medical literature for several years.

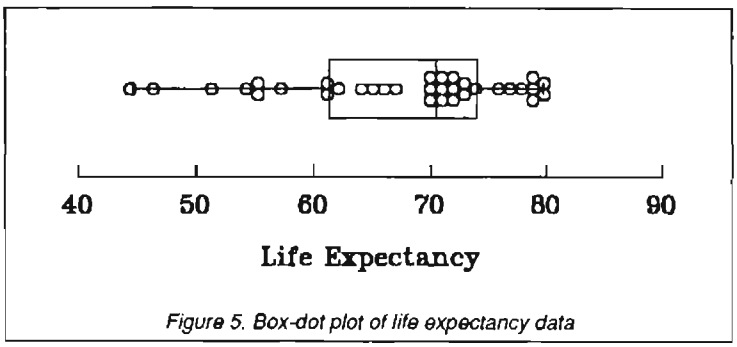


Figure 5. Box-dot plot of life expectancy data

TWO-VARIABLE GRAPHS

The most common two dimensional continuous variable data display is the scatterplot. Figure 6 shows an example for the life expectancy data. We have plotted the data for males against those for females. Enhancing scatterplots with smooths can sometimes reveal hidden structure. Like the dot-box plot, we can see the smooth and the raw data. Figure 7 shows the same scatterplot with a two dimensional kernel superimposed. This kernel reveals the skewness in the joint and marginal distributions of the data. Our eye can focus on either aspect of the display.

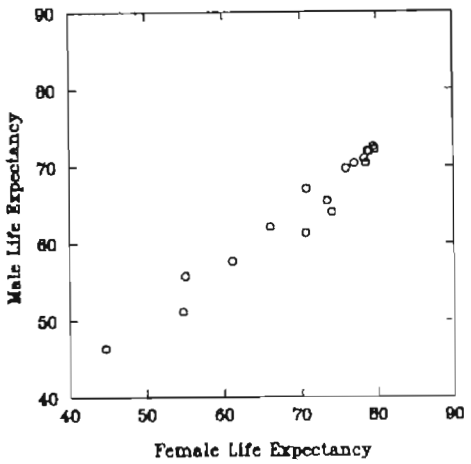


Figure 6. Scatterplot of male against female life expectancy

Unlike the kernel smooth for the histogram, the kernel is used here only to enhance perception of the data, not to conceal it or provide a potentially misleading summary. The contour lines are light enough so that the data easily show through.

Figure 8 shows how powerful this smoothing and data display can be. These data are birth and death rates per year per 100,000 people for 75 selected countries. The bivariate kernel contours are superimposed to show the joint sample distributions. Selected points are labeled. The zero population growth line at the left of the plot separates countries like Hungary, which are losing population, from countries like Guatemala, which are gaining rapidly. This graph reveals a disturbing nonlinearity and bimodality in world health statistics. Developed nations show varying birth rates but relatively low death rates. Underdeveloped nations have extremely high birth rates and high death rates. Some graphs elude parsimonious mathematical modeling. This is an example.

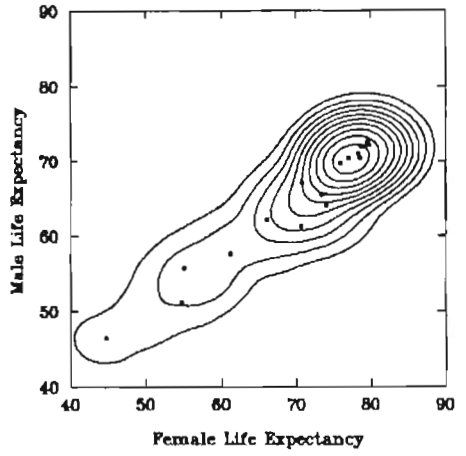


Figure 7. Scatterplot of male vs. female life expectancy with kernel smooth

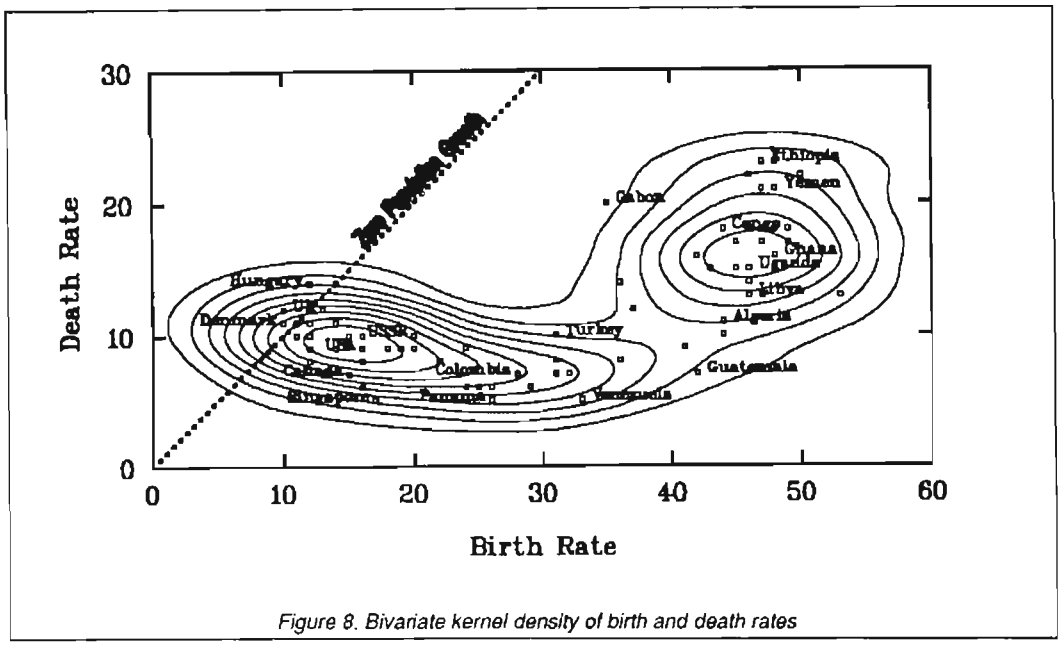


Figure 8. Bivariate kernel density of birth and death rates

MULTIVARIABLE GRAPHS

Most people think of 3-D displays when considering multivariable graphs. These are the popular graphs in computer magazines and they certainly sell software. There are even occasions when they can prove useful, particularly when the overall shape of a distribution or smoothing surface is of interest. As Becker and Cleveland (1991) have pointed out, however, statistical graphics are not in the business of creating real life scenes. Scientific visualization is fashionable now and has extensive and important applications. In statistics, however, we gain more by displaying multivariate data directly rather than by attempting to smooth them into some recognizable scene.

One of the most useful statistical displays is the scatterplot matrix (SPLOM). Like tree displays, SPLOMS are easy to understand for non-statisticians and people who have difficulty with spatial relationships. They are simply arrays of scatterplots. By placing all possible scatterplots in a single display, SPLOMs help us to see overall structure. Unfortunately, they do not reveal joint structure as do spin programs and 3-D visualization displays. High dimensional joint structure is obscure to 3-D programs as well, however, so this loss is bearable.

Figure 10 shows a SPLOM of our birth and death data, with an additional variable — health

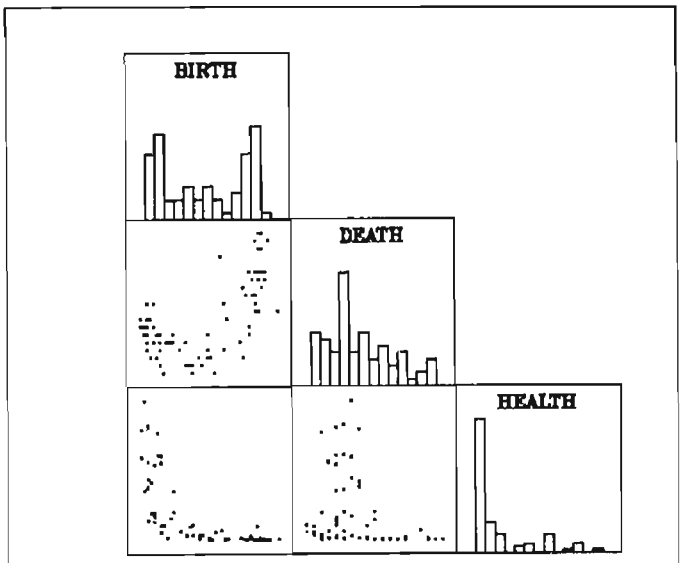


Figure 10. SPLOM with histograms.

expenditures for each of the countries, in U.S. adjusted dollars. The histograms on the diagonal indicate that the HEALTH data should be logged to reduce the positive skewness. Figure 11 shows

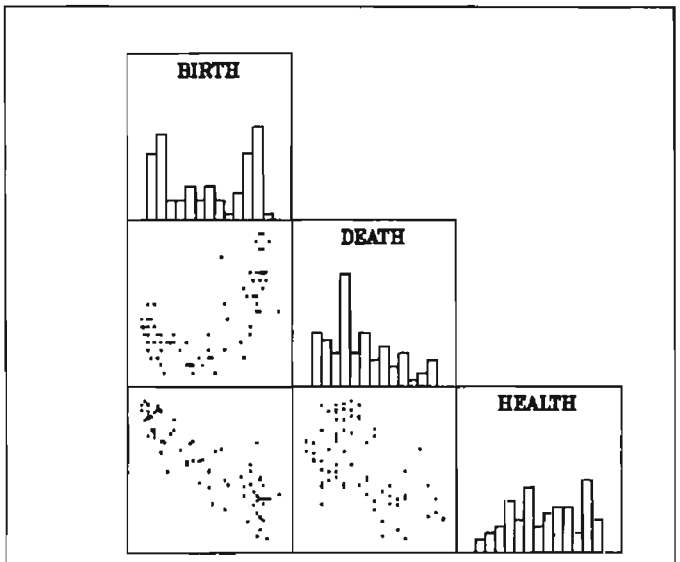


Figure 11. SPLOM of logged HEALTH

the same SPLOM after the transformation.

Now let's look at some enhancements of these scatterplots. Figure 12 shows the bivariate kernel densities superimposed on the same SPLOM. Now we see the bimodality apparent in Figure 8, but

it enters the other cells as well. In addition, we have used stripe plots in the diagonal cells instead

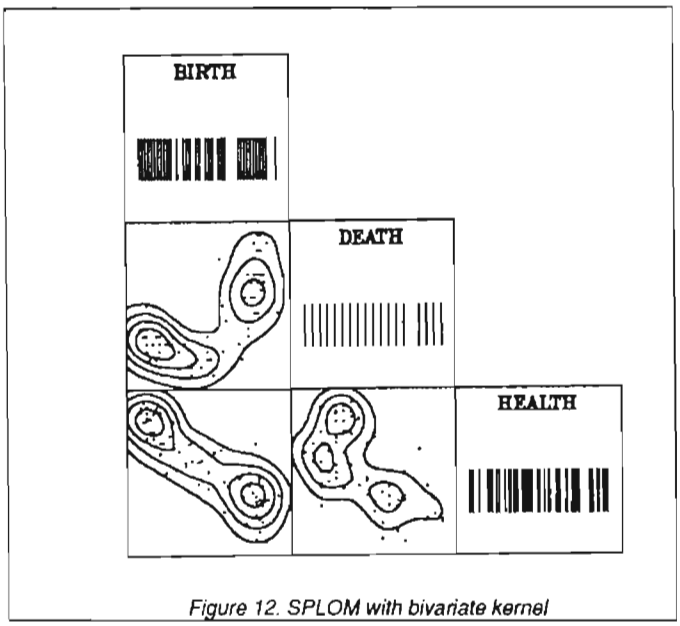


Figure 12. SPLOM with bivariate kernel

of histograms. Like dot plots, these density displays reveal the distribution of the actual data points. Other enhancements can be used to reveal different aspects of the bivariate structure. One of the most valuable is LOWESS (Cleveland, 1981). This nonlinear smoother is robust to outliers, so it is a good way to detect nonlinear trend in the bulk of the data. Figure 13 shows LOWESS curves superimposed on the same SPLOM.

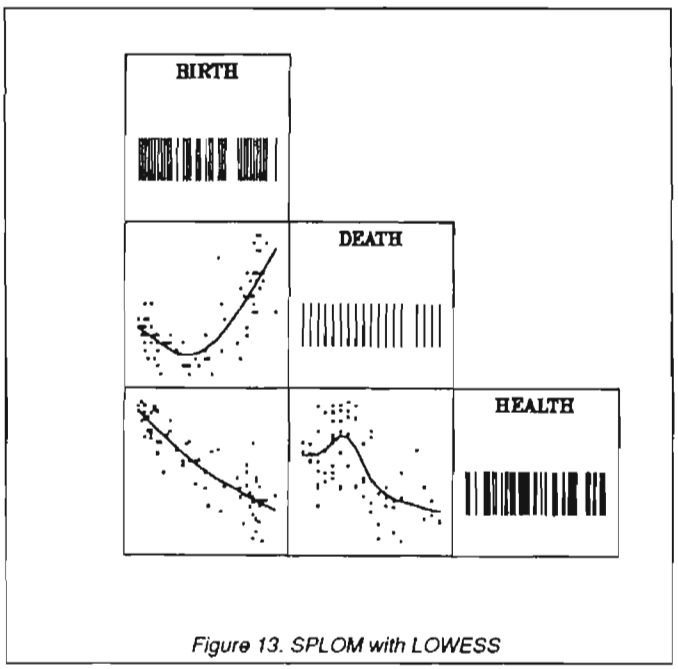


Figure 13. SPLOM with LOWESS

CONCLUSION

The graphs presented here are only a small sample of the kind which can be produced with a good statistical graphics package. All are black and white, although color has its uses. Particularly in presentations, color can be especially effective in distinguishing categories. Symbols in scatterplots can be drawn with different primary colors to reveal subgroups of the data, for example. In general, however, well designed black and white graphs can convey information succinctly and clearly. And if the data are displayed in the same graph whenever possible, it will be more difficult to deceive or convey the wrong impression.

REFERENCES

Becker, R., and W.S. Cleveland. (1991). "Take a Broader View of Scientific Vixualization." *Pixel*, 2, 42-44.

Cleveland, W.S. (1981). "LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression." *The American Statistician*, 35, 54.

Doane, D.P. (1976). "Aesthetic Frequency Classifications." *The American Statistician*, 30, 181-183.

Hartigan, J.A. (1975). *Clustering Algorithms.* New York: John Wiley & Sons.

Scott, D.W. (1979). "Optimal and Data-based Histograms." *Biometrika*, 66, 605-610.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* New York: Chapman & Hall.

Sturges, H.A. (1926). "The Choice of a Class Interval." *Journal of the American Statistical Association*, 21, 65.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

EFFECTIVE GRAPHIC PRESENTATION OF MARKET RESEARCH FINDINGS

Gordon Crowe
Gordon Crowe Associates (Canada)

INTRODUCTION

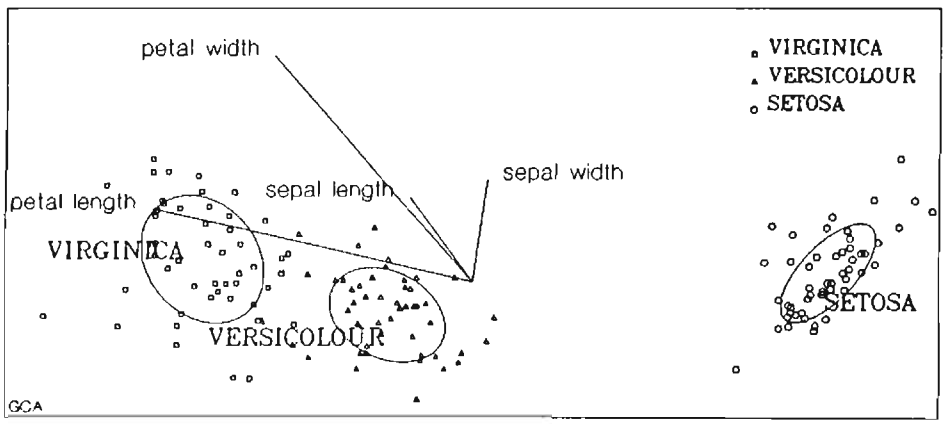
This paper presents some graph applications that I have found useful, and that I hope may provide a reference for extending your own graphic summarization of complex marketing research findings. Component plots, scatterplots, ladder plots, bar and line charts, and three-dimensional graphs that help summarize and communicate market research issues are presented. Because the example graphs are drawn from my market research work, some have been disguised to varying degrees. These graphs have been produced using the Systat statistics and graphics software.

Most of these graphs summarize considerable marketing information. Often such graphs should be presented to the viewer by unfolding the component information in several successive layers. Then once these component layers are individually understood, the relationships within the overall graph can become the focus of considerable marketing thought and discussion.

COMPONENT PLOTS AND PERCEPTUAL MAPPING

Perceptual maps generally show the point position of each group being mapped. The amount of variability around the point position of each group can be shown by plotting the reduced discriminant space point position scores for each of the original attribute rating sets. In turn, bivariate confidence interval ellipses can be plotted around the scores for each group as shown in Figure 1.

Figure 1 Discriminant map with scored data summarized using ellipses



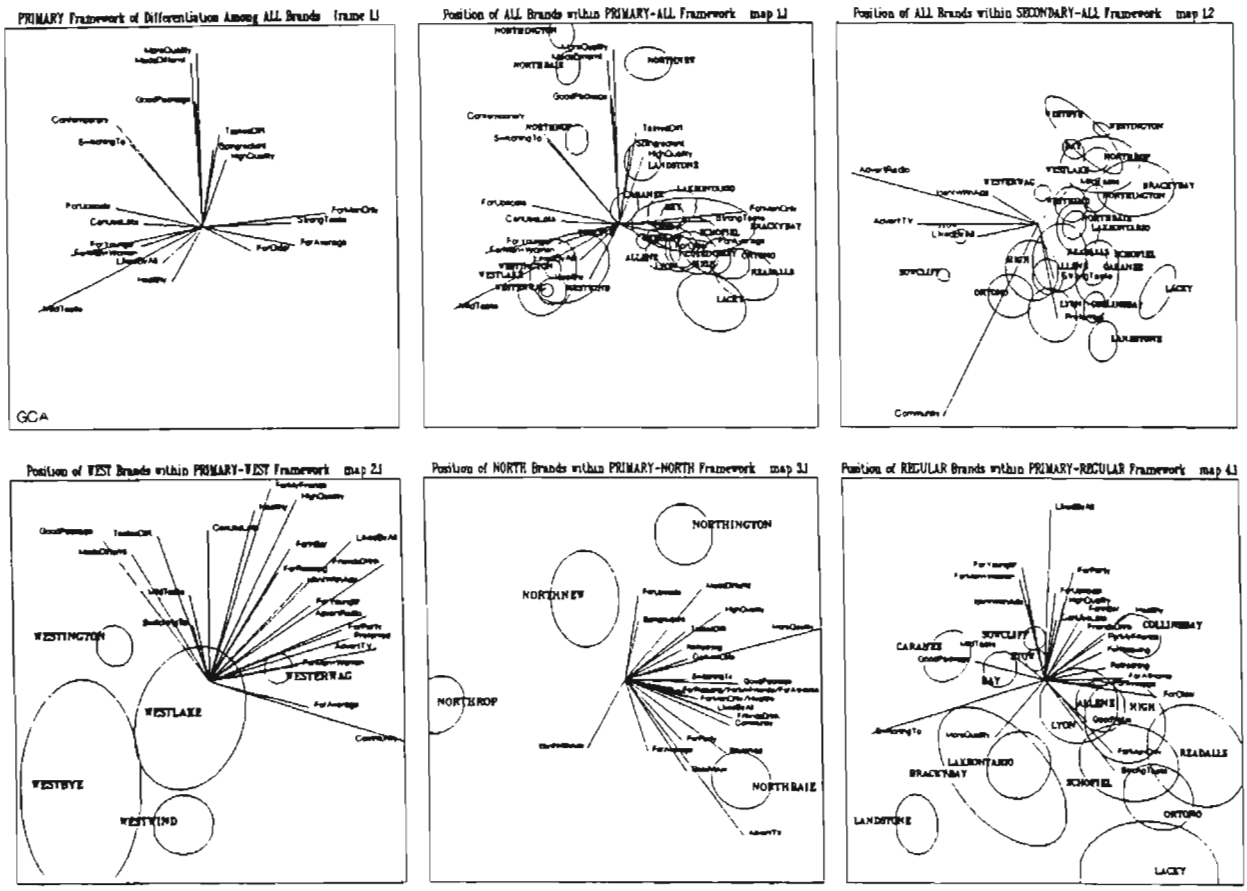
The inclusion of confidence ellipses allows better assessment of the degree to which the mapped groups indeed differ. They also help identify vague heterogeneous group images from more clearly homogeneous group images— and to the extent that the major versus minor axes of an ellipse varies — the attributes on which a group is less or more clearly defined can be discerned. Confidence ellipses are especially useful when viewed across multiple map studies or in subgroup maps within an overall market, and they should be mandatory in rolling wave-tracking maps.

Figure 1 shows a perceptual map that includes the position of each individual rating set, and an ellipse summarizing the point distribution of each group. This example is drawn from E. Anderson's 1935 Iris flower data which R. A. Fisher used in 1936 to describe discriminant analysis. Fifty observations across four attributes were obtained for each of three different Iris varieties. The interpretations of these ellipses illustrate several issues common to market research analysis: all three groups do differ; "Versicolour" is slightly more clearly positioned than is "Virginica"; "Setosa" is the most clearly positioned group in terms of "petal width" and "petal length," while being less clearly positioned specifically on "sepal width."

Sometimes it takes a number of perceptual maps to more fully describe a market. Most market research perceptual mapping starts with an overall look at all brands across all attributes. This provides the overall context for identifying the major brand subgroups and the overall distinguishing attributes for each subgroup. Then specific subgroups of brands are separately mapped to discern the points of difference among closely competing brands. Several iterations may be required/ undertaken at each level of mapping to remove obvious or overpowering attributes, or individual brands that are extremely distinct from the majority of other brands. Iterations may also be undertaken to understand particular marketing issues involving specific brand subgroups across specific attribute issues. This iterative development provides more focused marketing insight.

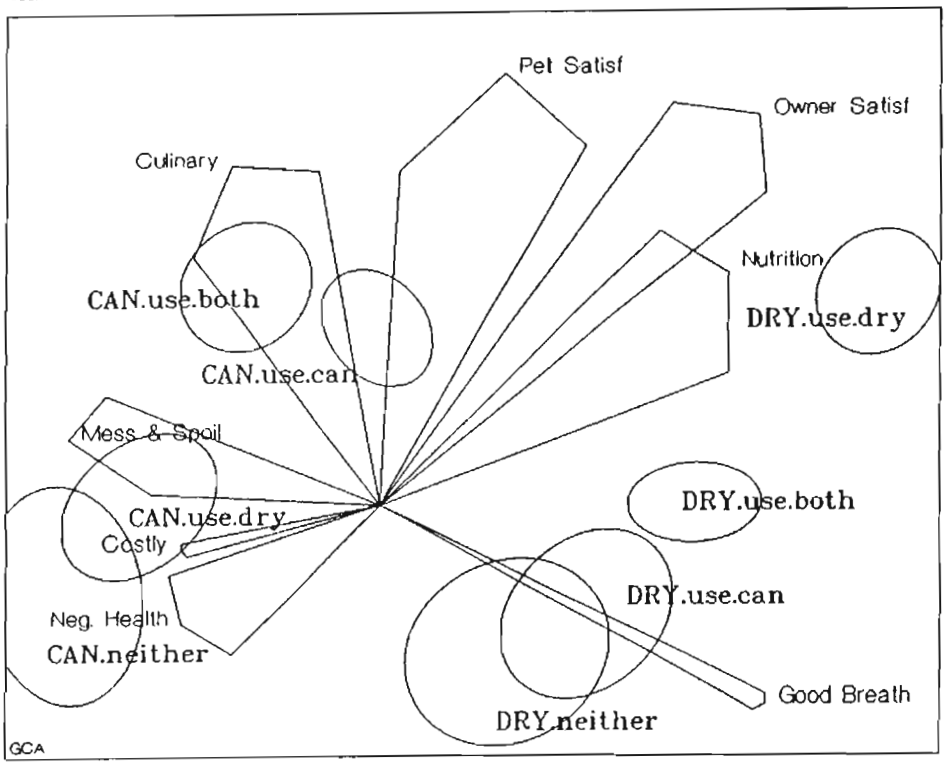
Figure 2 shows a few of the maps required to understand product attribute issues in a complex market. The upper left map shows just the attribute framework by which all brands are perceived to differ. The upper centre map then positions all of the brands within this attribute framework. This provides an overall market understanding but leaves considerable information still unexplained. Subsequent maps look at specific marketing issues in more detail. The upper right map again positions all brands, this time in terms of differentiation across a specific bundle of attributes. The three lower maps each look at the points of difference in terms of all attributes, but separately for each of the three brand subgroups identified in the upper centre map. For each additional map the attribute framework and brand configuration has been judgmentally rotated to match the overall map as much as possible. This assists understanding as successive maps are examined.

Figure 2 Some of the discriminant maps used to describe a complex market



Most perceptual mapping focuses on differences among several different objects that have been rated, usually a group of brands. The following two graphs show how perceptual mapping can help understand market research issues that are not brand related.

Figure 4 Discriminant map of product forms by usage subsamples
RELATIVE PERCEPTION OF CANNED AND DRY PET FOOD
TAKING INTO ACCOUNT JOINT USAGE ACROSS BOTH FORMS



This map shows the strengths and weaknesses of each form as perceived by those with various usage levels across both forms. An interesting point is how the "Can" form image does not align with the "Dry" form image among the users of only one form versus the users of both forms. Two attributes were excluded from this map and were discussed separately. Those attributes, "best when mixed" and "easy to mix," would have dominated the map, suppressing much of the other useful information that shows in Figure 4. Also note the many original attributes have been judgmentally summarized into a smaller number of summary factors to simplify the presentation of this map. The original study first presented the full-attribute framework and moved to the reduced summary-attribute framework in several stages, prior to presenting the reduced framework with the group positions.

Figure 3 Discriminant map of one set of ratings by subsamples
RELATIVE IMPORTANCE OF AIRLINE ATTRIBUTES TO BUSINESS TRAVELLERS
ACCORDING TO LENGTH OF HAUL AND FARE CLASS USUALLY TRAVELLED

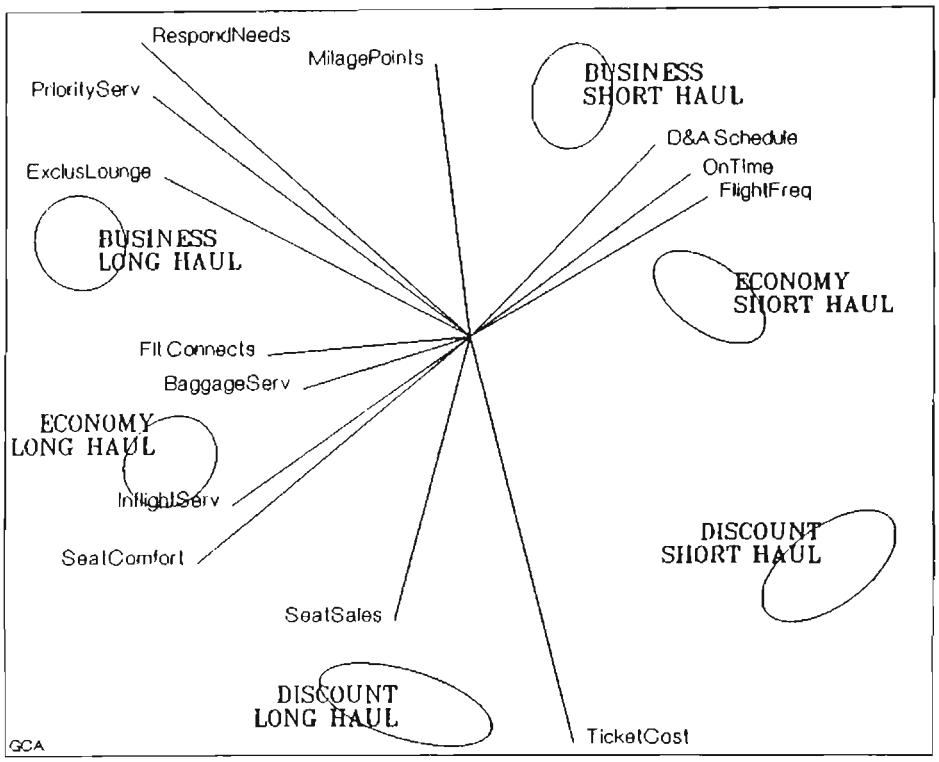


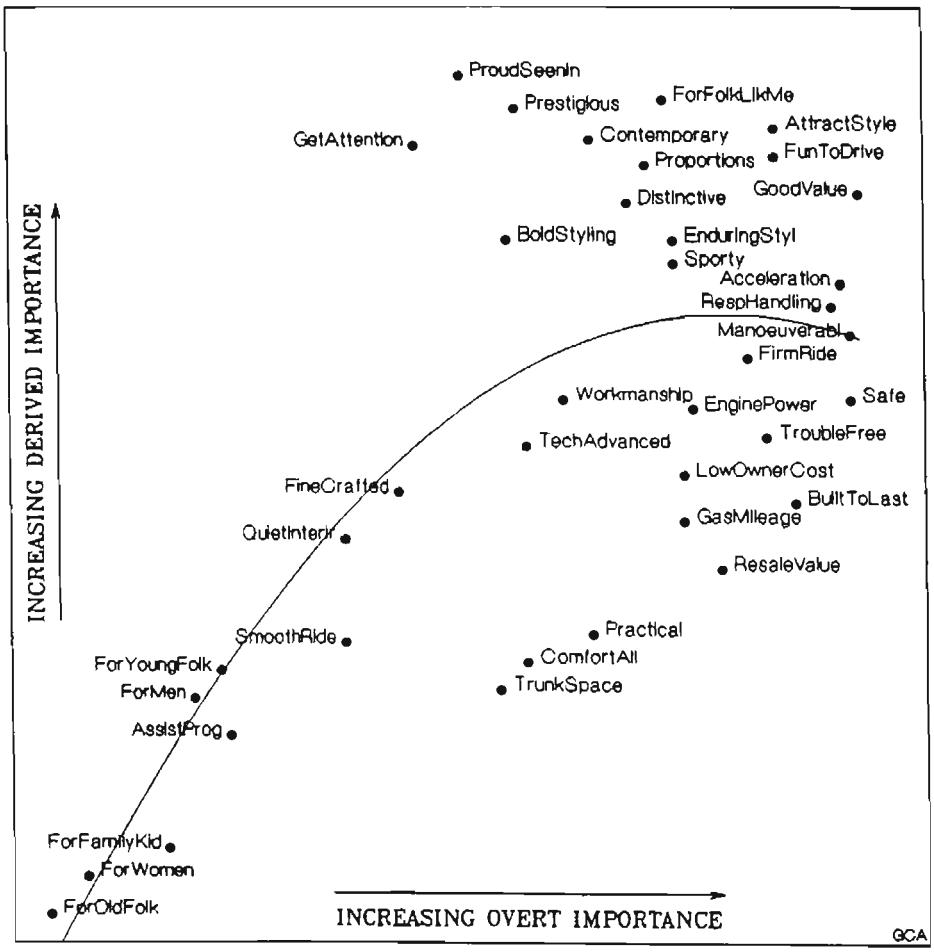
Figure 3 deals with a single set of importance ratings. The graph shows how subgroups of the sample differ on these ratings. The sample consists of business air travelers and is stratified by short versus long haul flights, and by fare class. Attribute importance was obtained cognizant of length of haul, and usual fare class was separately determined. The effects due to length of haul and due to fare class are individually apparent, as are the interaction effects.

Figure 4 deals with product form by degree of usage of each form. Although the original pet food attribute rating data was gathered at the brand level, for this graph it is collapsed to the form level. Also the ratings sets for each respondent are classified according to joint usage across the two main forms, "Can" and "Dry," and usage of other forms is ignored. The eight groups being contrasted are the two forms cognizant of four combinations of joint usage.

SCATTER PLOTS

Figure 5 is a scatterplot of derived importance versus overt rating scale importance for the entry level sports car market. Note the strong differences in attribute importance based on the two methods. Overt importance emphasises rational priorities, while derived importance may be closer to the emotive priorities underlying product choice within the market. The line provides a further reference for understanding how much the two importances vary for any attribute. The further the perpendicular distance from the line, the greater the discrepancy across the two measures.

Figure 5 Bivariate comparison scatterplot with quadratic regression line
COVERTLY DERIVED VERSUS OVERTLY STATED IMPORTANCE OF VEHICLE ATTRIBUTES



Scatterplots are useful for attribute prioritization in positioning studies. Attribute importance can be plotted by attribute discrimination F ratios to show relationships between importance and brand differentiation. The graph can be further extended to show attribute importance by attribute delivery, with the size of the plotted points indicating discrimination.

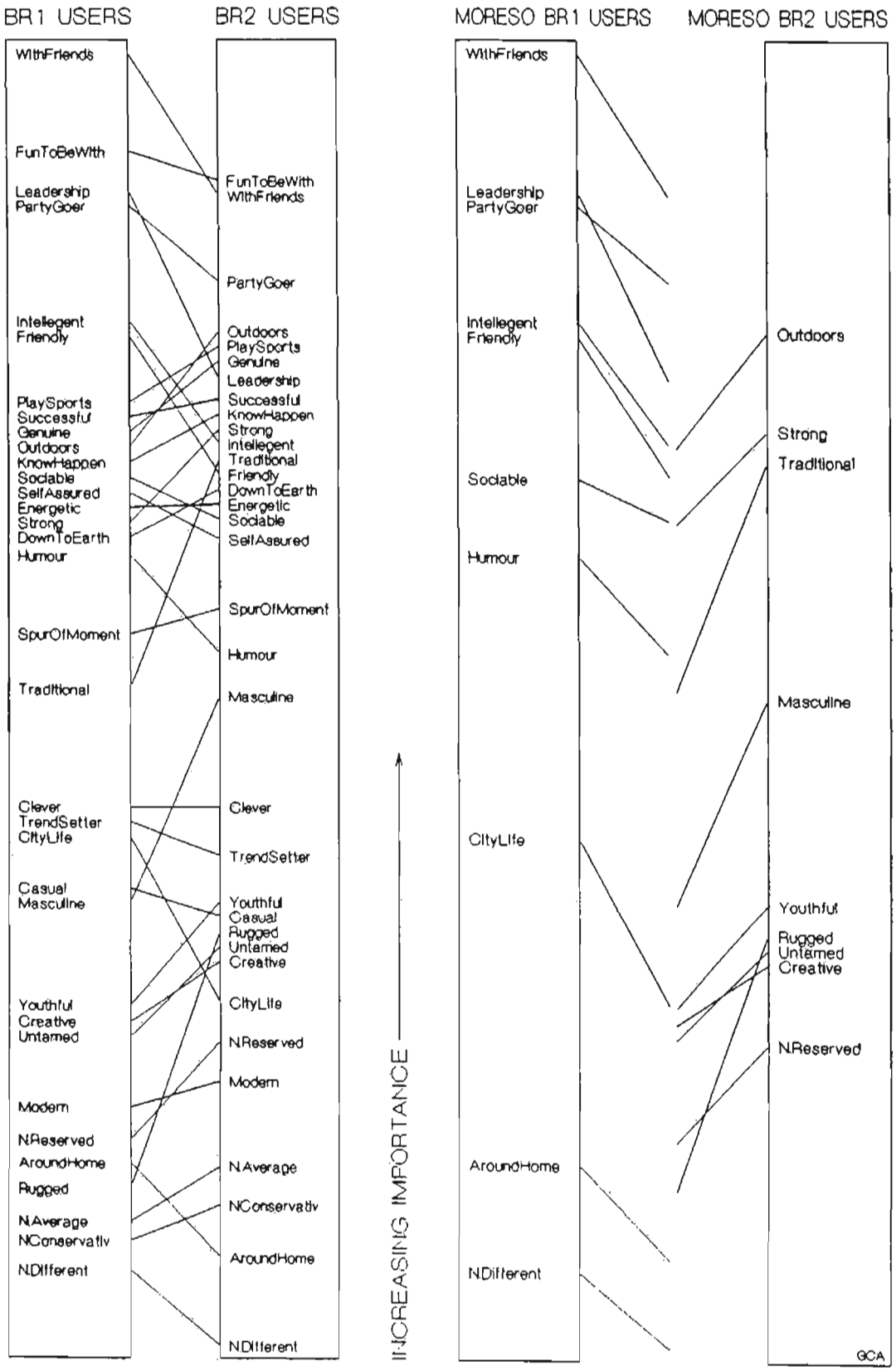
LADDER PLOTS

Derived importance analysis is generally an aggregate analysis across the total sample. This may be appropriate for homogeneous markets, or in studies where screening qualifications *a priori* determine the segment of interest. However, for large scale broadly based image studies, derived importance can be calculated separately for various subsamples. In such situations, the derived importance scores for various subsamples can be clearly contrasted using text ladder plots.

Figure 6 shows derived importance of personality attributes for two major competing brands. The left pair of graphs show all attributes positioned in terms of the importance metric. The right pair of graphs show only the significant points of difference between the two brands being contrasted. Although these brands are remarkably similar in terms of product attributes and users of both brands are similar in terms of product attribute importance that both brands deliver, the personality imagery desired by users of each of the two brands shows clear differences. "BR1" users tend to associate with urban thoughtful social imagery, while "BR2" users tend to relate to macho outdoors independent imagery.

Figure 6 Text based ladder plot comparing subsamples

DERIVED IMPORTANCE OF PERSONALITY ATTRIBUTES FOR USERS OF TWO COMPETING BRANDS

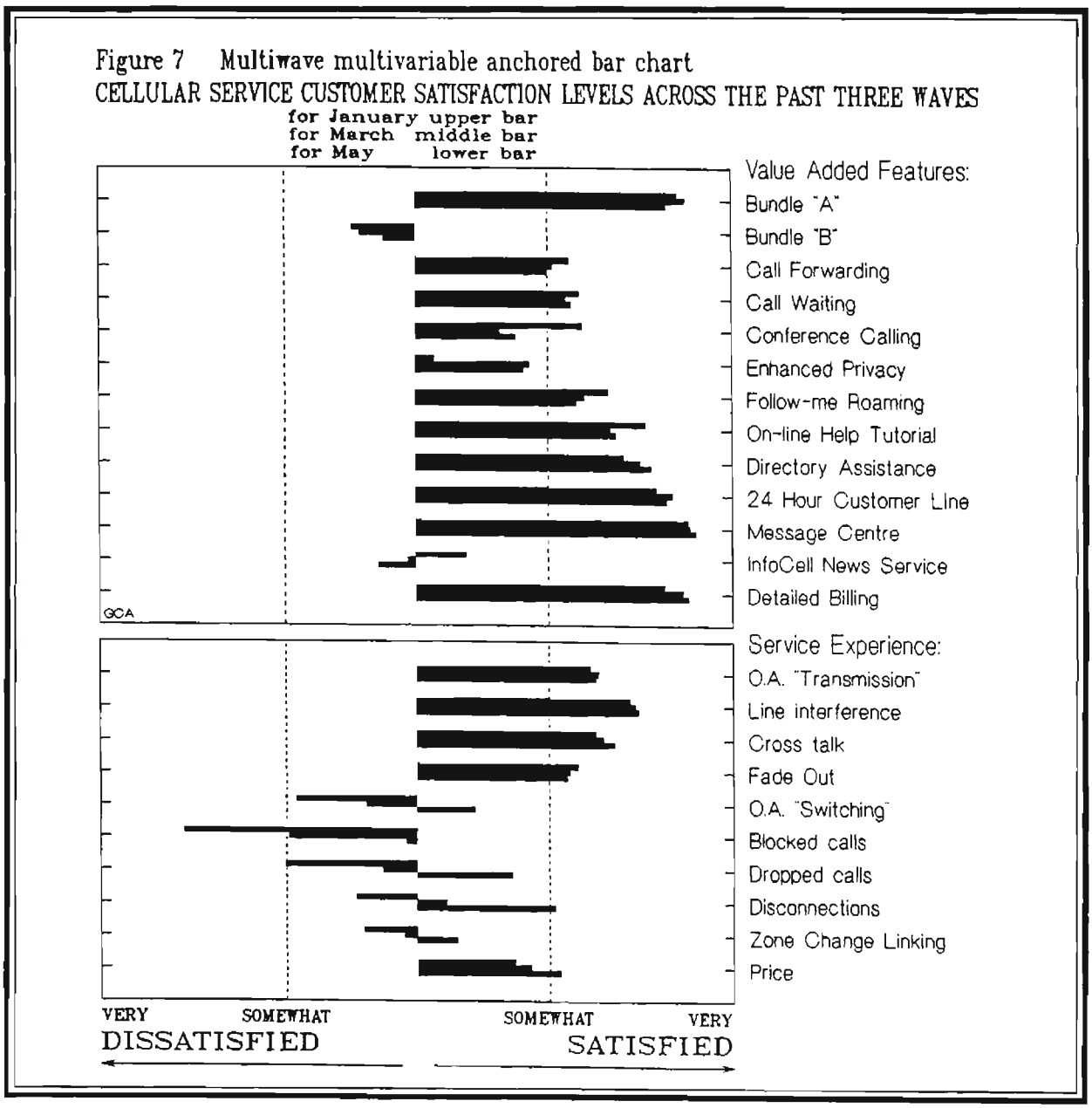


Ladder plots are also a good way of showing points of difference among attribute-based cluster analysis attitude segments. The metric can be the average level for each segment on each attribute, or it can be the extent to which each segment proportionately more or less wants each attribute, relative to the market average. Each method provides different summary information.

BAR AND LINE CHARTS

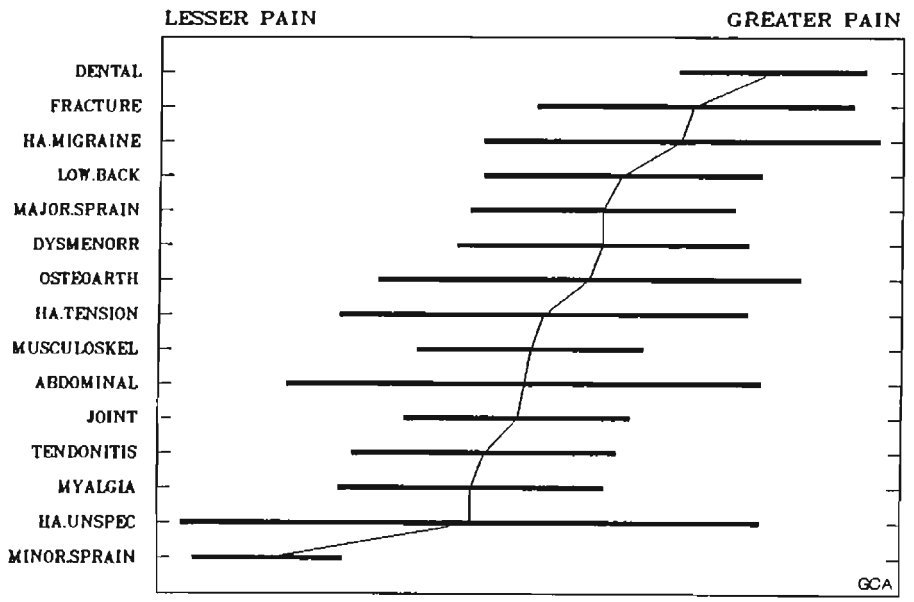
Bar charts are particularly useful in several situations. Figure 7 shows three waves of customer satisfaction levels across a factor-sorted list of attributes. The graph is mid-scale anchored and has quality control lines at the disaster level and the desired standard level. The results of recent switching technology improvements are readily discernable, as are other strengths, weaknesses, and points of change.

Figure 7 Multiwave multivariable anchored bar chart



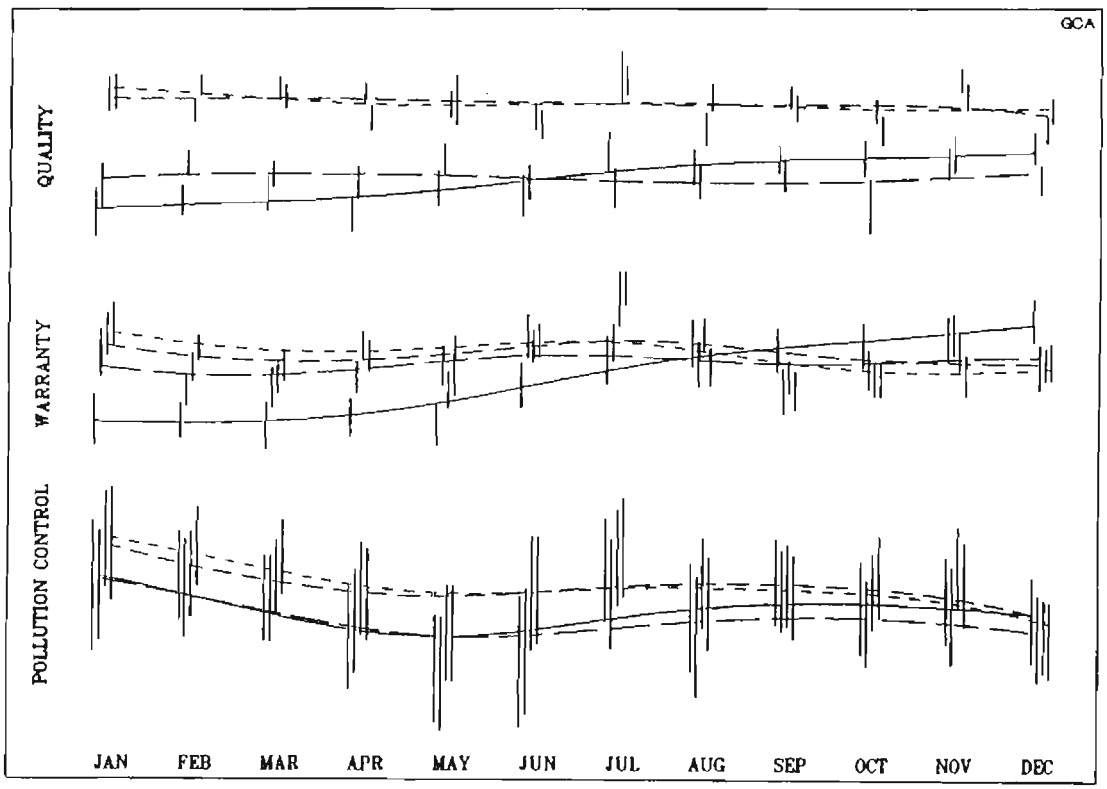
For data that can be described by average scores, some measure of variance around the average score often is critical to thorough understanding. Figure 8 shows the average level of pain for various conditions as well as the range of pain covering 90% of cases within each condition. This type of graph is particularly useful when presenting conjoint utilities and in tracking wave situations. Further detail of the underlying distributions can be shown using Tukey box and whisker plots, and dot plots.

Figure 8 Sorted mean and variance bar chart



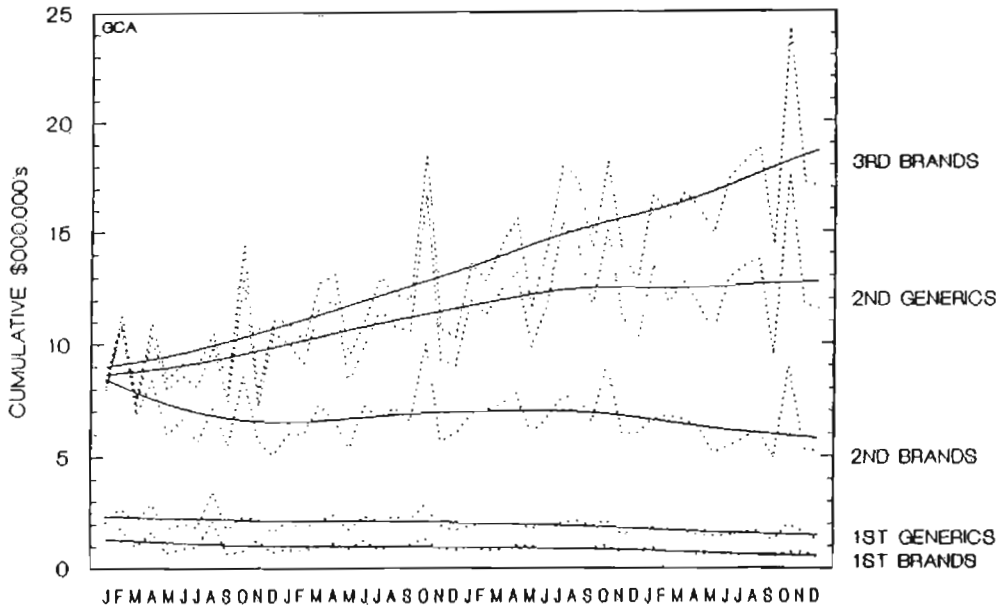
Bars can be placed on line graphs to provide confidence interval estimates on time series data. Figure 9 shows several attributes that are measured using continuous tracking. The daily data have been graphically smoothed in a manner similar to having actually run centre-weighted several-week-window time series. To provide an indication of real shifts in manufacturer perception, this graph also shows confidence interval bars. These are based on the standard error of monthly cumulated data, and presenting them without overlap necessitates a slight starting point shift for each manufacturer. This graph shows that the actual improvements made to "warranty" by "Mazda" have been noticed in the marketplace, and that they have produced some corollary benefit in terms of general "quality" perception. Note the much larger error bars on "pollution control" and how all manufacturers shift together.

Figure 9 Smoothed series with staggered monthly confidence limit bars



Demand curves can be smoothed across observed data values to provide a clearer overall picture. Figure 10 shows market growth over 48 months, stacked across the five groups of competing products. The dotted lines show detailed monthly variation and seasonal effects, while the solid line shows overall trends.

Figure 10 Stacked market share by month line graph and dwls smoothing

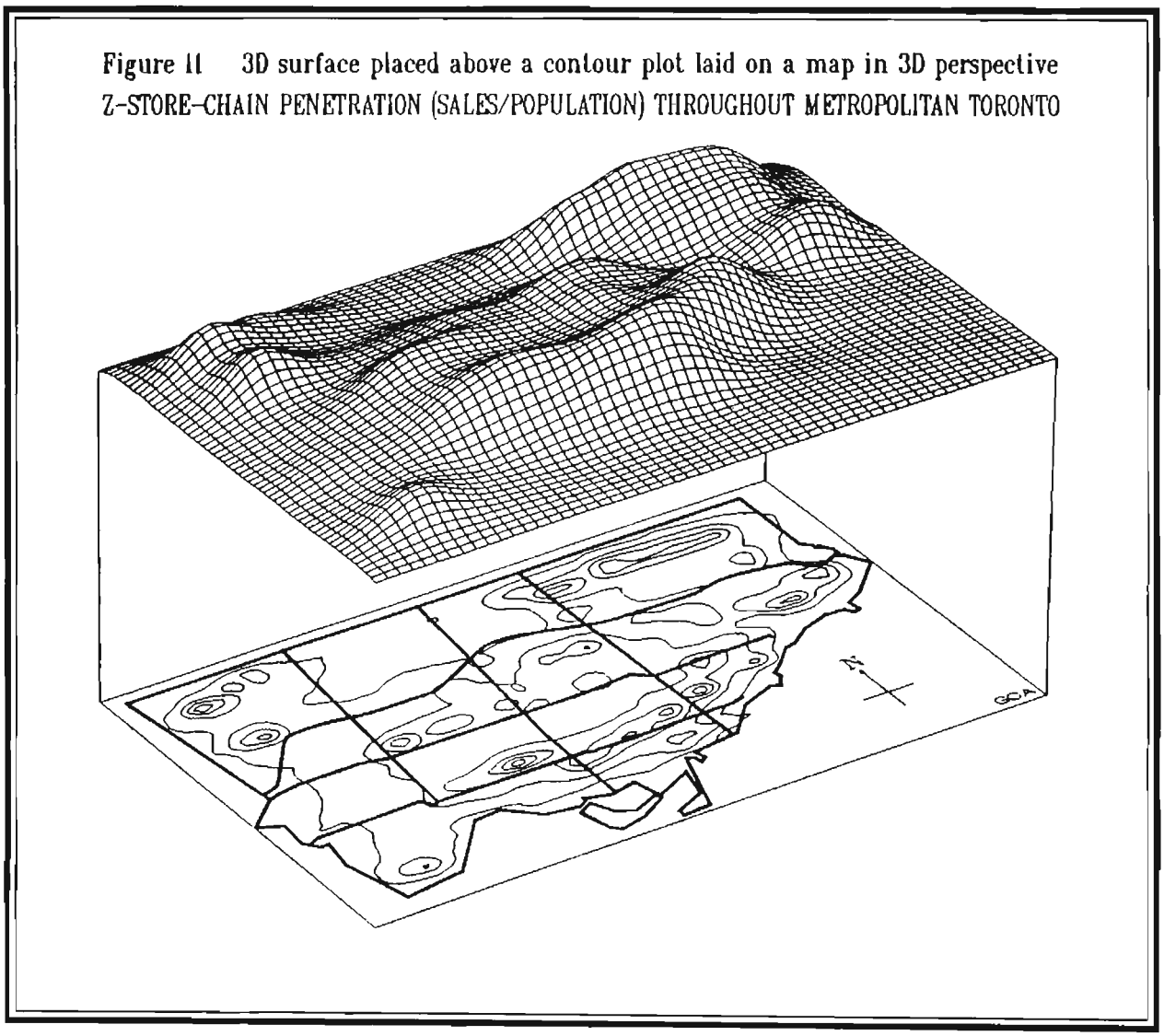


THREE DIMENSIONAL GRAPHS

Three-dimensional graphs, when rotated on the computer screen, provide diagnostic understanding for various multivariate analyses. Segmentation insights are gained by spinning the point cloud of respondent principal components loadings prior to cluster analysis. Factor and discriminant analyses alternate rotations are readily viewed by spinning vectors and scored data.

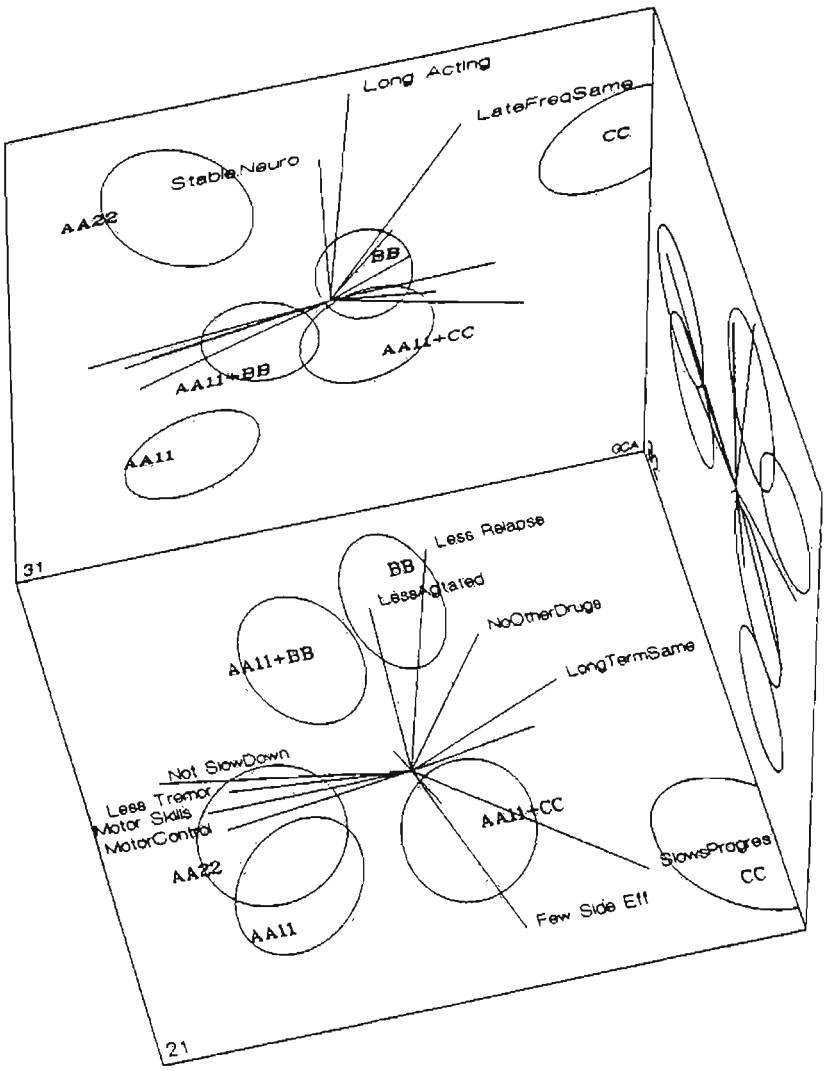
Unfortunately, point and vector positioning is seldom accurately interpretable when a three-dimensional perspective is projected onto a static page. However, three-dimensional surfaces projected onto a page are much more easily recognized. Surfaces are particularly useful when showing how characteristics vary across a geographic area.

Figure 11 shows the sales penetration for a store chain throughout the city of metropolitan Toronto. The hills and valleys of the surface show the areas of greater and lesser customer penetration. The contour plot laid onto the city map provides further detail for geographically locating various penetration levels. For this graph the primary data elements are geocoded latitude and longitude coordinates. These coordinates are then grouped into many small geographic areas. Then the variable of interest is plotted by the latitude and longitude of each geographic area. In this graph, the variable of interest is by-area sales penetration, consisting of cumulative customer sales volume for each area, adjusted by population count for each area. Because the sales data contain sampling error, the variable of interest has been smoothed using distance-weighted least squares for both the surface and the contours. Useful geographic surfaces may show survey data, census data, geophysical data, or some combination of these as in the above example.



Three-dimensional graphs are also useful in explaining relationships among various dimensions in nonmetric multidimensional scaling, correspondence, factor, and discriminant based analysis. Figure 12 shows the relationships among loadings on three dimensions of a perceptual map. However, the representation consists of three two-dimension maps, that are placed in three-dimensional relationship to each other on separate facets of a cube. The viewing angle for this graph is positioned from the upper right of the cube to provide a perspective that portrays the greater variance in the first two dimensions, the lesser variance in the third dimension, and to ignore the dimension three-by-two facet that is already adequately explained in the other two facets. As well, the attributes are only labeled on facets where they provide additional information. When plotting more than the first two dimensions, prior varimax rotation of the dimensions greatly assists in identifying the dimensionality.

Figure 12 Components plotted in 3D perspective onto 2D facets
RELATIVE POSITIONING OF SIX DRUG COMBINATIONS BY PHYSICIANS



SUMMARY

These graphs have been produced and printed using the Systat statistics and graphics software. The more complex graphs are developed in several steps. First the data are assembled, and sometimes the data require transformation. Next the component layers of a complex graph are separately developed. Then the layers are overlaid. Finally, labeling and titling complete the final graph.

Graphs which contain considerable marketing information often should be unfolded in successive layers as well, both in reports and in presentations. Then once the component layers are individually understood, the relationships within the overall graph can become the focus of considerable marketing thought and discussion. Such graphs provide a concentrated focal point for dealing with complex issues.

Generally black-and-white graphs suffice, and they are the medium for most reports, including this conference proceedings. Colour is sometimes used to distinguish among groups when working at the computer screen or preparing presentation slides. These same graphs could be pulled into editing packages for dramatic ad agency presentations, but the basic design concepts remain the same.

ALTERNATIVE APPLICATIONS OF PREFERENCE MODELS TO CUSTOMER SATISFACTION RESEARCH

Carl T. Finkbeiner
National Analysts, Inc.

Customer satisfaction is undeniably a key objective in the development and delivery of high quality goods and services. The measurement of customer satisfaction and the use of those measures in planning and execution are considered fundamental to the practice of Total Quality Management (TQM). The Malcolm Baldridge National Quality Award lists Customer Focus and Satisfaction as one of its core criteria, with 22% of point values in the examination items being associated specifically with the assessment and analysis of customer satisfaction (National Institute of Standards and Technology, 1992). And the Profit Impact of Market Strategy (PIMS) study is often cited as evidence that perceived quality is correlated with higher profits: for competitors viewed as superior by customers, the return on sales and return on investment are both about twice that of competitors viewed as inferior (Buzzell & Gale, 1987).

As a consequence of its importance, customer satisfaction measurement and modeling has captured a great deal of attention. Customer satisfaction measurement alone is of use to a business that wishes to know more about how it is perceived by customers. However, by itself, measurement is not diagnostic. You know very little about what you should do to improve whatever level of customer satisfaction you have achieved, unless you add something to the measurement. A very potent addition is the use of a model to help prioritize components of performance and to assess the potential consequences of modifying that performance.

Although customer satisfaction modeling has recently taken center stage because of its role in such management disciplines as TQM, Quality Function Deployment, and the House of Quality, it is not really new at all. In fact, since satisfaction is a form of preference, customer satisfaction modeling is nothing more than preference modeling dressed up in a different language. For decades, variations on preference models have been used extensively in market research and related fields such as psychology:

- Preference scaling (Thurstone, 1927; Coombs, 1964)
- Choice modeling (Thurstone, 1945; Bock & Jones, 1968; Luce, 1959)
- Expectancy-value modeling (Rosenberg, 1956; Fishbein, 1967)
- Preference regression (mentioned in Bass & Wilkie, 1973, and Cooper & Finkbeiner, 1983, though applications were common long before either reference)
- Conjoint analysis (Debreu, 1960; Luce & Tukey, 1964)

Because customer satisfaction modeling has roots in so many different areas, it can be confusing as to just what is being modeled and then as to what can be done with the model once you have it.

The purpose of this paper is to (hopefully) reduce some of this confusion by considering a number of common preference modeling approaches to customer satisfaction and evaluating them on a set of criteria. In some cases, the evaluations will lead to some suggestions for extending the approaches.

My intention is not to be exhaustive here, but rather to cover some of the more popular current approaches. My hope is that, after reading this paper, the user of customer satisfaction models will be better able to evaluate the applicability of a particular preference modeling approach for his or her purposes in a specific context.

OBJECTIVES OF CUSTOMER SATISFACTION MODELS

One or more of the following objectives pertain in most practical applications of customer satisfaction modeling.

1. Assess performance of a company(/product/service) from the customers' point-of-view. This assessment is a determination of how customers feel the company compares against some standard (either an abstract standard or relative to other companies). This includes both an overall assessment as well as perceptions along important dimensions of performance and service.

2. Establish priorities among aspects of product and service delivery. Perhaps the single most important outcome of customer satisfaction modeling is information about what the company must work on to improve satisfaction optimally. Priorities may be established either in the abstract or within the constraints of the costs to achieve the improvements.

3. Predict changes in customer satisfaction as a result of performance changes. An effective way to establish priorities is to assess the direction and extent of change in customer satisfaction resulting from particular proposed changes in performance. Prediction of the consequences of hypothetical change is a useful component of any model and, for present purposes, is taken to be a defining characteristic of a customer satisfaction model.

   Along with the ability to predict hypothetical changes comes the ability to estimate degradation in satisfaction due to a slackening in performance (risk) as well as the magnitude of increased satisfaction arising from improved performance (opportunities).

4. Identify competitors' weaknesses and advantages. The relative performance of competitors (overall and on key dimensions) as well as their risks and opportunities is necessary for a complete assessment of a company's position in a market.

5. Link concrete company actions with customer perceptions of performance. Once customer perceptions of performance on key dimensions are established, along with the relationship between these perceptions and overall satisfaction, it is also useful to establish associations between the perceptions and concrete objective actions which a company knows how to take. It is difficult, if not impossible, for a company to take steps to improve satisfaction based on "soft," broadly specified attributes, such as "Expensive" or "Responsive"; it is far better to base actions on more concrete attributes such as "Raise the price $10" or "Decrease response time to requests by 4 hours."

6. Predict customer behaviors as a result of performance changes. Just as concrete company actions must be linked to customer satisfaction, so is it also helpful if concrete customer outcomes are linked. If we can associate customer behaviors (such as choices, complaints,

and usage changes) with a complete satisfaction model and if costs can be assigned to product or service changes, it becomes possible to conduct cost-benefit analyses for various company actions.

In the following sections, a series of approaches to measuring and modeling customer satisfaction will be described and then judged in terms of whether or not, and how, they accomplish the above objectives. To provide a context within which to illustrate these approaches, the following example problem will be used, where possible.

EXAMPLE PROBLEM

A manufacturer of metal alloys (not the real product category) wishes to better understand its customers' views, needs and wants with respect to the services it provides. Since the product here is essentially a commodity manufactured to the specifications of the customer, customer services are the main aspect which the manufacturer wishes to investigate, although a few price and product quality issues are also included. A sample of 103 customers — buyers at manufacturing companies using the alloys in their manufacturing processes — completed a mail questionnaire.

Five companies account for nearly all of the product sold in this category. These companies are referred to as A - E, with company A being the sponsor of the research.

The primary measure obtained in the questionnaire is an overall satisfaction rating for companies A - E on a 1 - 10 scale with 1 being "Extremely Dissatisfied" and 10 being "Extremely Satisfied." The mean ratings for the five companies are:

OVERALL MEANS FOR 5 COMPANIES

Company	Mean Overall Satisfaction
A	6.62
B	6.35
C	6.06
D	7.04
E	6.63

While these data are of immediate interest, they are not at all helpful to company A in determining how to surpass companies D and E. For that information, more is needed.

Thirty-four attributes were identified as being of relevance to company A, including such things as "Responsiveness to inquiry for assistance during customer's product development," "Fairness of price changes," "Consistency of product quality during customer's manufacturing," and "Technical support of customer's marketing."

It is typically the case that the attributes of interest to the sponsor of research may be organized into a natural hierarchy. For example, we found in prior research that the broad category of "Supplier rep helpfulness" (at the primary level in the hierarchy) was comprised of more specific attributes at the secondary level of the hierarchy, such as "Frequency of rep contact," "Technical knowledge of rep," "Coordination of supplier's support services," and "Frequency of rep changes."

There are a number of approaches to formulating this hierarchy and to establishing which types of attributes are at the primary, secondary, and, in some cases, even at a tertiary level. Coverage of these approaches is not within the purview of this paper. Suffice it to say that this hierarchy is useful in organizing the large number of attributes often identified in these studies.

Since there is some risk of unwittingly (and unnecessarily) creating over-much redundancy if attributes are mixed at very different levels of specificity, the practice of creating and using a hierarchy of attributes also enforces a discipline on the research designer which reduces that risk. We typically select about 5 - 10 primary attributes and 2 - 10 secondary attributes for each primary. In the present example, we had 8 primary attributes with 2 - 6 secondary attributes for each of the 8.

We often find it useful to obtain ratings of satisfaction with each competitor on the primary level attributes, along with ratings of importance of those attributes to the customer. Together, these data provide a relatively simple first cut at performance pluses and minuses, presenting a far less complex picture than that of the more detailed secondary attributes. Furthermore, as we shall see, breaking overall satisfaction into its component parts can be helpful in organizing and guiding subsequent research. In our example problem, given the purposes of the client and the expected respondent burden, an independent decision was made to not obtain ratings on the primary attributes. Nonetheless, we can still use the example to illustrate most of the modeling approaches in this paper.

There are (at least) three general types of measures which are commonly useful for secondary level attributes. Parallel to the primary attributes, we obtain importance ratings and satisfaction ratings for each competitor on each secondary attribute. It is also useful to obtain measures of level of performance on each secondary attribute, as well. That is, in order to be able to act on an attribute rating for your company, it is helpful to know in concrete terms where customers perceive your performance to be, not just how happy they are with that performance.

A useful adjunct to the above ratings is some version of conjoint analysis, particularly if you are contemplating a performance change that is not currently present in the market. The disadvantage of conjoint in this context lies in its inability to easily handle the large number of attributes typical of most customer satisfaction studies.

To illustrate from our example, consider the secondary level attribute "Frequency of rep contact." For this attribute we obtained the following ratings:

- "Importance to Your Overall Satisfaction": a rating of this attribute on a 1 - 10 scale with 1 being "least" and 10 being "most" important

- "Level of Performance of Key Suppliers": a rating of "Once a week," "Every two weeks," "Once a month," or "Less than once a month" for each of companies A - E
  (Note the level of performance measure is stated in fairly specific concrete terms)

- "Satisfaction with Each Level of Performance": a rating on a 1 - 10 scale (with 1 being "Extremely Dissatisfied" and 10 being "Extremely Satisfied") of satisfaction with each level of performance on the attribute (note that analogous information, combined with importance scores, is used by ACA (ACA System by Sawtooth Software) to get initial partworths that are sometimes used as final partworths, as in the case of telephone-administered ACA

- "Satisfaction with Supplier Performance": the same 1 - 10 scale rating as above, but on each of companies A - E (in the present example, this rating was inferred from the preceding two ratings, for example, if company A is rated at "Once a week" by a respondent and if "Once a week" received a 7 rating of satisfaction, then we conclude that the respondent would rate company A at or near 7 for satisfaction on this attribute)

Thus, from each respondent, we obtained one importance rating on "Frequency of rep contact," a level of performance rating for each company with which the respondent was familiar (up to 5), a satisfaction rating for each of the four attribute levels, and a measure (or derived score) of satisfaction for each company. This yields a total of 15 possible scores for this one attribute, with the first 10 being actual measures and the last 5 being derived.

We usually do not attempt to obtain level of performance ratings on the primary attributes. These attributes are generally "soft" and do not lend themselves sensibly to level ratings. For example, the primary attribute corresponding to the "Frequency of rep contact" attribute is "Supplier rep helpfulness." It is difficult to imagine any measure of a company's performance on this primary attribute which isn't highly correlated with (and hence redundant with) a satisfaction rating.

Note that these aren't the only ways to obtain these evaluations from respondents. It is not uncommon to use rankings instead of ratings for any or all of the above ratings. In addition, an almost infinite variety of psychological scaling techniques have been used, such as Thurstone Case V paired comparisons, categorical judgment scaling, magnitude estimation, and constant sums.

And, besides the methods described above,

- Importance is sometimes measured:
  - In absolute terms ("Not at All Important" to "Extremely Important") rather than relative terms ("Least Important" to "Most Important")
  - By reference to specific levels of performance, as in ACA, for example, "How important is the difference between 'Once a week' and 'Less than once a month'?"
  - Via conjoint, regression or other analyses of satisfaction ratings, such as "Benefit/ Loss" analysis to be discussed later

- Level of performance is sometimes measured using:
  - Agree/Disagree ratings of one (usually) extreme level of performance, for example, "Rep contacts me frequently"
  - "Degree" scales, for example, a 1 - 10 scale where 1 means "Infrequently" and 10 means "Frequently"

- Satisfaction with levels of performance can be measured using:
  - Full profile conjoint analysis
  - Pairwise partial profile comparisons (for example, ACA)
  - Paired attribute tradeoff analysis
  - Discrete choice analysis

- Satisfaction with a company's performance on an attribute is sometimes assessed:
  - By reference to an "ideal" or "expected" level of performance, for example, have customer rate the "degree" of frequency of rep contact and then rate the "degree" of frequency at which he/she would like that contact ideally
  - By another variant on satisfaction ratings, namely, "Minimal Performance" ratings; for example, after having rated the level of performance of companies A - E on "Frequency of rep contact", have the respondent rate the level of performance, if any, at which he/she would carry out some specific negative action, such as complaining to the Better Business Bureau or switching to another brand

It has been my experience that, so long as you exert some common sense in choosing an approach to measurement of these four types of information, there is a certain amount of acceptable flexibility in that choice. I find that the measures we used in the above example are as valid and reliable measures of importance, level of performance, and satisfaction with performance levels as you are likely to find for the purpose of modeling customer satisfaction. In many studies, we also have used direct ratings of satisfaction with attribute performance and have generally found the results quite useful.

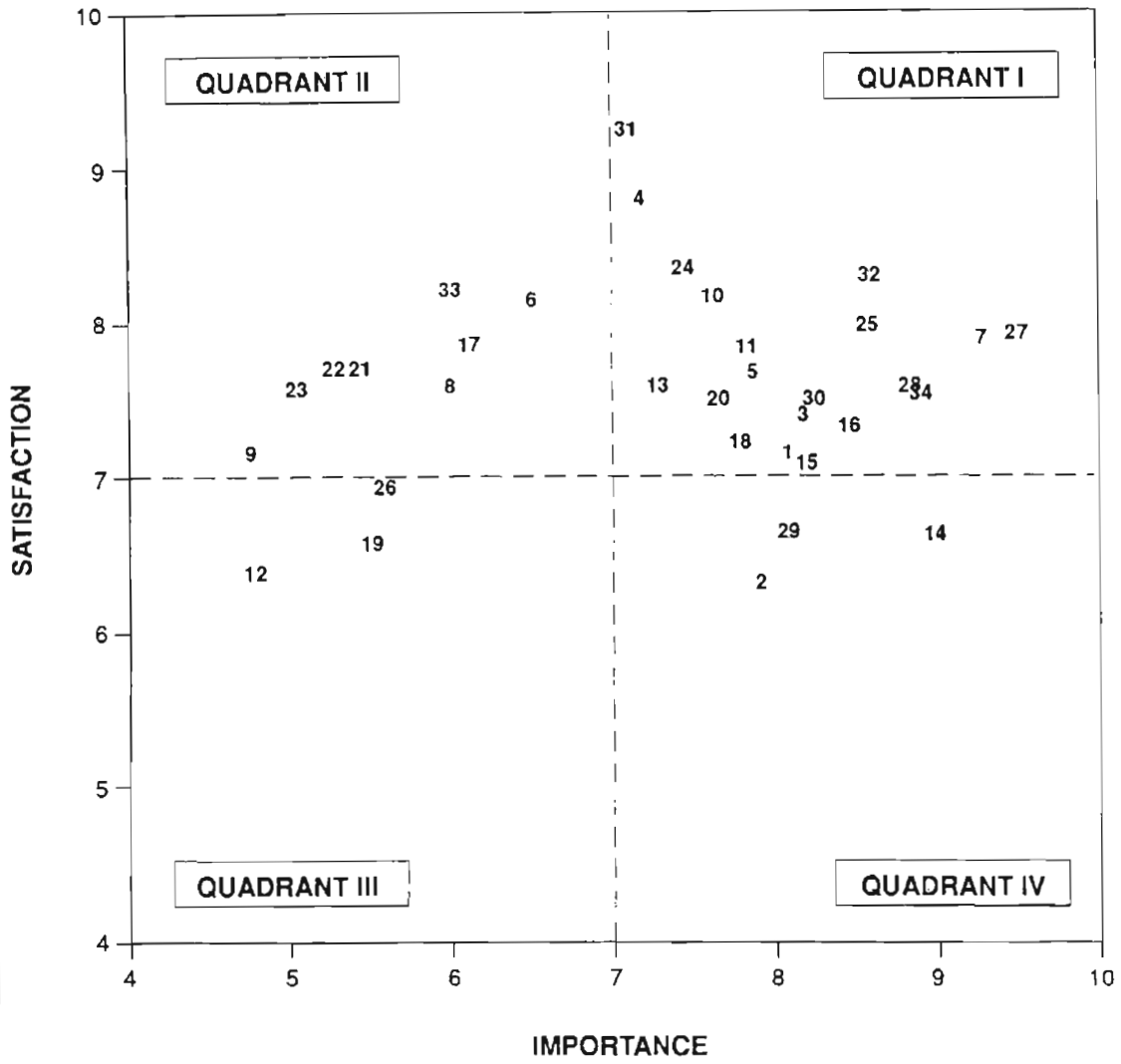
However, our methods, while not unusual, are also not entirely typical. After a review of materials from about 20 market research and consulting firms, it appears to me that the most common practice in the industry is to obtain (in addition to an overall satisfaction measure) two basic measures on each attribute: absolute ratings of attribute importance and direct ratings of satisfaction with each company's performance on the attributes. In my opinion, this is a fairly minimal and potentially limited (hence, limiting) information base. Absolute importance ratings tend to be clustered at the high end of the scale for all attributes: after all, given that the number of attributes is usually large to begin with, we do not purposefully include anything we know to be unimportant. And, while we have found direct ratings of satisfaction to be quite useful in many studies, they do not give a complete picture of performance, it being difficult for a company to relate satisfaction ratings to past or possible future actions.

Having now described the example problem to be used for illustrative purposes, as well as the alternative kinds of data which are sometimes collected, we turn to a discussion of some approaches to modeling customer satisfaction.

SATISFACTION/IMPORTANCE CHARTS

These charts represent the most common analysis performed on customer satisfaction surveys, although, as we shall see, the charts aren't really a modeling approach as such. The ratings of satisfaction with attribute performance are paired with the corresponding importance ratings and plotted against one another. Figure 1 shows the results for company A. Such plots may be created for each of the five companies.

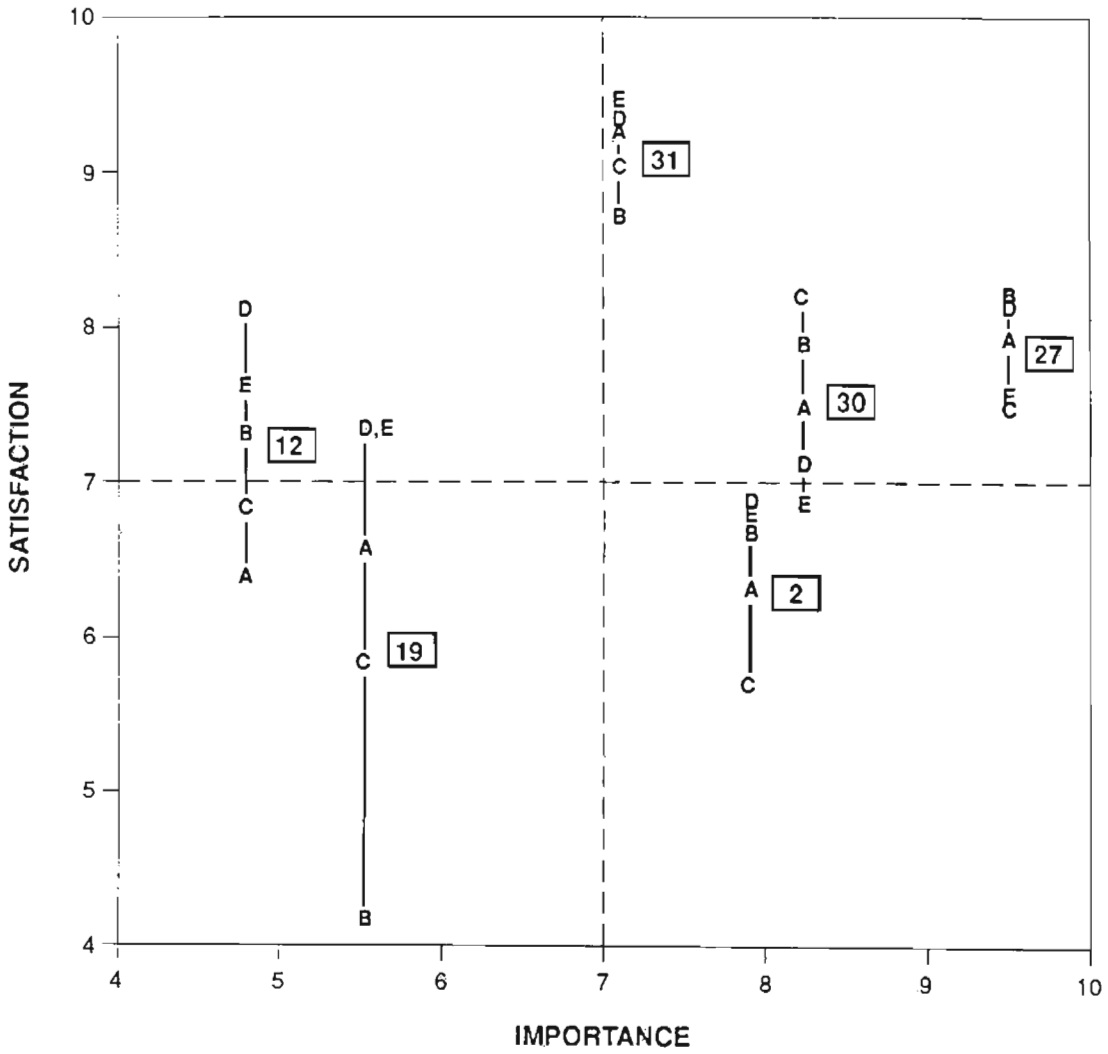
Figure 1
SATISFACTION/IMPORTANCE CHART
COMPANY A -- ALL ATTRIBUTES



This plot is then divided into quadrants, usually based on some midpoint of ratings, as here. These quadrants are interpreted to identify vulnerabilities and opportunities. Attributes in quadrant I are important, but performance is already high on those; attributes in quadrant II may represent overkill, having high performance on less important attributes. Although performance is lower on quadrant III attributes, they may be easy to give up on since they are also less important. Depending on the performance of competitors, quadrant IV attributes may represent either opportunity or vulnerability for company A, being more important attributes on which A is not performing as well.

Comparisons across companies may be made by plotting all companies on a single plot. Since these plots would get very cluttered with all attributes represented, usually only a few attributes are selected for one plot, as in Figure 2.

Figure 2
SATISFACTION/IMPORTANCE CHART
ALL COMPANIES -- SELECTED ATTRIBUTES



Not surprisingly, the companies are tightly grouped on attribute 27 (a product quality attribute), as we would have expected of the main competitors in a commodity market. This attribute is also a given for success in this market, being the most important of all attributes. But it looks like there is an opportunity for company A here, since its rating is 7.9, with the highest rated product at 8.2, and since there is room at the top of the scale for improvement. Of course, product quality in a metal alloy is one of the more difficult attributes to improve on, but it is not clear that available technology has topped out on this particular aspect of quality.

Attribute 31 (ordering time) is an example of an attribute on which it is easy to perform well, and so everyone does, even though its importance is only in the mid-range. Attribute 19 (technical information support for sales and marketing) is one of the less important attributes and shows tremendous spread on companies' willingness to provide this assistance. One strategy for company A might be to try, through marketing and education campaigns, to increase the importance of attribute 19 among customers.

The best company, D, generally is seen to do even the little things right, except on attribute 30 (price changes). Company A could improve on most attributes, but attribute 2 (timeliness), represents an opportunity for A: no one is performing superbly on this moderately important dimension and A is not that far from the best companies now.

The above sampling of interpretations illustrates the kind of richness available from these charts.

In an attempt to convert the two satisfaction and importance ratings into a single index of opportunity, the two scores are sometimes multiplied together. These opportunity scores are then used to prioritize the attributes by sorting them numerically. This is risky business because rankings derived from the product of two numbers depends on the scaling used for each separately, as is shown in the following example.

Consider two attributes, X and Y, where X has the lowest possible importance and Y the highest, but where a given company obtains the highest possible satisfaction on X, and the lowest on Y (see the left panel in the table below). In other words, X is in quadrant II (upper left) and Y is in quadrant IV (lower right) of the satisfaction/importance chart, indicating that the company may have overkilled on X but is underperforming on Y. Measuring importance and satisfaction with 1 - 10 scales as we did in the present example, X and Y would both have opportunity scores of 10 (that is, 1 x 10 and 10 x 1, respectively), as shown in the left panel in the table below. There is no reason why we could not have used a -5 to +5 scale for rating satisfaction, as in the middle panel, or a 0 - 100 scale for importance as in the right panel below. And opportunity scores would then be calculated by multiplying corresponding importance and satisfaction scores.

EFFECTS OF RESCALING ON OPPORTUNITY SCORES

	Both 1 - 10			Sat. -5 to +5			Imp. 0 - 100		
	Imp.	Sat.	Opp.	Imp.	Sat.	Opp.	Imp.	Sat.	Opp.
X	1	10	10	1	5	5	0	10	0
Y	10	1	10	10	-5	-50	100	1	100

As can be readily seen, the opportunity scores for X and Y indicate quite different rankings in all three cases. Which one is right? It depends upon how much weight you wish to give importance versus satisfaction in establishing priorities. One resolution of that dilemma is to use that scaling which correlates best with overall satisfaction ratings: that is, preference regression, to be treated later. Another resolution is to turn to "benefit/loss charts" or to conjoint analyses for prioritization of attributes. These techniques will also be discussed later.

In many ways, however, it is perfectly acceptable from the point of view of identifying opportunities for service improvement to simply stick with the satisfaction/importance charts. These charts have the great advantage of being simple to understand and of not straying far from the raw data. They clearly present an assessment of the company's perceived performance, as well as that of its competitors.

While it is possible (though tricky, as mentioned above) to establish priorities among attributes using this approach, it is incomplete as it stands. Because it is not a predictive model, the approach cannot estimate changes in overall satisfaction resulting from performance changes on attributes. And because of that fact, even when costs for carrying out performance changes are available, it is not possible to conduct a formal cost-benefit analysis. Such analyses are highly desirable when establishing priorities regarding which attributes are worthwhile for the company to address and which are not.

The following chart indicates my judgments as to the ability of satisfaction/importance charts by themselves to achieve the six objectives described earlier.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts
1. Company Performance Assessment	+
2. Attribute Prioritization	•
3. O'all Sat. Resulting from Perceived Perf. Changes	-
4. Competitor Risks & Opportunities	+
5. Co. Actions Link to Perceptions	-
6. Behaviors Resulting from Customer Sat.	-

+ = Achieves - = Fails • = Partial

The difficulties in establishing priorities based on the satisfaction/importance charts has led some researchers to another type of analysis.

BENEFIT/LOSS CHARTS

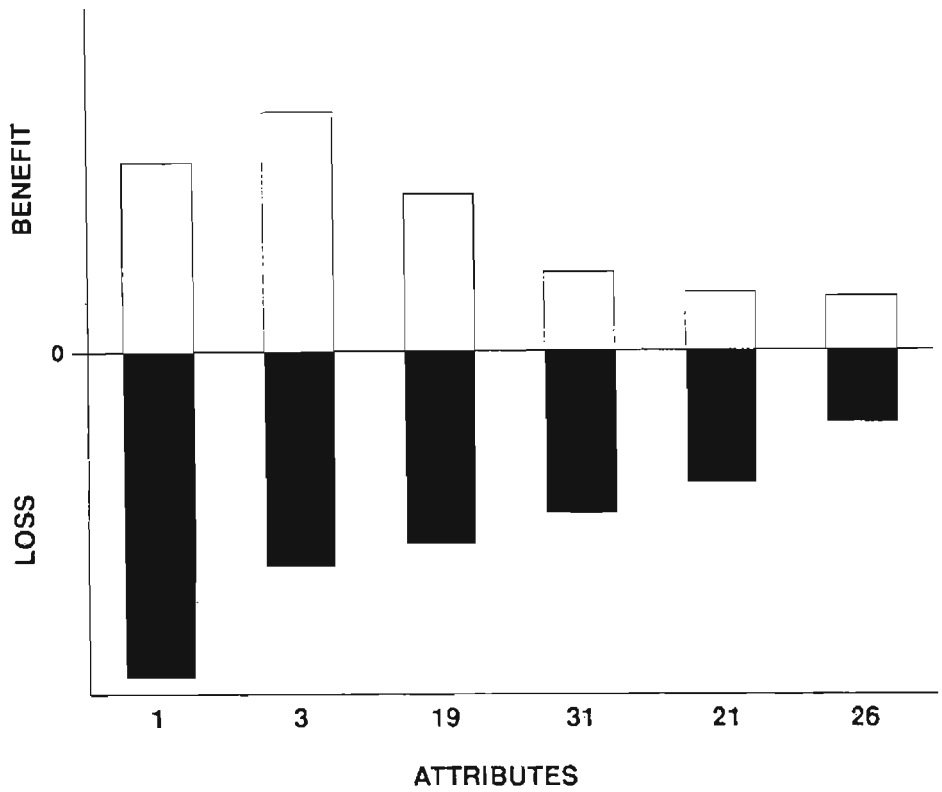
These charts go under a variety of names, including "risk/reward analysis" and "penalty/reward analysis," although you should be aware that not all graphs which look like benefit/loss charts are actually computed in the way I am about to describe.

This analysis uses only the overall and the attribute satisfaction ratings. Some aggregated index of overall satisfaction is obtained for two sets of respondents: those rating the company above and those rating it below "average" on an attribute. The aggregate index for the above average group is called the "benefit" and the index for the below average group is called the "loss." This pair of indices is computed for each attribute. The calculation can be done using means, medians, or end box percentages (the percent in the top box for the above average set or the percent in the bottom box for the below average set). Of course, as with any scale rating, the choice of measure of "averageness" for the attribute and of aggregated overall satisfaction may have an effect on the conclusions drawn. In some cases, a measure of "expected performance" is used instead of the "average."

Judgment must also be exercised in deciding whether to carry out the calculations across all companies rated, or for selected companies separately. That choice may depend on sample size considerations. Furthermore, if the analysis is on one company at a time, then, for your company, you are able to represent only the relationships between whatever attribute performance you are now providing and the overall satisfaction which currently exists given that performance. Broadening the analysis to include the entire market at least represents a wider range of performance and so should be better able to represent opportunities.

To illustrate with our metal alloy example, we use means and we use the ratings across all companies. For instance, of the 359 ratings of companies by customers on attribute 1 (responsiveness), we find that 227 ratings were at or above the mean and 132 were below the mean for that attribute. The above average group had mean overall satisfaction ratings of 7.4 and the below average group mean overall satisfaction was 5.0. Since the grand mean is 6.6 for these groups, we might conclude that being above average on responsiveness results in a .9 increase in mean overall satisfaction (the "benefit"). But, being below average on responsiveness results in a 1.6 decrease in mean overall satisfaction (the "loss"). We can compare this result to that of several other attributes in the benefit/loss chart shown in Figure 3.

Figure 3
BENEFIT/LOSS CHARTS



We see here that attribute 1 apparently is strongly associated with overall satisfaction in that this attribute has one of the highest "benefits" and the highest "loss." Furthermore, we see that its benefit and loss are asymmetric, with substantially more loss associated with being below average than the benefit associated with being above average. This is most markedly the case with attributes 21 and 31. Attribute 3 (technical support) has slightly more benefit for being above average than loss for being below average.

The benefit/loss charts are often used to assess the impact potential of attributes for the purposes of prioritizing them. They usually provide somewhat different results than the direct importance ratings. For example, whereas attribute 27 (a product quality attribute) received the highest direct rating of importance, its impact from the benefit/loss charts is far from the highest (not shown in Figure 3, but approximately the same as attribute 21). Conversely, where attribute 1 (timeliness) has the highest impact in the benefit/loss charts, it ranks 15th out of 34 on the direct importance ratings.

Which one to trust? Intuitively, I would believe the direct ratings first. While direct ratings have a tendency toward motherhood-and-apple-pie (that is, socially acceptable responses), they at least

allow the respondent to indicate importance unconstrained by the world as it is currently, with companies performing at their present levels. As indicated before, benefit/loss charts are limited to the world-as-it-is for its inferences about the world-as-it-could-be.

In addition, I have also observed that it is easy for people to misinterpret benefit/loss charts as providing a prediction of the consequences of changes in performance, for example, to assume that a one unit increase in performance on attribute 1 (Figure 3) produces less improvement in overall satisfaction than the decrease in overall satisfaction which would result from a one unit decrease in that attribute.

Consider a very simple example in which 60% of respondents rate both an attribute and overall satisfaction at 3; the remaining 40% rate both at 7, with a grand mean of 4.6. Note that we have constructed a perfectly linear relationship between the attribute and overall satisfaction — the correlation between the two is exactly 1.0, implying that a one unit increase in the attribute rating yields a one unit increase in the satisfaction rating and a one unit decrease in the attribute rating yields a one unit decrease in the overall satisfaction rating.

This is a very symmetrical relationship: the effect of a change in one direction is identical (in magnitude) to the same change in the opposite direction. However, carrying out the benefit/loss calculations indicates an asymmetry in the relationship: in this sample, being above average yields a 2.4 point improvement in overall satisfaction (7 - 4.6), but being below average yields a 1.6 decrease (4.6 - 3).

The benefit/loss calculations merely show the overall satisfaction presently associated with being above or below average on an attribute. Because ratings above average on the attribute are not necessarily the same distance above as the below average ratings are below, benefit/loss charts don't necessarily provide a true picture of the relationship between overall and attribute satisfaction. This holds regardless of which measure of "averageness" or whether means, medians or percentages are used. The only way to correct this shortcoming is to construct benefit/loss charts from calculations estimating the change in overall satisfaction resulting from a change that is "equivalent" (for example, a one unit change, or a 10% change) across both benefit and loss and across attributes, as well. This is accomplished via some form of predictive model, such as regression or choice modeling, but since that is essentially a different approach (to be treated in the next section) using the same type of graph to display results, we won't consider it to be a pure form of benefit/loss analysis as defined here.

While benefit/loss charts are relatively easy to construct and they carry with them the data for assessment of company performance, they are so easy to misinterpret and misuse, that they should probably not be used for establishing priorities. And, as with satisfaction/importance charts, this approach is not a predictive model, so results of change in performance cannot be simulated and formal cost-benefit analyses are not possible. As usually implemented, benefit/loss charts also provide no competitive analysis.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts	Benefit/ Loss Charts
1. Company Performance Assessment	+	+
2. Attribute Prioritization	•	•
3. O'all Sat. Resulting from Perceived Perf. Changes	-	-
4. Competitor Risks & Opportunities	+	•
5. Co. Actions Link to Perceptions	-	-
6. Behaviors Resulting from Customer Satisfaction	-	-

+ = Achieves - = Fails • = Partial

The above table represents my opinion that, in general, satisfaction/importance charts are safer and more informative than are benefit/loss charts. (I must hasten to add here that not all graphs that look like Figure 3 are based on the same calculations as I have described — the preceding caveat applies only to benefit/loss charts as described herein.) Kept in proper perspective, benefit/loss charts may be useful tools at times. However, they do not satisfy the need to represent the effects of changes in attributes.

PREFERENCE REGRESSION

Using exactly the same data as are used in benefit/loss charts, we turn now to the very common use of regression analysis applied to preference data. Overall satisfaction is, of course, the dependent variable and satisfaction with performance ratings are the explanatory variables. The basic model is:

$$y_i = b_0 + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_K \cdot x_{iK} + e_i$$

where y_i is overall satisfaction with a company for observation i, b_k is a regression coefficient for attribute k derived in the analysis, x_{ik} is the satisfaction rating on attribute k for observation i, and e_i is the error term. All of the machinery of ordinary regression is available, including goodness-of-fit

assessment and significance testing. I will assume here that the reader is familiar with regression analysis, but if not, there are a number of good introductory texts on the subject, such as Cohen & Cohen (1975).

As with benefit/loss charts, regression models may be estimated for each company separately or across all companies together, treating the respondent's ratings of a company as a single observation, with the respondent potentially contributing more than one observation to the analysis. The considerations in this choice are much the same as before: inadequate sample size may eliminate the option, and a single company cannot provide as much variation as when all companies are considered together, thereby limiting the coverage of the regression model.

Of course, when the same respondent contributes multiple observations, the observations may no longer be independent, creating the possibility of correlated errors. Such a situation violates a common assumption in the significance testing usually applied to regression, although the basic regression model itself does not require independence to be applicable.

We note that Agree/Disagree ratings or "Degree" ratings on attribute performance can also be used in regression. However, I find that those who commonly use regression don't often use these types of measures because they don't predict as well as do satisfaction ratings on the attributes. This is probably due, at least in part, to "halo" effects in the satisfaction ratings. It is also due to the fact that linear models are probably better suited to attribute satisfaction ratings: more of an attribute (in degree) is not necessarily linearly better (for example, more sales contacts may be irritating), but more satisfaction on that attribute is generally uniformly better.

As an aside, we acknowledge that there are many variants of preference regression which will not be explicitly covered here. For example, some form of direct or implicit regression analysis is used in perceptual mapping to project overall satisfaction into a map. Also, discrete choice models, such as logit and probit, may be thought of as types of regression models in which the dependent variable is choice, rather than overall satisfaction. Whether the models are linear or non-linear, metric or monotonic, many of the same issues hold as for ordinary preference regression.

The great attraction of preference regression is that it derives the impact of an attribute on overall satisfaction (that is, the regression coefficient) from the data, rather than asking the respondent for a direct measure of impact, as in importance ratings. The coefficient for an attribute is directly interpretable as the amount of change in overall satisfaction predicted from a one unit change in attribute satisfaction, all other attributes' satisfaction ratings being held constant. Attribute priorities are often established on the basis of regression coefficients (frequently indexed or percentaged or rescaled in some way), and a host of approaches to presenting or using these coefficients is adopted. One common use is to multiply each coefficient (or some rescaling of it) by the mean attribute satisfaction rating, providing, in effect, that component of the mean overall satisfaction rating due to each attribute.

To illustrate preference regression with a relatively small example, we calculate the regression model using a subsample comprised of a particular kind of respondent. This produces 133 observations of ratings of a company by a customer. While this sample size is not large, it is enough for illustrative purposes.

The squared multiple correlation, R^2, is .72 in this sample. The associated F-test shows this to be highly significant statistically, although this statistic is probably affected by the non-independence

of the observations. That value of R^2 would usually be taken as quite acceptable in customer satisfaction research.

However, a better measure of the goodness of fit of the model is provided by the so-called "jackknife cross-validity coefficient," P^2. To compute this measure, each observation is dropped from the sample one at a time, the model is re-estimated without the observation in the sample and then applied to the dropped observation's x values to produce a predicted value for y for that observation. This process is repeated for each observation in the sample and the resulting predicted values are correlated with the actual y values. The square of that correlation is P^2. In effect, P^2 represents the ability of the model to predict for other samples besides the present one, that is, to cross-validate.

In our example, P^2 is .35, indicating considerably less validity than might have been suggested by the R^2 of .72. This lack of stability is further driven home by the regression coefficients themselves.

SELECTED REGRESSION COEFFICIENTS

Attribute	Coeff.
1	.15
2	.20
3	-.00
12	.09
19	.00
21	-.06
26	.07
25	.33
27	-.03
30	.01
31	-.14

This list shows a rather disturbing and completely counter-intuitive result: negative coefficients. A negative coefficient implies that increasing attribute satisfaction will result in a decrease in overall satisfaction. Furthermore, priorities established directly from these coefficients are not consistent with any other method nor always consistent with common sense. For example, attribute 27 (product quality) was rated as the most important attribute, even in this subsample, but is here shown to actually have a negative contribution to overall satisfaction. Remembering that the ratings on attribute 27 were generally on the high side, multiplying the negative coefficient by the mean rating for, say, company A will only exacerbate the problem.

Because of the relatively small sample size here, few of these coefficients (and none of the negative ones) are statistically significant. However, apart from the statistical test's independence assumption being violated in these data, the simple fact that there are negative coefficients renders the predictions of this model implausible. I would hate to try to tell a client of mine that increasing product quality will lead to a decrease (even a small one) in overall satisfaction. Furthermore, with larger sample sizes, I routinely find statistically significant negative coefficients when I attempt a straightforward preference regression model such as this one. Forcing the negative coefficients to zero and re-estimating the model doesn't help much either: new negative coefficients invariably crop up and we eventually drive the R^2 to very low levels while eliminating many variables from the model.

The cause of this problem is common in preference regression: multicollinearity. Multicollinearity arises when the explanatory variables are correlated with one another, as is often the case with attribute satisfaction ratings. Multicollinearity is usually measured with an index called the "condition number" (Belsley, 1991). As a rule of thumb, a condition number under 5 is preferred, and a value under 30 is required in order to avoid multicollinearity problems. In our example, the condition number is 180.

The possible sources of multicollinearity are numerous, including non-independence of observations, the "halo" effect, and the natural association of attributes in marketed products. Shrinkage estimation (Farebrother, 1978) — such as ridge regression (Hoerl & Kennard, 1970) or principal components regression (Massy, 1965) — is the preferred solution to the problem, resulting in a reduction in goodness of fit and a flattening of the regression coefficients (although at least P^2 doesn't deteriorate so dramatically). However, shrinkage estimation is complex and not readily available or usable in most statistical packages. Even when shrinkage estimation is used, we are still left with other deficits of preference regression, to be described in the remainder of this section.

For either objective of establishing attribute priorities or of predicting overall satisfaction, the pure, unaltered preference regression model applied to naturally occurring data is unsafe. With proper caution and adjustment, however, preference regression may sometimes be used to predict overall satisfaction. If the regression model is linear, changes in attribute satisfaction can be simulated by changing the attribute mean, both for your company or for your competitors. If the model is non-linear, other changes in the distribution of attribute satisfaction, besides its mean, must also be specified in order to simulate performance change.

The use of the model for predictive purposes must be kept in proper perspective, however. Limitation of the data to the world-as-it-is creates just as much of a restriction (if not more) for preference regression as it does for benefit/loss charts. Because of this limitation to the world-as-it-is, extrapolation of attribute performance beyond whatever is represented in the marketed products is not safe. For example, if no differences exist between companies on price, then price cannot account for any of the variance in overall satisfaction. If, however, one company changes its price significantly from all the others, then price will certainly become an important variable.

If no company is distinguishable from any other on an attribute, leading to a zero relationship between the attribute and overall satisfaction, should we really conclude that the attribute is unimportant and that we can ignore it? Particularly if customers give a high direct rating of importance to the attribute and leave room for improvement in their satisfaction ratings for that attribute, doesn't this sound like a strong opportunity?

Consider attribute 27 (one aspect of product quality) in our example problem: it has the highest rated importance at 9.5 and about the least spread across companies on mean attribute satisfaction ratings (a range of .6 on the 1 - 10 scale and not statistically significant), with a regression coefficient of -.03. I would strongly recommend that company A, the sponsor of the research, ignore the regression result and aggressively pursue improvements on this dimension of product quality.

As a consequence of this dependence on the world-as-it-is, preference regression has a tendency to rediscover sameness in company performance and to recommend the known, safe solutions to performance improvement.

Because predictions (and hence, simulations) can be made, this is the first model in this paper which can be used to carry out a formal cost-benefit analysis. If the cost of changing performance on each attribute is known, then the costs of producing the same increase in overall satisfaction by improving performance on different attributes can be calculated. These differential costs should be taken into account when deciding which attributes to work on first.

In all forms of preference regression, the regression coefficients are taken to be the same for every observation. That is, there can be no individual differences in the impact of an attribute on overall satisfaction. Given what we know about the prevalence of segmentation, particularly segmentation on preferences and values. I expect some differences among individuals in their responses to attribute changes, and so, having a coefficient be constant across individuals is not desirable.

Preference regression is also typically linear, that is, the impact of an attribute is the same whether performance is poor or good on that attribute. Among other things, this implies that a one unit increase in attribute performance results in the same magnitude increase in overall satisfaction as the magnitude of the decrease that results from a one unit decrease in the attribute performance. Non-linear models involving interactions or "diminishing returns" effects are certainly possible, but are rarely used.

We now expand our table comparing methods against objectives to include preference regression. In my opinion, preference regression must be treated with a great deal of caution.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts	Benefit/ Loss Charts	Pref. Reg.
1. Company Performance Assessment	+	+	•
2. Attribute Prioritization	•	•	-
3. O'all Sat. Resulting from Perceived Perf. Changes	-	-	•
4. Competitor Risks & Opportunities	+	•	+
5. Co. Actions Link to Perceptions	-	-	-
6. Behaviors Resulting from Customer Satisfaction	-	-	-

+ = Achieves - = Fails • = Partial

Preference regression suffers from the effects of multicollinearity, attribute weights which do not allow individual differences in the impact of attributes, and limitation to the world-as-it-is rather than as it could be. We turn now to a model which is not as susceptible to these liabilities.

EXPECTANCY-VALUE MODELS

Actually, expectancy-value models are not all that common in customer satisfaction modeling, although they are sometimes used. The tempting calculation in which attribute importances are multiplied by attribute satisfaction ratings is the first step in expectancy-value modeling.

In the basic model as it has been adapted in the market research literature, these arithmetic products of importance and satisfaction are simply added up and the composite is taken as a measure of overall satisfaction. However, since this calculation isn't really a model, but rather a definition, I find it necessary to add to the equation a simple linear rescaling of the composite, so it at least has the right scale and is centered in the right place compared to overall satisfaction. Having done that, the expectancy-value model looks a lot like preference regression:

$$y_i = b_0 + b_1 \cdot (I_{i1} \cdot x_{i1} + I_{i2} \cdot x_{i2} + \ldots + I_{iK} \cdot x_{iK}) + e_i$$

where all terms are as defined for preference regression and I_{ik} is observation i's importance rating for attribute k. The I's, in effect, replace the regression coefficients in the preference regression model, thereby producing individual differences in the impact of an attribute on overall satisfaction. Again, there are many variants on this basic model, but this one is probably most typical.

Note that the mean of a product of two variables does not equal the product of the separate means of the variables. Because of this, the mean of y (overall satisfaction) predicted by the above model is not the same as the sum of the products of the mean I's (importances) and mean x's (attribute satisfaction ratings), even when those means are rescaled using the b's in the model. Therefore, the expectancy-value model cannot quite be thought of as a composite of the information in satisfaction/ importance charts, although they are obviously related in some ways.

Applying the above model to our example problem, where K is 34 and the sample size is 133, results in an R^2 of .39, highly significant statistically, albeit substantially lower than the .72 for preference regression. The value of b_1 is positive, which means that an increase in satisfaction on any attribute will predict an increase in overall satisfaction, as we would hope. (Unfortunately, in the interest of client confidentiality of the model, I cannot give the values of the coefficients here.)

Because of low degrees of freedom in this model (1), the jackknife cross-validity coefficient, P^2, does not decline much from R^2: P^2 is .37, indicating about the same cross-validity as for preference regression where P^2 is .35. Consistent with this evidence of improved stability is the condition number of 7 (compared to that of 180 for preference regression).

As an aside, we note that missing data may be more than just a nuisance problem in customer satisfaction research. If a large number of respondents do not rate a company on one attribute, it seems likely that the attribute is irrelevant to those people. Consequently, the attribute should play no role in those customers' models. Of course, you could estimate a separate regression model for those people, but if there are a large number of different patterns of missing data, this can be impractical. This problem applies as much to preference regression as it does to expectancy-value

modeling. A number of other solutions to the problem exist (Little & Rubin, 1987), and, rather than get bogged down here with what is a very knotty side issue, I will assume for this discussion that all observations have all importance and attribute satisfaction data present. However, this issue should not be taken lightly in the actual implementation of preference regression or expectancy-value modeling.

The basic expectancy-value model is really not quite appropriate. We must bear in mind the considerations raised in the section on satisfaction/importance charts concerning the dangers of multiplying importance and satisfaction ratings. We noted there that the outcome depends very much on the somewhat arbitrary choice of scale values used in the ratings. Indeed, because the rating scales we use provide interval scale data, we can consider any linear rescaling of the ratings without loss or distortion of information in the scale:

$$s = a_0 + a_1 \cdot r$$

where r is the original rating, s is the rescaled rating, and a_0 and a_1 are rescaling parameters. Applying this type of rescaling to both the importance and the attribute satisfaction ratings in the basic expectancy-value model above yields the following model, called the rescaled expectancy-value model:

$$y_i = b_0 + b_1 \cdot [(a_0 + a_1 \cdot I_{i1})(c_0 + c_1 \cdot x_{i1}) + \ldots + (a_0 + a_1 \cdot I_{iK})(c_0 + c_1 \cdot x_{iK})] + e_i$$

This model must be reduced to four parameters to eliminate indeterminacy so that parameters can be estimated uniquely. After doing so, the model can be estimated by multiple regression.

Individual differences in value are still retained in this model in the form of the individual importance ratings.

By including importance ratings in this model, where importance is assessed without reference to the specific performance of current companies, we have achieved some degree of independence from the world-as-it-is. Of course, because attribute satisfaction is measured only on existing companies, there is probably, in the relatively few parameters of the model, still some element of dependency on the world-as-it-is. Certainly, as we shall see in the next section, the expectancy-value model does not eliminate dependency on the world-as-it-is as much as does conjoint analysis, which systematically varies attribute performance without constraint by the actual performances of existing companies or by the interrelationships among attributes.

Applying the rescaled expectancy-value model to our example problem, we obtain an R^2 of .50, higher than the basic model and significant statistically, but still lower than the .72 for preference regression. The implied signs of the b_1 and c_1 coefficients are both positive, again as we would hope. For the rescaled model, P^2 is .48, indicating better cross-validity than for either the basic model or for preference regression. The stability of parameter estimates is still acceptable, though not as good as for the basic model, as evidenced by the condition number of 28.

When examining the residuals of prediction here, we find, as we often do in these studies, that the model under-predicts at the high end of the scale (for example, 10's are predicted to be lower) and over-predicts at the low end (for example, 1's are predicted to be higher). This effect is found for

preference regression as well. It is due to the truncation caused by "end-effects" on direct ratings: the limitation to 1 at the lower end of the scale and to 10 at the upper end. This truncation effect is stronger in a composite than in a single variable, producing the effect mentioned. It is possible to modify either the preference regression or the expectancy-value model to be non-linear to account for this effect and so, produce more accurate predictions. However, this is difficult to do properly so that the resulting model does indeed improve both its predictions and its cross-validity, and even when that happens, the improvement in accuracy is often very modest. Consequently, non-linear models are not always necessary, although they should always be tested for, and consideration should be given to their use when appropriate.

It is commonly the case that a customer may have different awareness of and varying certainty about company performance on different attributes. One way to capture this in the expectancy-value model is to add a factor for each attribute to account for these effects:

$$y_i = b_0 + b_1 \cdot [w_1(a_0 + a_1 \cdot I_{i1})(c_0 + c_1 \cdot x_{i1}) + \ldots + w_K(a_0 + a_1 \cdot I_{iK})(c_0 + c_1 \cdot x_{iK})] + e_i$$

This model is referred to here as the extended expectancy-value model. Once again, the parameters of this model are indeterminate without imposing constraints on them.

This extended model now looks similar to the preference regression model, except that the coefficients (w) are weighting the product of the rescaled I's and x's, rather than the x's alone. This similarity to preference regression implies that the model should produce substantial improvement in R^2.

However, like preference regression, the price is the risk of negative coefficients and an expected loss of stability in parameter estimates and in cross-validity. The instability problem can be solved by the introduction of ridge regression or some other form of shrinkage estimation. The negative coefficients can be eliminated by imposing the inequality constraint that all w's be non-negative. Both of these solutions are difficult to accomplish computationally, and may or may not be worth the trouble.

In addition, also like preference regression, improved R^2 is achieved with increased reliance on the world-as-it-is in determining the model. While we have retained some independence from this reliance by including the importance ratings in the extended expectancy-value model, we have definitely worsened matters as compared to the rescaled model. The extent to which the extended model is more like preference regression or more like the rescaled model in this regard, depends on the extent of the flattening effect on the coefficients caused by shrinkage estimation and the non-negativity constraint: the more nearly equal are the w's, the more the extended model will be like the rescaled model.

In our example problem, an extended model applied to the data yielded results which can be compared to those of the rescaled and basic expectancy-value models and to the preference regression model in the table below.

COMPARISON OF 4 MODELS

	R^2	P^2	Condition Number
Extended Expectancy-Value	.55	.52	20
Rescaled Expectancy-Value	.50	.48	28
Basic Expectancy-Value	.39	.37	7
Preference Regression	.72	.35	180

All coefficients in the extended model are positive, as we have required.

It is questionable from these results whether the extended model is a worthwhile improvement over the rescaled expectancy-value model, especially given the extended model's partial reliance on the world-as-it-is. However, based largely on cross-validity considerations, both the extended and rescaled models are preferable to the basic model or to preference regression.

If the mean rating for satisfaction on an attribute (say attribute k) is increased by one unit, without any changes in other attributes, then it can be shown that the change in mean overall satisfaction predicted by the rescaled expectancy-value model is:

$$\Delta_y = b_1 \cdot c_1 \cdot (a_0 + a_1 \cdot I_k)$$

and the change predicted by the extended model is:

$$\Delta_y = b_1 \cdot w_k \cdot c_1 \cdot (a_0 + a_1 \cdot I_k)$$

where I_k in both equations is the mean importance rating. In other words, the impact of an attribute on overall satisfaction is linear with the mean importance rating for that attribute. As a consequence, the rescaled expectancy-value model prioritizes attributes in exactly the same way as do the satisfaction/importance charts. The extended model will prioritize attributes differently insofar as the w's are not all equal to one another.

Because of the preceding equations, the expectancy-value model predicts the consequences of company actions, so long as those actions can be translated into an implied change in mean attribute satisfaction ratings. If attribute satisfaction is measured by direct ratings, this translation could involve a fair amount of guesswork, unless additional research is carried out to map company actions into attribute satisfaction.

Bear in mind also that any company action which would cause a change in satisfaction with one attribute, may well have an effect on satisfaction with other attributes as well. It may not be realistic to assume that we can ever change satisfaction ratings on one and only one attribute, which is what is implied in the above equation for Δ_y.

If the expectancy-value model is based on data across all companies in the survey, you can investigate and predict opportunities and vulnerabilities of both your company and of your competitors. Although it is based on performance data, an expectancy-value model, by itself, does not provide an explicit company performance assessment — you will still need something like satisfaction/importance charts for this.

Nonetheless, with the predictive model and with some way to map changes in concrete company actions to changes in satisfaction-with-attribute(s), we have the possibility of conducting cost-benefit analyses to help judge whether a particular service improvement is worth the cost, or conversely, whether the savings from reducing service are worth the bad will.

With expectancy-value modeling we have obtained much more stable predictions of overall satisfaction and have improved the stability of the parameter estimates. Expectancy-value models, if properly constructed, should provide a valuable tool in customer satisfaction research, my opinion on this being expressed in the following table.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts	Benefit/ Loss Charts	Pref. Reg.	Exp.-Value
1. Company Performance Assessment	+	+	•	•
2. Attribute Prioritization	•	•	-	+
3. O'all Sat. Resulting from Perceived Perf. Changes	-	-	•	+
4. Competitor Risks & Opportunities	+	•	+	+
5. Co. Actions Link to Perceptions	-	-	-	-
6. Behaviors Resulting from Customer Satisfaction	-	-	-	-

+ = Achieves - = Fails • = Partial

While expectancy-value modeling reduces dependency on the world-as-it-is for its parameters, compared to preference regression, it still retains some element of the world-as-it-is in its estimates, particularly the extended expectancy-value model. We turn now to consideration of a popular model which is influenced very little by the current attribute performance of existing companies.

CONJOINT ANALYSIS

There are many forms of conjoint analysis and related methods, including:

- Full-profile conjoint
- Partial-profile conjoint (as in ACA)
- Pairwise tradeoff analysis
- Discrete choice analysis
- Hybrid conjoint analysis

These methods are covered in depth elsewhere, including in many present and past papers at this conference. I do not intend to provide that coverage here, but rather will provide only a cursory description.

Conjoint requires attributes to be defined in terms of very specific concrete attribute levels, for example, levels of "Once a week," "Every two weeks," "Once a month," and "Less than once a month" for the attribute "Frequency of rep contact." Overall satisfaction ratings are obtained for systematically varied attribute bundles. Satisfaction with performance levels on the attributes (usually called utilities or partworths) is derived from those ratings. Conjoint produces a model for estimating overall satisfaction from specified levels of performance on attributes. This model may have different parameters (that is, partworths) for every respondent. The conjoint model may be used with a discrete choice model to produce share estimates for companies simulated by the attribute bundles which characterize their products.

Because we control the profiles of attribute levels assigned to the attribute bundles rated by the customer, there should be little or no problems with multicollinearity in conjoint, in contrast to preference regression.

Unlike the previous methods, conjoint analysis does not rely at all on ratings of existing companies in computing its parameters, the partworths. Because of this, we free ourselves as much as it is possible to do from the limitations of the world-as-it-is. More than any other method, conjoint encourages the discovery of the new and revolutionary in products and services because it allows us to explore levels and combinations not anywhere represented in the current market.

Nonetheless, we still need to relate our model to company performance if it is to be useful. A convenient and useful way to accomplish that is to obtain ratings of the companies on the attribute levels used in the conjoint. These ratings of the companies are then used directly in the conjoint model to predict overall satisfaction incorporating customer perceptions. Since customers usually

vary in their perceptions of a company's performance on an attribute, there is a distribution of ratings across the attribute levels. This presents something of a problem, because traditionally, conjoint uses for every respondent, a fixed level for specifying a simulated company's attribute performance. Either of two approaches is typically used in the present situation to insert customer ratings of company performance on an attribute:

- Use the modal category of the rating as a fixed level of performance for that company for every customer.

- Ignore tradition and insert each customer's unique perceptions of attribute performance for a company.

The first approach has the advantage of ease and fitting in to traditional conjoint simulation software more easily. The second approach has the advantage of greater realism, in that it more fully reflects the variability in customer perceptions, which variability in turn contributes to the distribution of predicted overall satisfaction. This last point is particularly germane when using conjoint in a discrete choice model for estimating shares, since the variance of predicted overall satisfaction, and not just its mean, affects estimated shares.

To simulate changes in a company requires the user of a conjoint model to specify changes in attribute performance for that company. For the first approach to inserting customer perceptions, the user only need specify a change in the modal response. However, in the second approach the user must specify changes in the distribution of ratings across attribute levels. One common approach here is to move everyone up (or down) a number of levels equal to the number of levels that the user would like to simulate, truncating at the end of the scale. For example, if the distribution on an attribute is 20%, 30%, and 50% for levels 1, 2, and 3, moving everyone down (toward level 1) by one unit would result in a distribution of 50%, 50%, and 0% for the three levels. Other, more complex procedures exist, whereby the user can fully specify every level of the entire distribution.

As to which of the two approaches to use for inserting customer perceptions, the tradeoff is between convenience and realism, and is often not an easy decision to make. For estimating shares, the difference is probably great enough that realism should prevail; for estimating mean overall satisfaction, it may not matter as much.

Conjoint analysis typically calculates a variable called an "Importance Score" for each attribute. This Importance Score usually reflects the amount of difference an attribute could possibly make in overall satisfaction, if the attribute changed from the worst to the best possible level. However, that amount of change is usually not the amount possible for any given company: almost every company has some customers who rate its current performance above the worst level or below the best level of any given attribute. Since the Importance Scores can't really represent the impact potential for a given company, I do not recommend using these scores in prioritizing attributes for that company.

It would make more sense to vary a particular company's performance on the attribute from its current position to one level higher (or lower) or to calculate how much of an increase in overall satisfaction would occur if the company's performance were improved to the most attractive level of the attribute. (Of course, the issues discussed above about methods for inserting customer perceptions and then for simulating changes pertain here.) This approach to assessing the impact potential of an attribute is particularly relevant when the conjoint model for estimating satisfaction is

used in a discrete choice model to estimate shares or probabilities of choice for the companies. In this context, the partworths are not linearly related to the outcome variable and so the Importance Scores have no particular meaning at all.

In contrast to preference regression or expectancy-value models, conjoint does not operate effectively with "soft" attributes, such as "Rep friendliness." Since most of the primary level attributes, in the hierarchy of attributes alluded to earlier, are actually quite "soft," conjoint does not work well on primary attributes. It is better suited to secondary attributes, which can usually be stated in fairly concrete terms.

Unfortunately, there are usually many more secondary attributes than conjoint can handle comfortably. Full-profile conjoint should not be made to handle more than about 8 - 10 attributes. In ACA, I find respondents get confused and rebellious somewhere around 15 attributes. Other approaches (involving split sampling or hybrid conjoint) may handle more attributes, but they would all likely choke on the 34 attributes used in our example, and that isn't even a particularly large number of attributes for a customer satisfaction study.

The use of conjoint in predicting shares for scenarios in which companies change their performance is unique among methods described so far, in that this is the one model which attempts to predict behavior and not just overall satisfaction. As will be seen in the next section, predicting behaviors is an important extension of customer satisfaction modeling.

It must be pointed out, however, that the shares that conjoint predicts often bear little resemblance to actual market shares, based as they are only on preferences, and not including such factors as awareness and availability. Furthermore, there are other customer behaviors besides purchase probabilities (measured by shares) which are important, like complaining to a regulatory agency or the Better Business Bureau, or switching to another brand. Conjoint should get credit for attempting to predict behaviors, but it still isn't the complete picture.

Another important aspect of the problem is the influence of company actions on customers' perceptions of company performance on attributes. Conjoint's use of specific concrete levels of attribute performance makes associating actions with perceptions much easier than is the case with attribute satisfaction ratings on a 1 - 10 scale. However, there is another complicating factor. To illustrate, consider that improving the mean response time to requests for technical information from 1 day to 2 hours may influence customers' perceptions of performance on other attributes besides "response time to inquiries": for example, perceptions of "friendliness of reps," "response time on service requests," and "speed of order filling," may also be affected. Furthermore, we don't even know for sure that, just because the company actually does decrease its response time from 1 day to 2 hours, that customers will necessarily perceive the change.

Nonetheless, conjoint analysis is the first place where the last two criteria (5 and 6 in the table below) are addressed in any way at all. In view of its other advantages (no multicollinearity, independence of world-as-it-is), and its few disadvantages (possible difficulty specifying attribute changes, difficulty with "soft" attributes and with large numbers of attributes), I find conjoint compares favorably with the other methods.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts	Benefit/ Loss Charts	Pref. Reg.	Exp.-Value	Con-joint
1. Company Performance Assessment	+	+	•	•	•
2. Attribute Prioritization	•	•	-	+	+
3. O'all Sat. Resulting from Perceived Perf. Changes	-	-	•	+	+
4. Competitor Risks & Opportunities	+	•	+	+	+
5. Co. Actions Link to Perceptions	-	-	-	-	•
6. Behaviors Resulting from Customer Satisfaction	-	-	-	-	•

+ = Achieves - = Fails • = Partial

MARKET EXPERIMENTS AND BEHAVIORAL REGRESSION

To this point, as is appropriate given the customer orientation of customer satisfaction research, we have focused largely on the internal judgments and decision processes of the customer. It is helpful to think of the customer as a system with inputs and outputs. The models described thus far have been models of the internals of the customer system. However, as indicated previously, to complete the picture, we need to understand more about the effects of inputs on the system, and the relationships of the system to its outputs.

Strictly speaking, the techniques used for examining the inputs and outputs are outside the scope of customer satisfaction modeling, which is an internal "system" model. I cover them here only briefly, but I would stress that obtaining these results is an extremely important piece of the whole puzzle.

The inputs to the system are the products and services provided by a company. Actions by the company regarding those products and services will therefore affect the system and, to successfully plan for customer reactions, the company must understand the effects caused by its actions. In the models we have discussed, the point at which company actions connect with the system are through

customer perceptions of performance on attributes, as measured by ratings of level or degree of performance, or as measured by satisfaction with that performance.

Full understanding of this connection requires that the response of customer perceptions to variation in company actions must be measured. The analytic tools are either analysis of variance or regression analysis, with the dependent variable(s) being perceptions and the explanatory variable(s) being company actions. The data for these analyses are: distributions of perceptions on attributes associated with objectively measured company actions.

The company actions measured may be whatever actions naturally occur in the market place. In that case, confounding of the effects of different actions is quite likely — similar to what we found in preference regression when using perceptions of current performance on attributes to predict overall satisfaction.

More appropriately, the company actions may be controlled systematically, in effect by performing market experiments, in which the actions are experimentally manipulated for the purposes of observing their effects. Market experiments are difficult to carry out successfully, because it is virtually impossible to hold all relevant variables constant while manipulating only the variables (company actions) of interest. Competitors have a nasty habit of interrupting your experiments with their own company actions, thereby confounding their effects with yours. Nonetheless, market experiments are the most effective means of attempting to understand the distribution of perception resulting from your actions.

At the other end of the customer system, the output end, is the need to relate actual customer behaviors to overall satisfaction or other measures derived from our research, such as intentions to behave or conjoint share estimates. Modeling actual behaviors is critical if customer satisfaction models are to be effectively used, since predicting overall satisfaction is not, by itself, generally sufficient. Customer satisfaction models (the good ones) show you, in effect, how to make people happy with your products and services. Making people happy with your products and services is, after all, a means, not an end: the "end" being profitability. For commercial enterprises, if changes do not result in increased profitability in the long term, they have not been worth it. Cost-benefit analysis based on customer satisfaction as the "benefit" is only a partial solution: ideally, the "benefit" ought to be revenue or something closely related.

Regression analysis lends itself well to the problem of predicting behaviors and so, I refer to the analysis of outputs as behavioral regression. In its simplest form, this is regressing an actual behavior on overall satisfaction or other derived measure (such as conjoint share estimates). This may be done at the individual customer level (for example, by obtaining actual purchases of a product as well as the data necessary for a conjoint analysis) or at the aggregate market level (for example, by obtaining actual market shares from a secondary source and comparing them to outputs of the conjoint model).

Behavior, however, is not that simple: in most markets, we know that there are other determinants of behavior besides the preferences and perceptions measured in customer satisfaction studies. For example, awareness created by advertising, availability created by distribution systems, and situational constraints imposed by cash flow, the already installed equipment base, or the regulatory environment are but a few of the other influences which we don't measure with customer satisfaction models.

Accommodating these effects requires recourse to econometric-type regression models in which customer behavior is the dependent variable and all effects, including, but not limited to, customer satisfaction model predictions of overall satisfaction, intention or behavior, are treated as explanatory variables. The regression model used here may well be a non-linear one, such as the multinomial logit model described for the individual level in Train, *et al.* (1986).

It is also possible to use data at aggregate levels in behavioral regression, relating aggregate behaviors (like actual shares or actual complaint counts) to aggregate measures on customers (like mean overall satisfaction) and to market measures on customers and products/services. To use aggregate models requires some level of disaggregation (for example, measurement of changes over time and/or over several subpopulations of customers) in order to have enough variation in the data to detect relationships. In my experience, aggregate level analyses seldom have enough data to be very rich sources of information about the relationship of behavior to customer satisfaction. I find it is generally preferable to carry out the individual level modeling.

To complete the comparison of approaches, I now include market experiments and behavioral regression in the comparison table.

COMPARISON AGAINST OBJECTIVES

Objective	Sat./ Imp. Charts	Benefit/ Loss Charts	Pref. Reg.	Exp.- Value	Con- joint	Mkt. Expmts	Behav. Reg.
1. Company Performance Assessment	+	+	•	•	•	+	+
2. Attribute Prioritization	•	•	-	+	+	-	-
3. O'all Sat. Resulting from Perceived Perf. Changes	-	-	•	+	+	+	-
4. Competitor Risks & Opportunities	+	•	+	+	+	-	•
5. Co. Actions Link to Perceptions	-	-	-	-	•	+	-
6. Behaviors Resulting from Customer Satisfaction	-	-	-	-	•	•	+

+ = Achieves - = Fails • = Partial

THE CUSTOMER SATISFACTION RESEARCH CASCADE

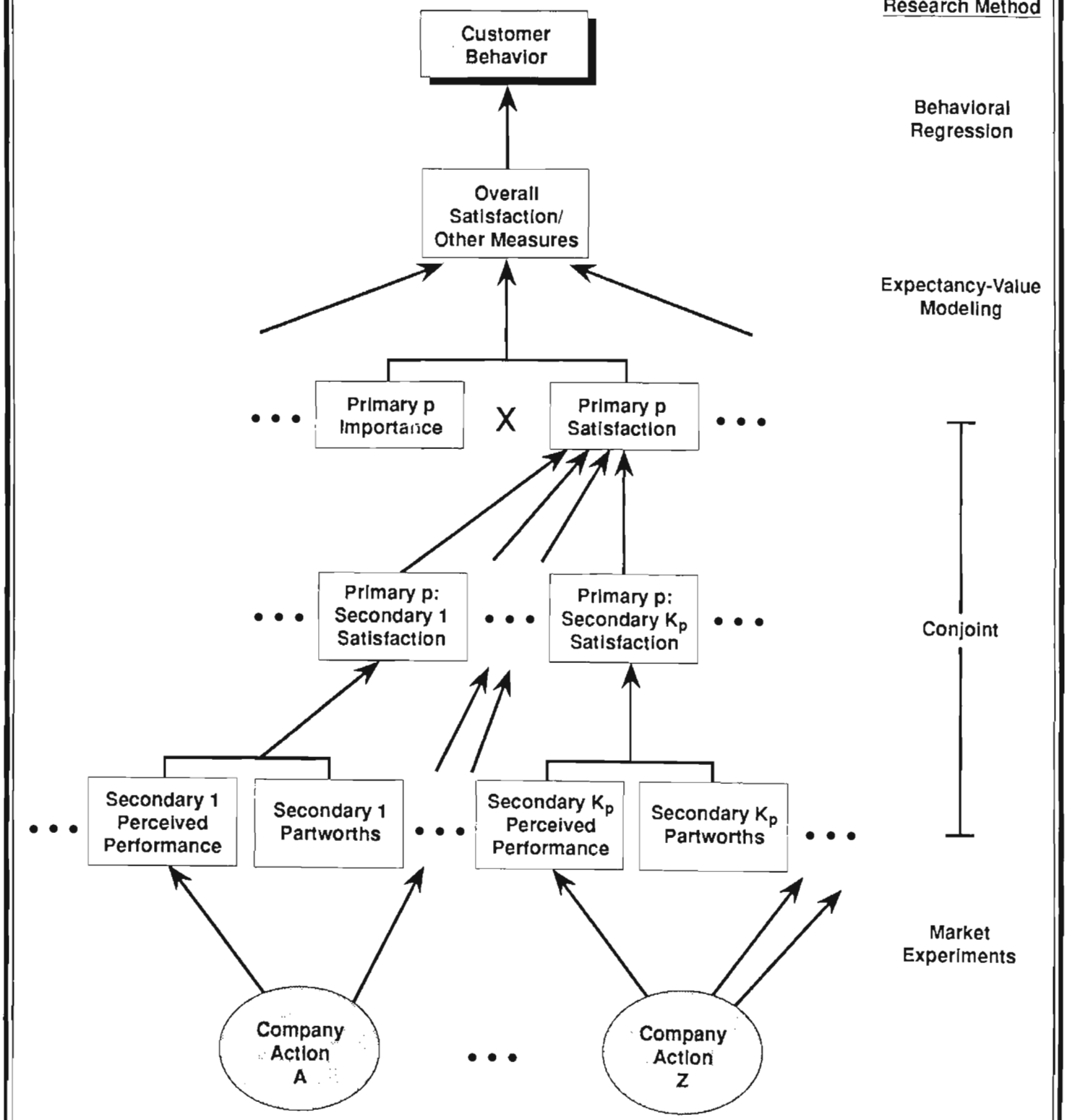
Of course, it would be impossible to construct a single study which simultaneously obtained all the data for a complete model of the system, its input links and its output implications. In fact, such a complete model may well be utterly impractical and impossible no matter how many studies are undertaken.

It is tempting to simplify the problem by ignoring the intermediary "system" in the complete model: treat the outputs (behaviors) as the dependent variable(s) and the inputs (company actions) as the explanatory variables in a regression model. Even this may be impractical, however, because a grand market experiment in which many company actions are explored in detail quickly gets to be very complex and costly. Often, however, a more modest version of this strategy in which only a

few variables are manipulated and customer behavior is observed may be a perfectly reasonable research approach. However, when the objective is to better understand the customer, we know this is a fairly sterile model. To really provide the kind of insight about customers required by management disciplines like TQM or QFD necessitates the modeling of that internal "system," customer satisfaction.

We see from the comparison table that several methods taken together can be made to cover most of our objectives. Practically, the problem must be broken down into pieces, starting with the system piece, the inputs and the outputs. Using behavioral regression, customer satisfaction models, and market experiments as separate research endeavors is viable if a unified model is used and some means of feeding one model into the appropriate next model is provided. Figure 4 shows an example of such a strategy, in which some of the approaches discussed form a kind of cascade of research, with one level establishing elements of the model and linking on to the next level.

Figure 4
RESEARCH CASCADE



With a research cascade, it is not necessary to build the whole model in one study: separate pieces can be built and then assembled, so long as they are designed to be interlinked. Each piece, by itself, is significant research and will be useful and helpful as a stand-alone entity. The complete model may never be finished, but you will find that every part of the attempt is invaluable to your ability to better serve your customers.

Building such a model is a difficult undertaking, requiring a long term commitment to the process, but then, that is characteristic of providing quality products and services in general. While I have emphasized that customer satisfaction is a means, not an end, and that the ability to base demand and/or profit improvement on a model is the ultimate test of its usefulness, I would also echo the Baldridge award in asserting that a focus on customer satisfaction is an important component of long term success. And, to paraphrase the advocates of Total Quality Management: customer satisfaction should be a journey, not a destination.

REFERENCES

Bass, F. M. and W. L. Wilkie (1973). "A Comparative Analysis of Attitudinal Predictions of Brand Preferences." *Journal of Marketing Research*, 10 (1973), 262-269.

Belsley, D. A. (1991). *Conditioning Diagnostics*. New York: Wiley.

Bock, R. D. and L. V. Jones (1968). *The Measurement and Prediction of Judgment and Choice*. San Francisco: Holden-Day.

Buzzell, R. D. and B. T. Gale (1987). *The PIMS Principles*. New York: Free Press.

Cohen, J. and P. Cohen (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: Wiley.

Coombs, C. H. (1964). *A Theory of Data*. New York: Wiley.

Cooper, L. G. and C. T. Finkbeiner (1983). "A Composite MCI Model for Integrating Attribute and Importance Information." in T. C. Kinnear, ed., *Advances in Consumer Research, Vol. X*I. Provo, UT: Association for Consumer Research, 109-113.

Debreu, G. (1960). "Topological Methods in Cardinal Utility Theory." in K. J. Arrow, S. Karlin, & P. Suppes, eds., *Mathematical Methods in the Social Sciences*. Stanford, CA: Stanford University Press, 16-26.

Farebrother, R. W. (1978). "A Class of Shrinkage Estimators." *Journal of the Royal Statistical Society, B*, 40, 47-49.

Fishbein, M. (1967). "A Behavior Theory Approach to the Relations between Beliefs about an Object and the Attitude toward the Object." in M. Fishbein, ed., *Readings in Attitude Theory and Measurement*. New York: Wiley, 389-399.

Hoerl, A. E. and R. W. Kennard (1970). "Ridge Regression: Biased Estimation for Non-Orthogonal Problems." *Technometrics*, 12, 55-67.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley.

Luce, R. D. and J. W. Tukey (1964). "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology*, 1 (1964), 1-27.

Massy, W. F. (1965). "Principal Components Regression in Exploratory Statistical Research." *Journal of the American Statistical Association*, 40, 234-256.

National Institute of Standards and Technology (1992). *Malcolm Baldridge Quality Award: 1992 Award Criteria*. Gaithersburg, MD: U.S. Dept. of Commerce.

Rosenberg, M. J. (1956). "Cognitive Structure and Attitudinal Affect." *Journal of Abnormal and Social Psychology*, 53, 367-372.

Thurstone, L. L. (1927). "A Law of Comparative Judgment." *Psychological Review*, 34, 273-286.

Thurstone, L. L. (1945). "The Prediction of Choice." *Psychometrika*, 10, 237-253.

Train, K., D. McFadden, and A. Goett (1986). "The Incorporation of Attitudes in Econometric Models of Consumer Choice." paper presented at the American Marketing Association Conference on Attitudes and Behavior, March.

Comment on Finkbeiner

William G. McLauchlan
McLauchlan & Associates, Inc.

As usual, Finkbeiner's paper represents an outstanding contribution to the proceedings of this conference. I'm particularly pleased to hear his remarks with respect to preference regression.

Preference regression is fraught with danger. Problems of multicollinearity, negative coefficients, and low R^2 values aside, there is, I believe, a much more fundamental issue with what Finkbeiner neatly describes as this "world-as-it-is" perspective. I am concerned when preference regression analyses are conducted because we choose not to believe our respondents when they tell us what is important to them. Interestingly, however, we are happy to use their satisfaction and agreement measures as a basis for modeling their preferences.

The balance of my comments relate to an alternative to the Satisfaction/Importance charts that Finkbeiner describes. Recognizing that what I will present is no more of a "model" than is the familiar quadrant chart, I believe that the analysis I describe below can be a useful way of integrating Importance and Satisfaction measures into what Finkbeiner calls an "Index of Opportunity."

GAP scores, as I will refer to them, illustrate the relationship between attribute importance and product satisfaction in a manner that is slightly different from the more traditional Satisfaction/ Importance Quadrant Analyses usually conducted using these kinds of data. In both analyses, the objective is to identify product characteristics that are important to the market where satisfaction is sub-optimal.

GAP Analyses are conducted from the perspective that satisfaction should be mathematically as high as average satisfaction but that satisfaction needs to be weighted by the satisfaction on all attributes and by the importance of the attribute, relative to the importance of all attributes. In other words, it is not sufficient to be at least average in satisfaction. Instead, satisfaction should be at least average when mathematically compared to importance and overall satisfaction.

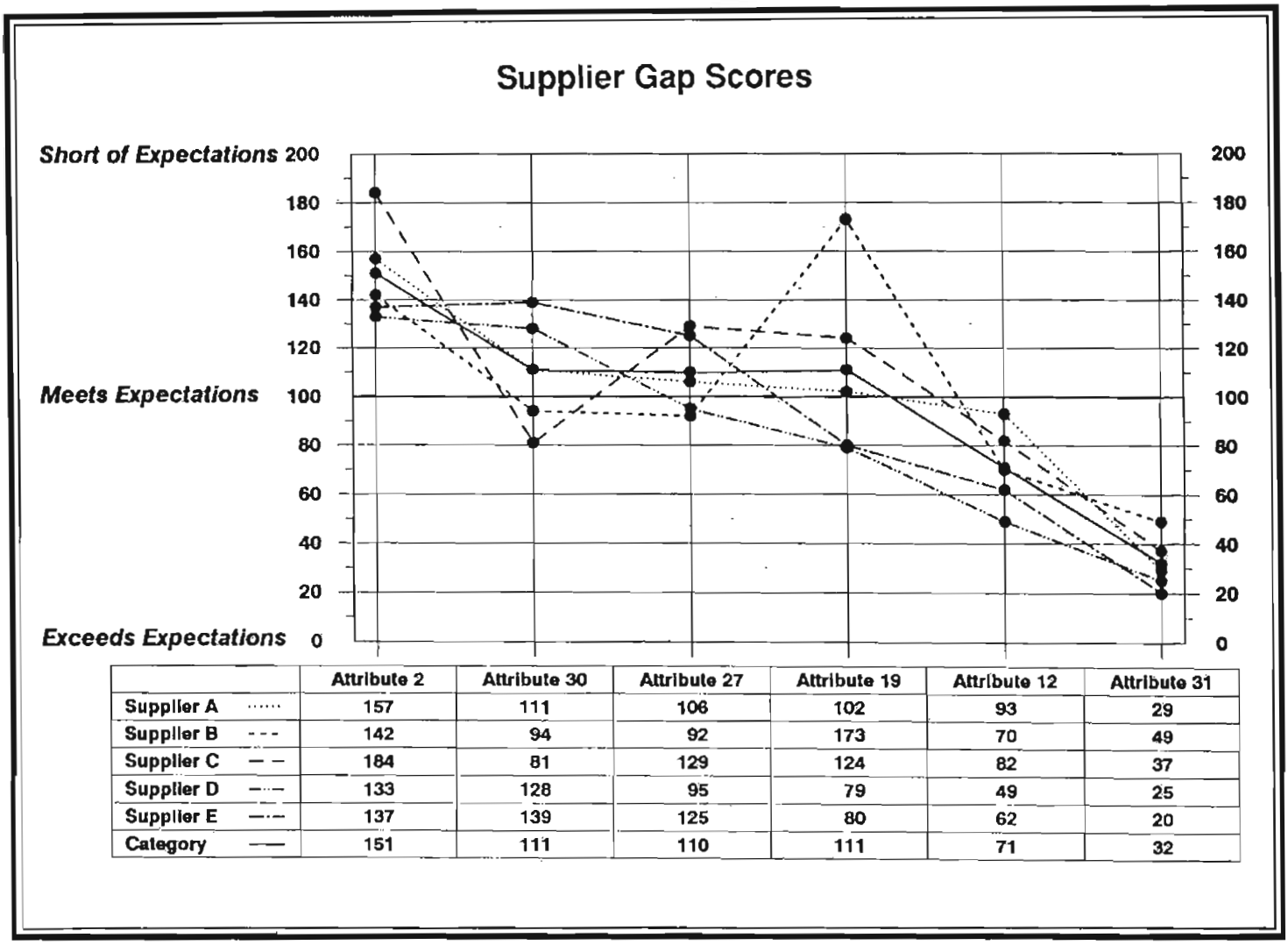
The GAP scores are calculated based on the following formula:

$$\text{GAP} = \frac{(\text{Attribute Importance} * (10 - \text{Attribute Satisfaction})) * 100}{(\text{Grand Mean Importance} * (10 - \text{Grand Mean Satisfaction}))}$$

where the "10" in the formula indicates "perfect" satisfaction on the 1 to 10 satisfaction scale. (The appropriate value should be substituted for different scales.) Conceptually, the formula is computing the distance from "perfection" in satisfaction on a given attribute relative to the distance that the average satisfaction is away from "perfection." This ratio is then weighted by the relative importance of the attribute.

Finkbeiner was kind enough to provide the coordinates for the plotted points in his Figure 2: Satisfaction/Importance Chart. Using these data, GAP scores were computed and have been plotted in the following chart.

Figure 1



	Attribute 2	Attribute 30	Attribute 27	Attribute 19	Attribute 12	Attribute 31
Supplier A ·····	157	111	106	102	93	29
Supplier B - - -	142	94	92	173	70	49
Supplier C — —	184	81	129	124	82	37
Supplier D —·—	133	128	95	79	49	25
Supplier E —·—	137	139	125	80	62	20
Category ——	151	111	110	111	71	32

As can be seen, the comments that Finkbeiner makes regarding manufacturer performance on the various attributes still apply when the GAP scores are examined. What may make the GAP approach more viable is the ability to plot many more attributes on a single chart and to discern more easily competitive strengths and weaknesses.

A COMPARISON OF RESULTS OBTAINED FROM ALTERNATIVE PERCEPTUAL MAPPING TECHNIQUES

Thomas L. Pilon
TRAC, Inc./University of North Texas

Perceptual mapping addresses the general problem of positioning objects (brands) and attributes in a perceptual space. There are numerous perceptual mapping techniques. There is some confusion among marketers and market researchers as to:

- the theoretical differences between these techniques,
- the appropriate applications for each of these techniques, and
- the proper interpretation of results from each of these techniques.

The purposes of this paper are to provide a layperson's introduction and to demonstrate the use of a few of the most popular techniques. As background, the first section of this paper will provide an overview of the uses of perceptual mapping. The next section will classify many of the available techniques according to their general approach. A representative technique from each general approach will be described and employed on a common database (transformed and re-scaled as necessary to be appropriate to the particular technique). Finally, differences in interpretation will be discussed. This paper will not address the issue of "ideal points."

USES OF PERCEPTUAL MAPPING

Perceptual maps are used in a number of different ways. They are used to:

- identify the perceived relative strengths and weaknesses with respect to an object's attributes,
- define the competitive set of objects that is perceived to be similar or dissimilar,
- determine the similarities and differences in perceptions across various market segments,
- identify repositioning opportunities and expose vulnerabilities of the current position, and
- uncover new product opportunities based on gaps in the current market structure.

In spite of obvious value of these uses, perceptual maps are still very widely misunderstood.

A FOUR-STEP PROCEDURE FOR PERCEPTUAL MAPPING

1. Identify All Objects to Be Evaluated

The most important issue in perceptual mapping is to generate a complete list of objects to be evaluated. Objects can be products, services, corporations, institutions, retail stores, cities, celebrity endorsers, political candidates/parties, events, and so on. It is critical that all "relevant" objects be included since perceptual mapping is a technique of relative positioning. Relevancy should be determined by the objectives of the research study. The perceptual maps resulting from any of the methods can be greatly influenced by either the omission of relevant objects or by the inclusion of inappropriate ones (Maholtra, 1987).

2. Select Decompositional or Compositional Approach

The choice of approach should be made early in the design of a research project since each approach has very different data requirements. The decompositional approach decomposes overall respondent impressions as to the similarity (or dissimilarity) of objects to form a multidimensional map. The compositional approach composes ratings on objects on attributes into a multidimensional map with considerably fewer dimensions than attributes. Each approach has advantages and disadvantages which will be discussed in the paragraphs below.

The Decompositional or Attribute-Free Approaches have two distinct advantages. First, they require only that the researcher collect information on respondents' overall perceptions of objects; a researcher does not need to generate a list of relevant attributes nor collect ratings on them. Secondly, since each respondent gives a full assessment of similarities across all objects, perceptual maps can be generated for individual respondents or aggregated to form a composite map.

These advantages result in some disadvantages, as well. First, it is difficult for the researcher to identify the basic dimensions by which respondents evaluate attributes. Furthermore, the researcher has little basis for determining both the dimensionality of the perceptual map and the representativeness of the solution since there are not any statistical measures of fit (Hair, Anderson, Tatham, and Black, 1992).

The Compositional or Attribute-Based Approach, unlike the decompositional approach, does produce an explicit description of the dimensions in perceptual space. The compositional methods also provide a direct method of portraying both attributes and objects on a single map.

A primary disadvantage of the compositional approach is that an exhaustive list of relevant attributes is required. If meaningful attributes are omitted, the map could be fallacious and misleading. In an attempt to avoid the omitted-attributes peril, it is not uncommon to create an unreasonable respondent task by including too many or extraneous attributes. The researcher must also make assumptions about how respondents connect and use attributes in order to combine the attributes on a map to represent overall similarity. In fact, one of the major differences across compositional methods is the method used to combine attributes. A final disadvantage of compositional methods is that the researcher is generally not able to produce individual maps.

3. Select Appropriate Technique Based on the Approach Selected

Each approach has numerous available techniques, each with its own specific advantages and disadvantages.

Decompositional techniques include many of those techniques that originated from Bell Laboratories in the late 1960s and early 1970s. A few of the more popular of these techniques include KYST, MDSCAL, INDSCAL, and ALSCAL. The selection of a specific method depends on:

- the nature of scale (ranking versus rating),
- whether similarities data, preference data, or both are obtained, and
- whether individual or composite maps are to be derived.

A discussion of the strengths and weaknesses of each individual technique is beyond the scope of this paper and is presented elsewhere. See Green, Carmone, and Smith (1989), or Schiffman, Reynolds, and Young (1981) for detailed descriptions.

Multivariate compositional techniques can be divided into three basic groups:

- conventional multivariate statistical techniques, such as discriminant analysis and principal components factor analysis, are very useful for developing a dimensional structure among numerous attributes and then representing objects on these dimensions. See Pilon (1989) for a discussion of the issues relating to the use of discriminant analysis versus factor analysis (principal components analysis). Johnson's (1971a, 1971b, 1987) articles include excellent discussions of multiple discriminant analysis.

- multidimensional scaling techniques which utilize attribute ratings as input. Perhaps the most popular of this class is MDPREF which will be described below.

- specialized perceptual mapping methods, most notably correspondence analysis, which provide perceptual maps with only nominally scaled data as input. Walkowski (1990) provides a very readable description of correspondence analysis as a non-parametric substitute for discriminant analysis. Also, Hair *et al.* provide an excellent overview of correspondence analysis. Finally, Mullet (1987) demonstrates some interesting applications.

4. Analysis and Interpretation of Results

This section will include a description of the inputs, mapping algorithm and interpretations of a few techniques that were chosen for their popularity and representativeness of the approaches and techniques that were discussed above.

Decompositional Techniques

MDSCAL is an extremely versatile nonmetric multidimensional scaling program written by J.B. Kruskal and Frank Carmone. Given a set of proximities (similarities) data, it derives a configuration of objects in a pre-specified number of dimensions.

MDSCAL plots only the objects and does not plot the attributes. MDSCAL places the objects in a space of pre-specified dimension so as to minimize "Stress," which measures the "badness of fit" between the configuration points and the data. It employs an iterative search procedure which moves all the points around until it finds the "best" configuration (minimizes stress). Typically, several maps are generated, each of different dimensionality. The stress is then plotted against the number of dimensions (called a scree plot). The dimensionality at which there is little improvement

in the goodness of fit when the number of dimensions is increased (the elbow in the scree plot) is carried into the interpretation stage.

For MDSCAL, as well as for most decompositional methods, the most important issue is the description of the perceptual dimensions and their correspondence to objective attributes. As mentioned above, a weakness of all decompositional techniques is the difficulty in identifying (labeling) the dimensions. In the past, most researchers have relied on subjective interpretations of the dimensions. However, more objective procedures do exist. PROFIT, a computer program/ algorithm, is a useful program which utilizes attribute ratings for each object and then finds the best correspondence of each attribute to the derived perceptual space.

One of the most powerful aspects of the classic multidimensional scaling techniques is that, even though the inputs are usually non-metric (rank-order data), the outputs are metric (comparisons of absolute distances between points are meaningful). Also, multidimensional scaling techniques usually, but not always, fit an appropriate model in fewer dimensions than multivariate statistical techniques such as principal components analysis and discriminant analysis (to a lesser extent).

As an aside, MDSCAL has been merged with a program known as TORSCA into a program called KYST. KYST includes the powerful initial configuration procedure from TORSCA as well as the great generality of MDSCAL.

Compositional Techniques

MDPREF is a scaling algorithm in which, like discriminant analysis, the objects are represented as points and attributes as vectors in the same space. While the MDPREF is metric since it utilizes Eckart-Young decomposition (a principal components analysis), it does not make the types of distributional assumptions often associated with classical multivariate statistical methods (like discriminant analysis or factor analysis). The solution is simultaneous rather than iterative, usually producing results that are more stable than other joint space models.

MDPREF produces a point-vector map to portray three sets of relationships:

- relationships among the objects: the nearer the objects are together in perceptual space, the greater the perceived similarity among them,
- relationships among the attributes: the closer the attributes are to each other, the stronger the associations among them, and
- most importantly, relationships among the objects and the attributes: the further the objects are in the direction of the head end of the attribute arrows, the more it is credited with possessing the attributes associated with each vector. To compare objects along any one vector, simply mark a position on the arrow by dropping a line perpendicular to the arrow from each object position.

Note that MDPREF is an extremely versatile program. When the inputs are paired comparisons, it is considered to be a decompositional approach; when the inputs are direct preference judgments it is considered to be a compositional approach.

Discriminant Analysis (Johnson, 1971a) requires:

- perceptions be homogeneous across respondents,
- attribute data be scaled at the interval level,
- attributes be linearly related to one another, and
- the covariance matrix (the amount of disagreement) be the same for each object,

but it appears to be particularly robust with respect to violations of these assumptions and solid in application.

While these assumptions may seem constraining, they do enable maps provided by Discriminant Analysis (DA) to have the following useful properties:

- given the additional assumption of multivariate normality, there is a test of significance for distance between any two objects,
- maps are not dependent on the inclusion (or exclusion) of objects. If an object was deleted from the analysis, the remaining objects would have the same relationships to one another and to the attribute vectors, and
- the solutions are unique; the technique does not have the problems associated with local optima, as do many non-metric techniques.

Discriminant Analysis derives the map by first finding the weighted combination of attributes which discriminates most among objects, maximizing an F ratio of between-object variance to within-object variance. The second and subsequent weighted combinations are found which discriminate maximally among objects, with the constraint that they all be orthogonal to one another. Having determined as many discriminating weighted combinations as can be found, average scores for objects on these discriminant dimensions may be used as coordinates to plot objects on a map. Attributes are placed on the map with respect to their relative correlations with the discriminant functions.

Discriminant Analysis produces a point-vector map which is interpreted in the same manner as MDPREF maps (see above).

For Discriminant Analysis, as well as for most compositional methods, it is important that the perceptual map be validated against other measures of perceptions, since the map is a result of the particular attributes that were specified by the researcher. Applying DA to a holdout sample and testing its classification hit rate is one way that this might be accomplished.

Correspondence Analysis is a recently developed interdependence technique that facilitates dimensional reduction and provides for perceptual mapping. Correspondence analysis (CA) differs from other interdependence techniques in its ability to accommodate both nonmetric and nonlinear relationships. It provides a multivariate representation of interdependence for nonmetric data not possible with other methods. In its most basic form, CA uses a crosstabulation of two variables as input. It then transforms the nonmetric data to a metric-level form and performs dimensional reduction similar to factor analysis. It also performs a form of perceptual mapping similar to multidimensional scaling, where categories are represented in the multidimensional space.

In its most general form, Multiple Correspondence Analysis (MCA) requires as input a crosstabulation of more than two variables in a multi-way matrix. MCA relates the frequencies for any row/column combination of categories to all other combinations based on the marginal frequencies. A value similar to a chi square is obtained and normalized, then a process much like factor analysis defines lower dimensional solutions.

As in factor analysis, the appropriate number of dimensions must first be identified. Derived eigenvalues are available to indicate the relative contribution of dimensions in explaining the variance in the categories.

Once the dimensionality has been established, the map, in which both brands and attributes are represented as points, can be interpreted. The proper interpretation correspondence maps is not as straightforward as one would hope (see Hoffman and de Leeuw (1992) for an in-depth discussion; Whitlark and Smith (1992) for a potentially useful alternative). In the manner in which correspondence analysis is usually conducted, the distances between brands can be interpreted directly:

- the closer two brands are to one another, the more similar they are perceived to be, and
- the closer a brand is to the origin, the "more average" it is perceived to be, and finally
- the closer a brand is to the edge of the map, the more unusual it is perceived to be.

In simple correspondence analysis (two-way contingency table), attributes can be compared to other attributes in an identical manner.

The correspondence of a brand to an attribute category is not, however, represented by the proximity of the categories. In other words, if Brand 1 is closer to Attribute A than Brand 2, it is not proper to interpret that Brand 1 has more of the attribute than Brand 2. Rather, if Brand 1 and Brand 2 are both near Attribute A, they can both be seen as having some of that attribute. Contrary to popular usage, the interpretation of the attribute points is guided by a centroid principle, not a distance principle:

- attribute coordinates are the "center of gravity," or centroid, of brand coordinates having the attribute,
- attributes with low marginal frequency will be plotted toward the edge of the map,
- attributes with high marginal frequency will be plotted nearer to the origin of the map, and
- in the case of Multiple Correspondence Analysis, a variable discriminates better than another variable to the extent that its category points are further apart.

As an aside, see Weller and Romney (1990) for a comparison of correspondence analysis to MDPREF.

CASE EXAMPLE

The data used in the example below were collected in a computer-assisted interview utilizing Sawtooth Software's Adaptive Perceptual Mapping (APM) software. Data from 250 respondents were used to generate the maps that are discussed below. Data were collected on seven brands across 12 attributes. Every respondent rated all brands that they were familiar with on all the

attributes. It was not possible to obtain permission to reveal the product category, brands, and attributes, but it is not really critical given the purpose of this paper.

In the maps below, the coordinates of the various solutions were very carefully plotted, flipped, and rotated to facilitate comparisons. Great care was taken to preserve the relative placements of the brands and attribute points and vectors on the maps. Furthermore, the maps were plotted and printed in a way to preserve the aspect ratios. Freelance Graphics for DOS was used to create the maps.

MDSCAL (Decompositional)
SYSTAT MDS (Multidimensional Scaling) produced by Systat, Inc. was used to generate the map shown in Figure 1. PC-MDS (Multidimensional Statistics Package) KYST (which includes the MDSCAL algorithm) maintained by Scott Smith at Brigham Young University was also used and resulted in a virtually identical map. Since MDS requires a similarities matrix as input, a correlation matrix was created using the SYSTAT CORR procedure. MDS automatically recognizes the type of matrix as well as the type of correlation, and handles it appropriately, if the correlation matrix was created using the CORR procedure, so it is a good idea to use it. MDS reported explaining 94.9% of the variance in two dimensions.

As mentioned above, these procedures generate maps which contain only objects (brands in this case); attributes are not included. The following relationships are easy to discern:

- brands 1 and 7 are perceived to be quite different from each other as are brands 2 and 5 and brands 6 and 5,
- brands 3 and 4 are perceived to be very similar to each other and somewhat similar to brand 5, and
- brands 2 and 6 are also perceived to be somewhat similar to one another.

MDPREF (Compositional)
PC-MDS MDPREF (Multidimensional Preference Scaling) was used to create the map shown in Figure 2. A 7x12 matrix of average ratings of brands on attributes was input. The MDPREF solution reported explaining 64.7% of the variance in the first dimension and 19.5% of the variance in the second dimension for a total explained variance of 84.2%.

This map shows both the brands (as represented by numbers) and the attribute vectors (labeled with letters). It is interesting to note that the brands are positioned very similarly to those in the MDS map with a few minor deviations:

- brand 7 is located much more towards the center of the map, instead of on the edge as in Figure 1, and
- brands 3 and 4 are still perceived to be very similar to each other, but they appear to be slightly more similar to brands 6 and 7 than to brand 5.

In addition to the brand locations with respect to one another, the following statements can be made about the relationships of the attributes to one another:

- attributes I and K are very different from each another, and somewhat different from all of the other attributes as well, and
- attributes A,B,C,E, and F are very similar to one another.

Finally, the following observations can be made about the relationships of the brands to the attributes:

- brands 1 and 5 are perceived as having attribute I, whereas brand 6 is seen as having attribute K,
- brand 2 is not seen as having attributes G, I, or K and is negative on the rest of the attributes, and
- brands 3, 4, and 7 do not have a well-defined image, since they are all positioned near the origin.

Discriminant Analysis (Compositional)

Adaptive Perceptual Mapping (APM) developed by Sawtooth Software Inc. was used to produce the map in displayed in Figure 3. Since the data were collected by APM, the data were already in the proper format for the APM analysis module. APM utilizes the individual respondent attribute ratings on brands, as opposed to MDPREF, which uses the average respondent attribute ratings on brands. By proportioning the F ratios for all dimensions and multiplying them by the total variance accounted for by the principal components, it was calculated that APM explained 38.3% of the variance in the first dimension and 24.5% of the variance in the second dimension, for a total explained variance of 62.8% in the first two dimensions.

As can be seen by comparing Figures 2 and 3, the results are quite similar for MDPREF and APM with a few exceptions:

- in APM, the brands are more spread out. This is the expected result since the objective of discriminant analysis is to derive the dimensions in such a way that they discriminate maximally across the objects,
- the position of brand 7 in the APM solution corresponds more with the MDS solution than with the MDPREF solution. This may be explained by the fact that respondents were the least familiar with brand 7. The different placement of brand 7 may have something to do with the way that the routines handle missing values,
- although attributes G, H, and J are pointed in somewhat different directions, the rest of the attributes are pointed in directions very similar to MDPREF. This result is not surprising since these attributes have the "shortest" attribute vectors in both maps, which means they are not represented well in the first two dimensions. All attribute vectors are the same length in n-dimensional space; they only appear to be shorter if they are pointing into or out of the two-dimensional plane that appears in the figure.

Correspondence Analysis (Compositional)
Mapwise from Market Action Research Software was used to generate the map displayed in Figure 4. PC-MDS CORAN was also applied and produced a practically identical map. The maps were created by recoding the attribute ratings into a two point scale (the brand has the attribute, or it doesn't). A 7x12 matrix containing the number of respondents that said that each brand had each attribute was used as input. Incidently, a second solution which utilized a matrix of mean attribute ratings yielded an almost identical map. Mapwise reported explaining 58% of the variance by the first axis and 28.5% of the variance by the second, resulting in a total of 86.5% of the variance explained in two dimensions.

Once again, this map is very similar to the others:

- the placement of the brands seems to correspond most with the APM solution,
- the attributes that point in a similar direction in the APM map are all relatively near one another in the MCA map,
- the attribute points that correspond to the "shortest" vectors in APM are closest to the origin,
- keeping in mind the "center of gravity", or centroid, rule of interpreting brand/attribute relationships, the brand/attribute interpretations are almost identical to those that one would make from the APM map.

FIGURE 1 - SYSTAT MDS

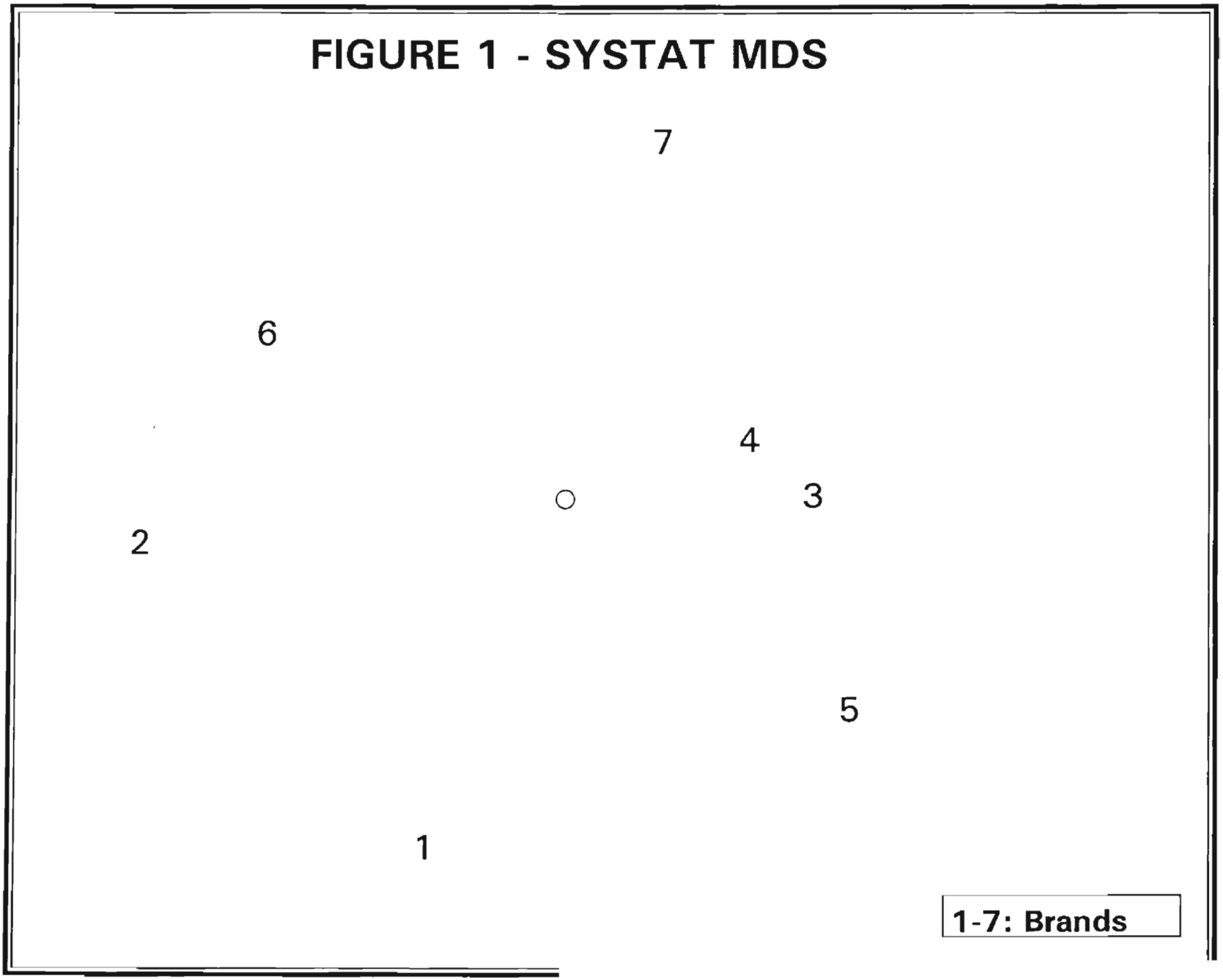
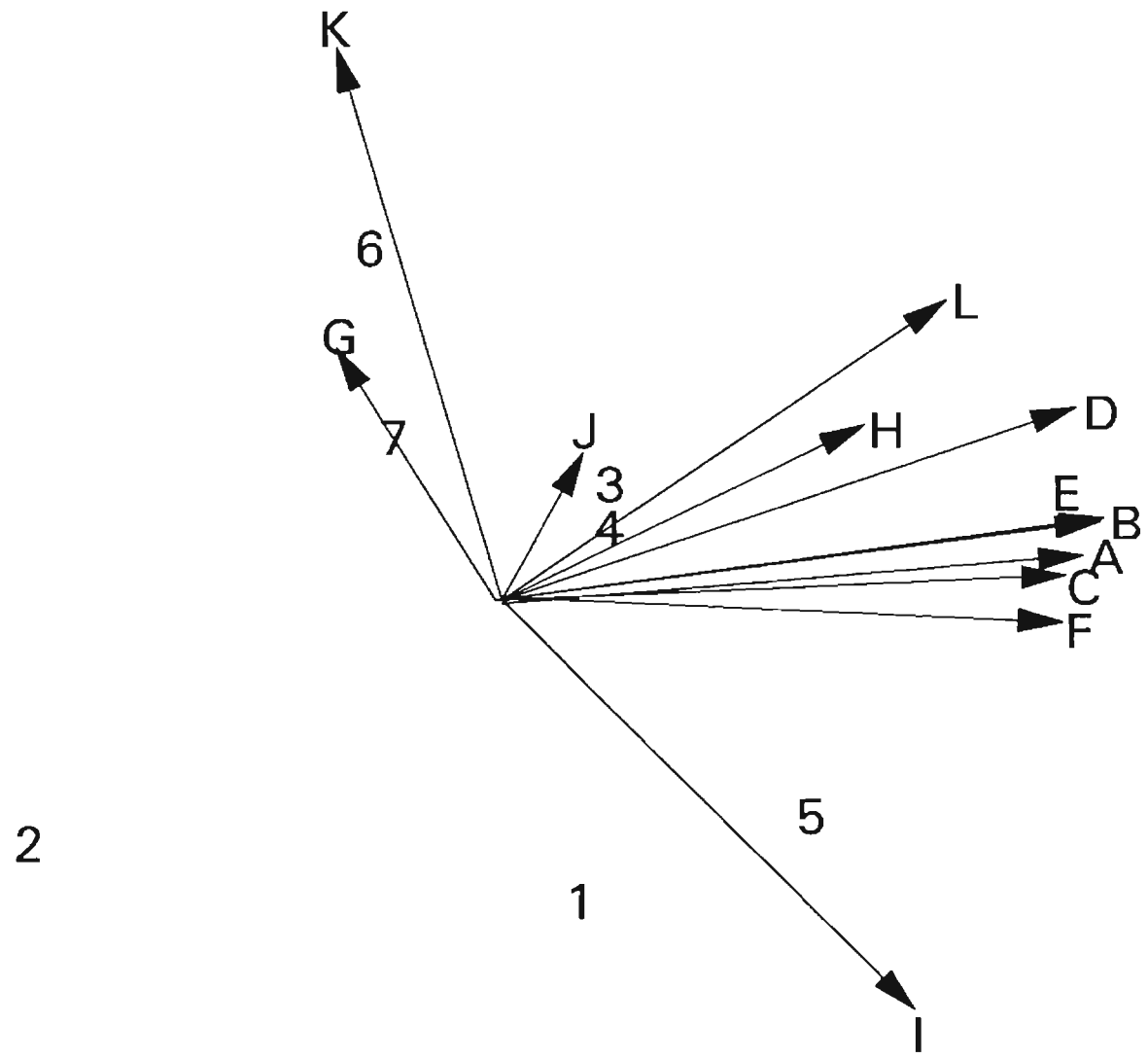


FIGURE 2 - MDPREF



A-L: Attributes
1-7: Brands

FIGURE 3 - APM

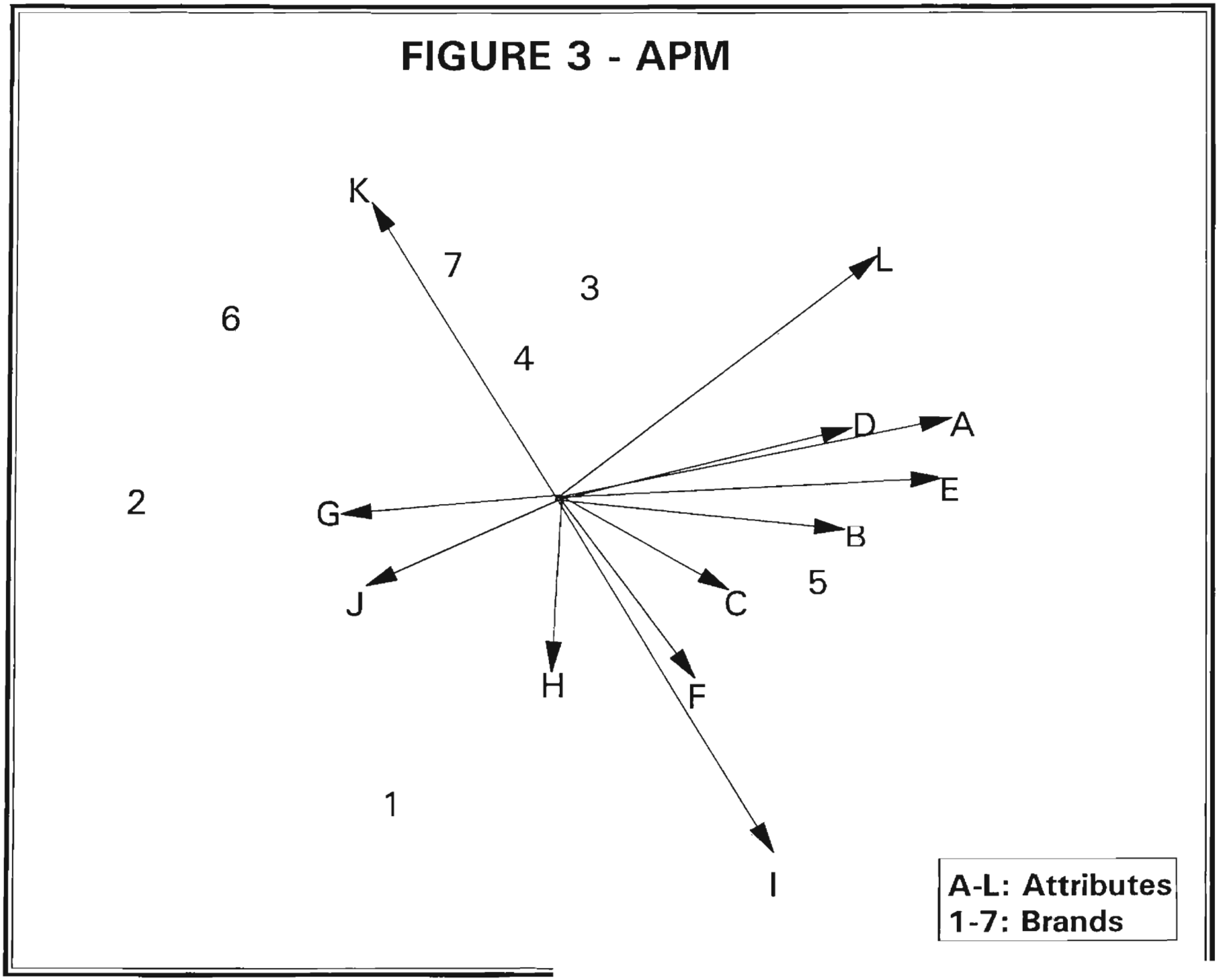
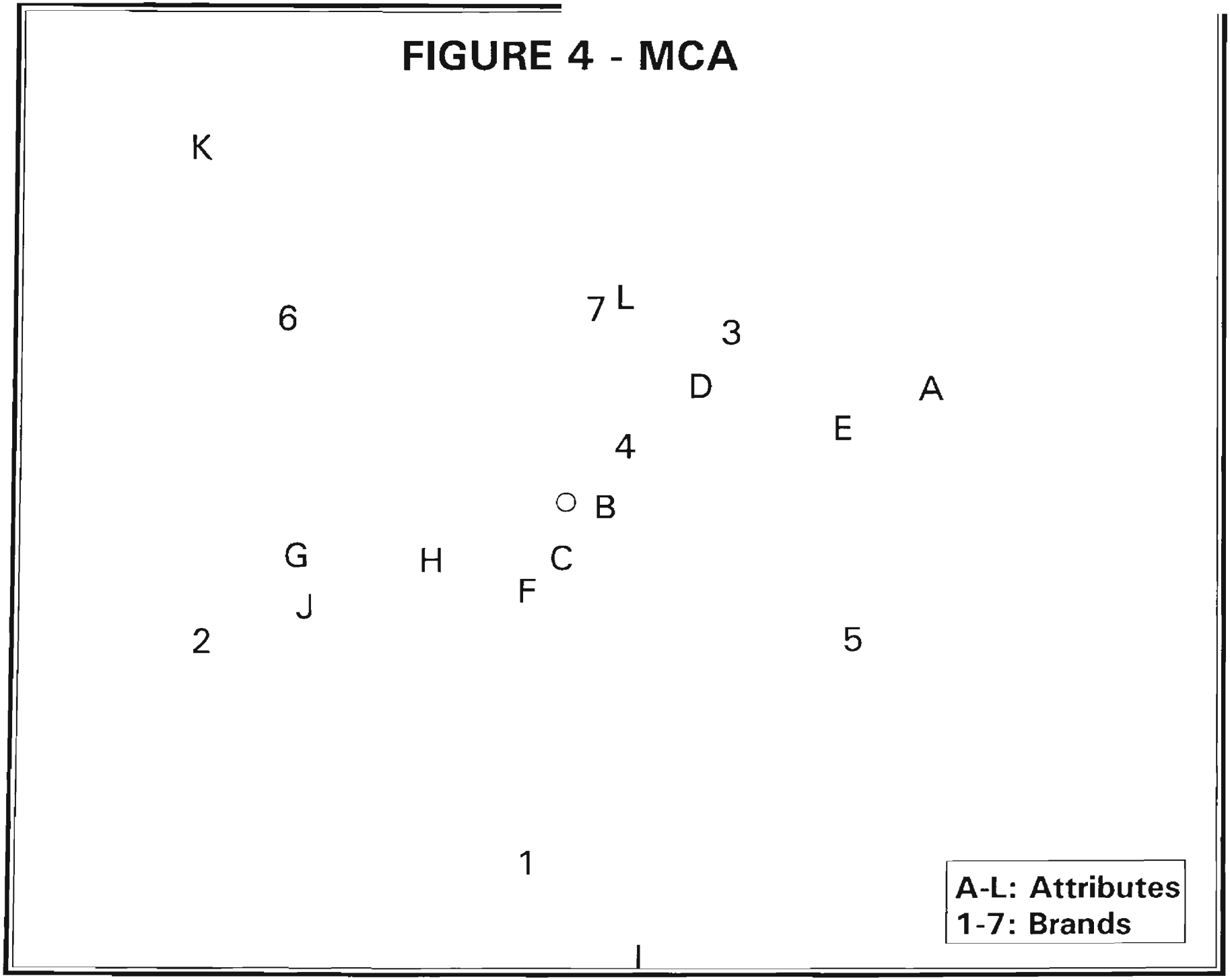


FIGURE 4 - MCA



DISCUSSION

The case example demonstrates that, for the most part, the four techniques yield very similar results. The differences can be largely explained by the differences in the level of inputs, the underlying objective of the techniques, and the way that missing values are handled. So, if the output is fairly similar, the question becomes how should one decide which technique to use?

One criterion for deciding which technique to use should be the ease of data collection. MCA allows the most flexibility here, in that it is hard to imagine data that can not be transformed in some way so they are appropriate for input. MCA is also particularly strong in being able to map data that have already been summarized and the raw data are difficult or impossible to obtain. APM, through its adaptive nature, has cleverly circumvented what used to be one of the biggest drawbacks of DA: the difficulty of collecting ratings on a large number of brands and/or attributes. However, the usefulness of the adaptive data collection algorithm is not limited to DA. It could be used to collect data to be used in the other techniques, as well.

Another criterion for deciding which technique to use should be the degree to which the theoretical underpinnings of the technique agree with the way the researcher desires to interpret the data. The goal of DA, which is to generate dimensions that will discriminate or separate objects as much as possible, seems most appropriate for most marketing studies.

Another criterion might be the confidence that a researcher has in the solution. Although some measures of statistical significance are available for MCA, DA is the only technique of the four that is truly statistical. Although MDPREF requires fewer assumptions about the nature of the data and still creates vector maps that look like discriminant maps, it lacks the statistical underpinnings of DA, and DA has been shown to be very robust with respect to violations of its assumptions.

There are many other more pragmatic criteria. Considerations such as sophistication of the audience, availability of software, and technical ability of the researcher are beyond the scope of this paper.

In conclusion, I like Discriminant Analysis. However, I value Multiple Correspondence Analysis for those situations in which DA is not usable or appropriate. Also, by using more than one technique, one generally gains considerable insight as to how to interpret each technique's output.

REFERENCES

Green, Paul E., Frank J. Carmone, and Scott M. Smith (1989). *Multidimensional Scaling: Concept and Applications*. Boston: Allyn & Bacon, 1989.

Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1992). *Multivariate Data Analysis with Readings, 3rd Edition*. New York: MacMillan Publishing Company.

Hoffman, Donna L. and George R. Franke (1986). "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research." *Journal of Marketing Research*, vol. 23 (August), 213-27.

Hoffman, Donna L. and Jan de Leeuw (1992). "Interpreting Multiple Correspondence Analysis as an MDS Method." unpublished working paper, University of Texas at Dallas, (January).

Johnson, Richard M. (1971a). "Market Segmentation: A Strategic Management Tool," *Journal of Marketing Research*. vol. 8 (February), 13-18.

Johnson, Richard M. (1971b). "Multiple Discriminant Analysis: Marketing Research Applications," in *Multivariate Methods for Market and Survey Research*, Jagdish Sheth, ed., 65-82.

Johnson, Richard M. (1987). "Adaptive Perceptual Mapping." *Sawtooth Software Conference Proceedings,* 143-158.

Maholtra, Naresh (1987). "Validity and Structural Reliability of Multidimensional Scaling," *Journal of Marketing Research*. vol. 24 (May), 164-73.

Moore, William L. and Edgar Pessemier (1992). *Computer Aided Marketing Planning*. In press.

Mullet, Gary M. (1987). "Correspondence Analysis: A New Tool for Image Studies." *Journal of Professional Services Marketing*, vol. 2(3) (Spring), 41-61.

Neal, William D. (1988). "Overview of Perceptual Mapping," *Sawtooth Software Conference Proceedings,* 151-163.

Pilon, Thomas L. (1989). "Discriminant versus Factor Based Perceptual Maps: Practical Considerations." *Sawtooth Software Conference Proceedings*, 166-182.

Schiffman, Susan S., M. Lance Reynolds, and Forrest W. Young (1981). *Introduction to Multidimensional Scaling*. New York: Academic Press.

Walkowski, Jeff (1990). "Correspondence Analysis: What, When, How." *First Annual Advanced Research Techniques (ART) Forum Proceedings*, Bill Neal, ed., 225-237.

Weller, Susan C. and A. Kimball Romney (1990). *Metric Scaling: Correspondence Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-075. Newbury Park, CA: Sage.

Whitlark, David and Scott Smith (1992). "Improving Correspondence Analysis: A Hybrid Property Model Approach." unpublished working paper, Marriott School of Management, Brigham Young University.

SOFTWARE REFERENCES

APM System for Adaptive Perceptual Mapping
Sawtooth Software, Inc.
1007 Church St.
Evanston, IL 60201
(708) 866-0870

Mapwise Perceptual Mapping Software
Market ACTION Research Software, Inc.
16 West 501 58th Street, Suite 21-A
Clarendon Hills, IL 60514-1740
(708) 986-0830

PC-MDS Multidimensional Statistics Package
Institute of Business Management
Brigham Young University
Provo, Utah 84602
(801) 378-5569

P-STAT
P-STAT, Inc.
230 Lambertville-Hopeville Rd.
Hopewell, NJ 08525
(609) 466-9200

SYSTAT
SYSTAT, Inc.
1800 Sherman Avenue
Evanston, Illinois 60201
(708) 864-5670

Comment on Pilon

Katie Klopfenstein
MarketVision Research, Inc.

This is a valuable paper, because it provides an overview of important topics for marketing research practitioners. Pilon's paper reviews the two basic types of mapping, the analysis methods and software available, their data requirements, and interpretation of perceptual maps generated through different techniques. I would like to focus on two areas where I have questions or would like to have more information, and then discuss other considerations in choosing a mapping technique.

QUESTIONS REGARDING THE PAPER

Although I understand why Pilon could not identify the product category, brands, and attributes, it would have enhanced what I learned from the paper if I had been able to interpret the maps with these in mind. As we all know, our judgment and the judgment of our clients are among the most important tools we have in interpreting a perceptual map. Because the category was not identified, it was difficult to assess which map was "best" in this case.

It was not entirely clear to me why it is inappropriate to develop associations between brands and attributes by looking at a map developed through correspondence analysis. With maps generated through other compositional techniques, we can drop perpendicular lines from the attribute vectors and rank order the brands with regard to their level of association with each attribute. I would like additional information regarding why it is not appropriate to do so with correspondence analysis.

OTHER CONSIDERATIONS IN CHOOSING A MAPPING TECHNIQUE

Although this paper provides theoretical reasons, advantages, and disadvantages of the two approaches (compositional and decompositional), I would like to list some additional, pragmatic considerations.

1. **How well the product category is known by your client.** If this is a new category for your client's firm, a decompositional technique may be more appropriate because it would be difficult (or impossible) to generate a comprehensive attribute list. In addition, it can be extremely informative to interview selected respondents after they complete a survey and ask them why two brands are similar (or dissimilar) and what basis they use for comparison.

2. **Rate of change in the product category.** If the product category is changing rapidly, it may be difficult to devise a salient attribute list. In addition, revolutionary change in a product category can introduce a new dimension and make our perceptual maps obsolete overnight. For example, consider a map of the toothpaste category before the introduction of Crest. Crest redefined the product category by introducing a new dimension, fluoride protection/cavity prevention. As another example, consider the computer industry before the introduction of the personal computer (PC). Before the development and evolution of PCs, physical machine size meant

computing power, and this is no longer the case. For a rapidly changing product category, a decompositional technique may be more useful.

3. **The sophistication of your audience and their ability to think creatively**. Maps generated through decompositional techniques can require more creativity in interpretation. This can be a benefit, because it encourages discussion about how respondents view the product category. However, some clients have difficulty working with a map where the dimensions are not already defined in some way.

LINKING CONJOINT ANALYSIS AND PERCEPTUAL MAPPING

Roger Gates *and* ***Mike Foytik***
DSS Research

INTRODUCTION

An obvious way to test the predictive validity of a conjoint solution involving an existing product or service category is to apply the results, via simulation, to the existing brands or product types in the market. The goal is normally to determine how well we are able to predict actual market shares for the brands/types. If the resulting simulations produce market share estimates that are fairly close to actual shares, then there is a tendency to conclude that the conjoint estimates are valid. Conversely, if the estimated shares differ markedly from the actual shares, then the conjoint estimates are viewed with suspicion.

One of the problems in implementing a test, such as the type described above, is the question of matching product/service profiles (particular configurations of attributes/levels) with actual brands or product types. In conjoint exercises we typically we have respondents evaluate product/service alternatives described in fairly objective terms. For our test simulations, we could objectively measure the brands or product types in the market on the attributes used for the conjoint exercise and use the attribute levels determined through the measurement process to represent each brand/ type. The problem with this approach is that the market may not perceive a particular product or brand to have the levels of each attribute that it actually possesses. Much of modern marketing is concerned with shaping consumer/buyer perceptions of brand/product types and for this, and other reasons, the perceptions of customers may not correspond with the actual characteristics of products/brands measured in an objective sense.

This paper grew out of a project recently completed by the authors in which this particular problem was encountered. The details of the particular application and the solution developed are provided below.

BACKGROUND

The application in question involved a project for a major provider of telecommunications services. The project dealt with the commercial market for telecommunications services and covered the four major types of telephone service used by business customers:

- Standard business lines
- Key systems
- PBX systems
- Centrex service

The goals of the research were to determine the service attributes used by business customers to evaluate alternatives, the relevant levels of these attributes, and the relative importance of each attribute level in the decision making process.

METHODOLOGY

The client had determined in their RFP that conjoint analysis would be the appropriate analytical technique to accomplish the goals of the research. We concurred in the assessment and developed a research design that included the following elements:

- *Qualitative research.* The goal of the qualitative research was to identify the attributes used by business customers in assessing telephone service alternatives and to determine realistic levels of these attributes for purposes of the research. To achieve this goal we conducted a series of eight focus groups (two with customers with each of the four types of service) and a series of in-depth interviews. On the basis of this research nine salient attributes of business telephone service were identified.

- *Quantitative research.* The goal of the quantitative research was to quantify the relative importance of the nine salient attributes and their associated levels identified in the qualitative research. The quantitative research proceeded into two phases:

  - Preliminary telephone interview. The first phase of the quantitative research employed a telephone survey. A probability sample of business firms in the area served by the client was purchased from Survey Sampling. Screening questions were used to determine which of the four types of service (standard lines, Key systems, PBX or Centrex) the respondent company had chosen as their primary telephone solution. After this was determined, respondents were questioned regarding telephone system usage and parameters, firmagraphics and other classification data. Finally, they were recruited to participate in a disk-by-mail survey. We determined whether or not they had an IBM-compatible PC and the floppy disk size for that PC. Those who did not have access to an IBM-compatible PC were asked to complete an extended version of the telephone survey which included Sawtooth Software's ACA System (for conjoint analysis) and APM System (perceptual mapping) modules.

  - Disk-by-mail survey. Those with access to an IBM-compatible PC (83.2 percent of all respondents) were sent a disk-by-mail survey the day following the initial telephone interview. They were offered a significant incentive to complete that survey. Completed disks were returned by 78.4 percent of these respondents. The disk-by-mail survey included two major sections. First, they were asked to complete an ACA exercise covering those same nine attributes. Finally, they were asked to complete an APM exercise covering the four basic types of service (standard business lines, Key systems, PBX and Centrex). In this exercise they were asked to rate the four systems on the same nine attributes.

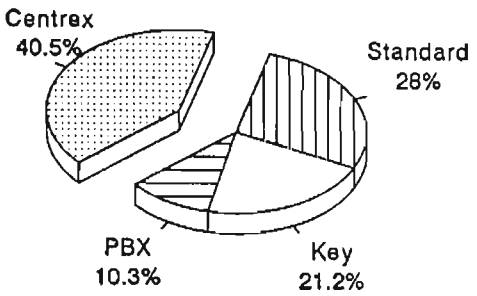
ANALYSIS AND RESULTS

The initial qualitative research indicated that there might be discrepancies between customer perceptions of Centrex service and the perceptions of knowledgeable sales reps and management personnel regarding Centrex. For this reason, the APM module was added to the questionnaire to

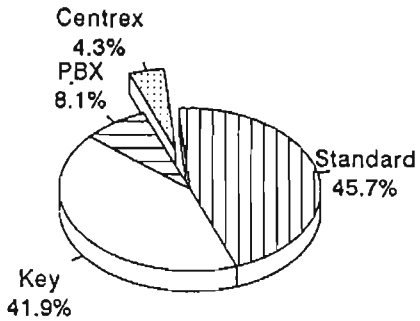
quantify customer perceptions of the four types of service. The APM data turned out to play a key role in the analysis for this project. Without these data many of the conclusions and recommendations of the study would have been impossible.

The ACA Choice Model 3 Simulator was used to predict shares for the four types of service using the estimated partworths and product profiles judged by management to describe the four types of telephone service. However, the initial market simulations produced estimates that were radically different from actual market shares. For example, Centrex service was known to have an actual market share of only 3.9 percent. However, the simulation using a product profile for Centrex judged by management to accurately represent the characteristics of this service on each attribute produced an estimated share of 40.5 percent. Results for Centrex and the other three types of service, showing actual market share and simulated share, are summarized in Figure 1. Keep in mind that the profiles used to represent each of the four types of service were selected by management. We were faced with the problem of explaining why our estimates deviated so widely from actual shares.

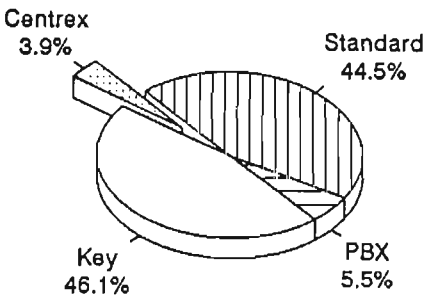
Figure 1
Predicted Market Share
Simulations on Management & Consumer Perceptions



Management Perceptions



Consumer Perceptions



Actual Market Share

Note: Products for ACA simulations using Model 3, which corrects for product similarities, were based on product designs from mgmt./consumer perceptions.

Evaluations of awareness and ratings of the four types of service on the nine attributes provided some clues. The awareness data suggested that business customers have the lowest level of awareness in regard to Centrex service. The ratings data suggested that business customers have a tendency to perceive Centrex service as being more expensive and less feature rich, by a wide margin, than is actually the case.

The solution was to use the APM results to pick product profiles for the four types of service that correctly represented customer perceptions of the characteristics of each service. This new set of profiles was used in the ACA simulator and the resulting shares and actual shares are also shown in Figure 1. The market share predictions based on the profiles chosen on the basis of customer perceptions of the four types of service closely match the actual market shares for the different services. Also note the substantial differences between the share estimates based on customer perceptions and the share estimates based on management perceptions. It should be noted that, in all cases, results have been adjusted to reflect the stratified sampling approach used in the research design.

APM RESULTS

Figures 2 and 3 show the disparity between management and customer perceptions of the four types of service on two basic dimensions: *Cost of System* and *Feature Richness*. Customer perceptions are based on the APM data collected in the disk-by-mail exercise. Management perceptions are based on a survey of client personnel with responsibilities for various types of business telephone services. Although standard business lines are perceived similarly by both groups, the differences in perceptions of the other three types of service (Key, PBX and Centrex) are substantial. Except for current Centrex customers, Centrex service is perceived as being the highest priced of the four telephone systems. Additionally, Centrex is perceived to offer much less in the way of features and options than either Key or PBX. Customer perceptions of Key and Centrex service are nearly opposite those of management.

Furthermore, when customer perceptions are analyzed by the type of primary telephone system the customer has, the differences are even more striking. Only current Centrex system users perceive Centrex to be feature rich and relatively less expensive than Key and PBX systems. In fact, Centrex customers viewed the system as lower priced than even standard business lines and more feature rich than PBX systems.

SELECTING PROFILES FOR ACA BASED ON CUSTOMER PERCEPTIONS

We used the APM results in several ways. First, ratings data gathered for the APM portion of the interview were used to evaluate the estimated partworths and to choose more realistic profiles for the ACA market simulations. Importance ratings for each attribute in APM were compared to the sums of utilities for each attribute. Adding the utilities for each level of an attribute produces a measure of the overall utility of that attribute. A sum of 100 is considered average. The rank order of the attributes and the magnitude of the differences between attributes are very similar using the APM importance ratings or the ACA sums of utilities.

FIGURE 2
Map of Consumer Perceptions of Telephone Systems

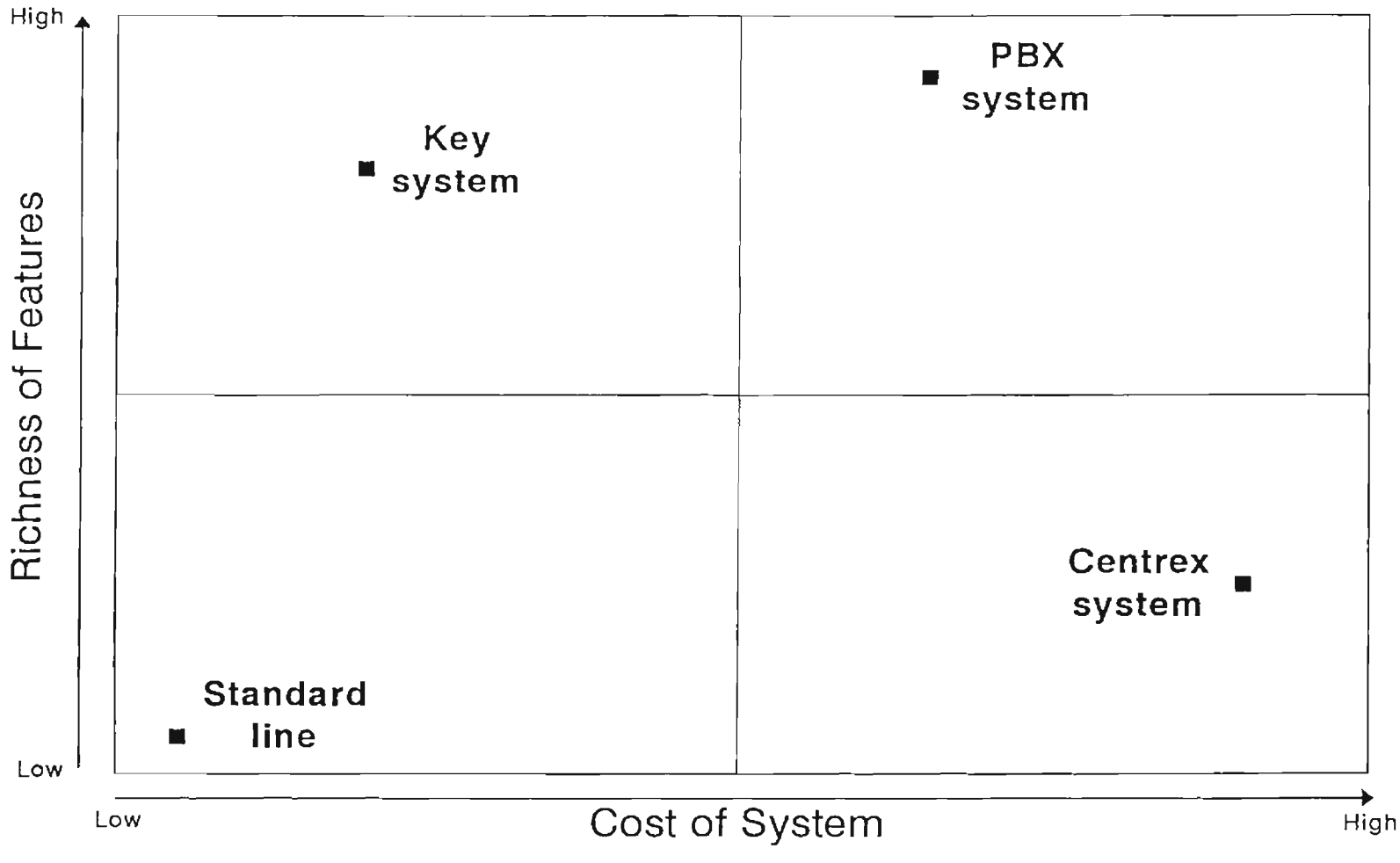
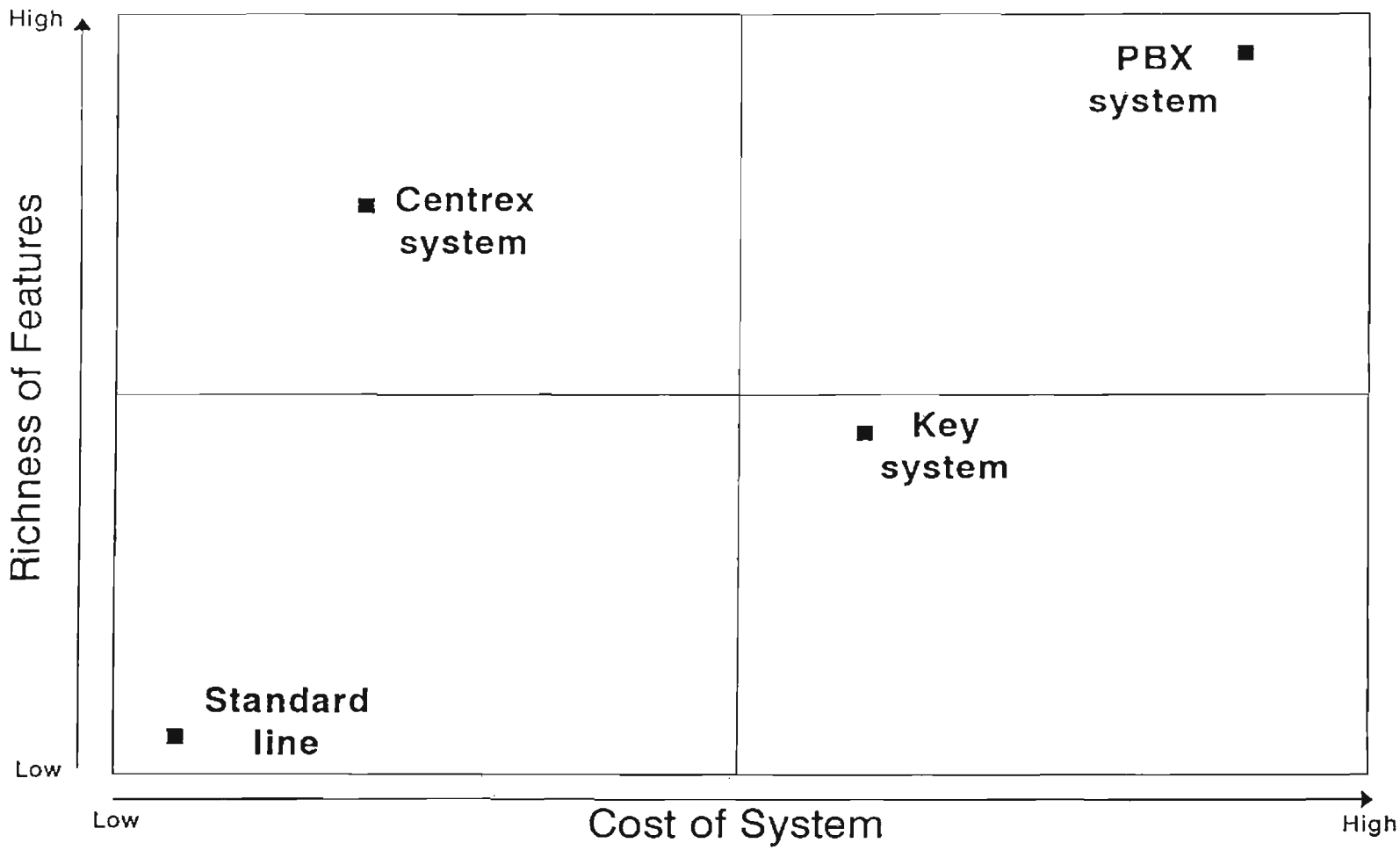


FIGURE 3

Map of Management Perceptions of Telephone Systems



Next, APM ratings of selected products on selected attributes were used to select appropriate profiles for each type of service for the ACA simulations. These ratings were instrumental in the selection of profiles that matched *customers' perceptions*. As noted above, customer perceptions differed greatly from management perceptions of the characteristics of each type of service. Without the perceptual data from APM, product specifications would have been based solely on management's interpretation of how each product should be defined.

Several steps were required to incorporate APM ratings of products on attributes into ACA product specifications. These steps are detailed below:

- **Step 1.** Rescale the average ratings of each product on each attribute to a 100-point scale where the highest rating (among all products and all attributes) is assigned 100 points and the lowest rating (among all products and attributes) is assigned zero points. This scales the ratings in a manner similar to the *presentation* form of utilities in ACA simulations and the scaling used by the POINTS program in ACA. The formula for scaling each rating is:

$$\textit{Scaled rating} = \left[\frac{\textit{Original value} - \textit{Minimum value}}{\textit{Maximum value} - \textit{Minimum value}} \right] \times 100$$

- **Step 2.** Compare the scaled ratings from step 1 to the average utilities computed by ACA (in presentation form). For each attribute, if the scaled rating of any product on that attribute is greater than the utilities of every attribute level, the APM ratings for that attribute must be scaled a second time. The purpose of this second scaling was to bring the scaled ratings within the range of each attributes' computed utilities. This guards against the use of extrapolated values for any attribute level. The scaling factor is applied to all service types, even those whose ratings are less than the maximum utility for any service type for that attribute. This step is necessary because of the very small average utilities assigned to attributes that have little influence on overall preference (where the sum of all attribute levels is well under 100). The scaling factor for Step 2 is:

$$\textit{Scaling factor} = \frac{\textit{Maximum attribute level utility}}{\textit{Maximum product rating}}$$

- **Step 3.** Once APM product ratings on each attribute are appropriately rescaled (Steps 1 and 2), the third step involves the interpolation of perceived attribute levels for each service type based on the rating of each service type on that attribute. If a scaled rating for a service type is exactly the same as the average utility for one of the levels of that attribute, the service type in question is specified using the corresponding attribute level. When a scaled rating falls between two attribute levels, the attribute level to be used in the service type specification for ACA must be interpolated. The linear interpolation formula is given below:

$$\textit{Interpolated level} = \left[\frac{\textit{Scaled rating} - \textit{Utility}_1}{\textit{Utility}_2 - \textit{Utility}_1} \right] + \textit{Attribute level}_1$$

- **Step 4.** The final step involves an evaluation of the profiles selected for each service type by means of the procedure described above. The computed attribute levels for each product should be reasonable and defensible in light of reality. If adjustment needs to be made (for example, service is described with an attribute level that is impossible for that service type or any other service type in the market), the relative differences between the service types on that attribute should be maintained. Adjustments must be made in the attribute levels for all the service types in the same direction and the same magnitude to maintain this relationship.

The service profile specifications that result from this four-step process should have the following properties:

- The highest rated service on any attribute will have an attribute level which gives it a higher utility than any other service on that attribute.

- The greater the difference between any two service-type ratings on an attribute, the greater the difference in the utilities associated with the attribute levels specified for each of the service types.

- No attribute levels will be extrapolated beyond the range of the predefined attribute levels.

- The greater the difference between utilities for any two attribute levels, the greater the difference in service-type ratings on that attribute that will be required to move from one attribute level to another.

CONCLUSIONS AND COMMENT

This research clearly suggests that, when dealing with existing brands or product types, researchers should look to consumer perceptions of the brands or product types in the process of selecting profiles to represent them in conjoint simulations. This is particularly true in those situations where there is strong evidence or reason to believe that there are significant differences between market perceptions of the product or service and the objective characteristics of the product or services. Because we work with objective descriptions in regard to attributes and attribute levels in conjoint exercises, there is often a tendency to assume that brands are perceived objectively by consumers. Incorporating perceptual mapping into the research design provides a reality check.

This is particularly true when management has strong, and possibly erroneous notions, regarding the way different brands or product types are perceived. The perceptual maps show how the brands or products are actually perceived and the data required to produce the maps provide a basis for selecting the appropriate profiles. A method for using the data to select profiles is outlined in the paper.

Comment on Gates and Foytik

Karlan J. Witt
IntelliQuest, Inc.

I would like to applaud the authors for undertaking an effort which is the ultimate expression of applied research: incorporating reality into the research process. In the case presented in the paper, reality is the perception of the respondents.

One point mentioned briefly in the paper that I would like to underscore is that the preference from a conjoint simulator may differ from market share. This may occur for many reasons:

- lack of awareness of products
- lack of awareness of features or benefits
- lack of availability of products
- local discounting or promotions
- incompatibility with installed equipment
- influence of others at the time of purchase (salespeople, VARs, consultants, friends, coworkers)

Any of these circumstances may cause differences in predicted and actual preference, but do not necessarily trigger a product change. Oftentimes the issue is one of communication rather than product design, and a focused communications campaign may be the appropriate response. Collecting both product and positioning information empowers the client to address either need.

One concern I have with surveys which incorporate both conjoint and perceptual mapping tasks is the length of the survey, and the resulting data quality. Both tasks are tedious for respondents and may wear them out more quickly than a simple time estimate for the survey would otherwise indicate.

I do, however, think there is value to combining the two. The biggest question, then, that the paper raised in my mind was the rescaling technique used to generate the base case specifications for simulations.

Although different techniques have been used in the past, the most intuitive to me is to measure both the conjoint and perceptual mapping on the *same* scales. For instance, if you use "above average service and support," "average service and support," and "below average service and support" in the conjoint exercise, then use that same scale for collecting perceptual data.

Below are examples of a common implementation of a perceptual question and an example of an alternative when combining conjoint and perceptual mapping:

An example of a standalone perceptual mapping question:

"How much do you agree that IBM personal computers have above average service and support?"

Strongly Disagree								Strongly Agree
1	2	3	4	5	6	7	8	9

An example of a perceptual mapping question used in conjunction with conjoint:

"How would you rate the service and support of IBM personal computers?"

Above average service and support
Average service and support
Below average service and support

The scale for perceptual data may be drawn out to capture greater differentiation and still reflect the conjoint attribute levels. For instance, using a 9-point scale where "1" is below average service and support, "5" is average, and "9" is above average. This is not suitable for all attributes, however, and needs to be evaluated for each study.

By collecting the evaluations on the same scales, you avoid any differences simply due to scaling, as well as anything introduced by the rescaling techniques. It lets the market tell you exactly how they perceive a particular brand or product in the terms you are using to simulate market preferences and responses to change.

Finally, the authors had the clients take the survey to compare results to the market. I suggest that this is a healthy exercise, and can become fun for the participants. IntelliQuest has had success devising a contest where different parties at the client firm respond to the survey as they expect the market to respond, and the party closest to the actual responses wins some prize.

While this process allows the client to become comfortable with the survey instrument and the methodology being employed, this also helps the client develop hypotheses early on in the analysis process, providing the researcher with theories to confirm or disprove with the market data.

I would like to see more of this type of work done to further explore the relationship of preferences from alternative measurement techniques, and to use the market's perceptions as base cases, or "reality," and to explore opportunities within that framework.

INTEGRATING CONJOINT RESULTS INTO DECISION MAKING

Louise Minor
Goodyear Tire & Rubber Company
Katie Klopfenstein and ***Robert V. Miller***
MarketVision Research, Inc.

OVERVIEW

As marketing researchers we are faced with three basic responsibilities. First, we must be certain that the research design we recommend is consistent with the overall needs of our clients. Second, we must be certain that the capabilities of the design are clear in that the research provides intended results and does not over-promise. Third, the results must be easily converted into action steps that drive management decision making forward in a clear, action oriented path.

All too often the concept of tradeoff analysis is understood only in a cursory way by our clients. That is, there is an understanding of what conjoint does fundamentally, but no real understanding of the power of the technique and often a misleading idea of the results and what can be done with these results to bring about change with the product or service under investigation. Moreover, much of the literature is technical and does little to increase understanding of how this technique can be used effectively in management decision making.

One of the best approaches to explaining conjoint analysis to non-researchers is to use a case study approach where examples of successfully executed conjoint studies can be discussed, including how the results, were used to make changes in the product or service. Unfortunately, the number of case studies showing problem definition, research implementation, and most importantly, use of results are few. Consequently, we can show non-researchers only a modest number of examples of how this technique is used and can provide only limited information and examples on how results have been used in management decision making.

The purpose of this paper is to discuss how conjoint research can be conducted within a company that is unfamiliar and somewhat uncomfortable with the technique and to show how the results can be incorporated into the management decision-making process.

This research project was conducted during the summer of 1991 and deals with an industrial product that is presently considered a commodity. The overall goal of the research was to determine if the physical product could be differentiated from competing products in a meaningful way. As a result, physical attributes of the product as opposed to non-physical attributes (such as customer service) were investigated as a means of product differentiation. All interviews were conducted with high level managers (both customers and non-customers) in a plant environment. Through this case study we will demonstrate how conjoint research can be made more actionable by carefully managing the research project process and the way results are presented.

GAINING UNDERSTANDING OF CONJOINT

Demonstrating why conjoint analysis was the best technique for the product differentiation project was quite challenging. Given that many levels of management and staff had to either support the project and/or utilize the results, the justification of the conjoint methodology was more than an one-time effort. In addition, conjoint was not a widely recognized technique and required explanation. Of those who were aware of this research method, a concern existed that it was too complicated and too expensive.

This project was assigned by the vice president of the division. It reflected his response to the business manager's, the marketing manager's and the business team's assessments of the product's market position and their corresponding requests for assistance. Given that so many individuals had a vested interest in this project, the conjoint approach had to be introduced and sold to individuals in the order of the organizational chart, from the vice president to the business team members.

GAINING CREDIBILITY AND "BUY IN"

With the conjoint analysis being proposed and discussed multiple times, a rather successful informal presentation evolved. It was based upon the most common questions and objections that arose, "We already know what our customers want. Why should we go through all this?" Further probing of these issues uncovered that the division was keenly aware of which product attributes their customers desired. In fact, more than 22 attributes were identified by internal participants. However, which attributes would have the greatest impact upon market share and overall profitability could not be determined.

Explaining how conjoint analysis would more closely replicate the customers' decision-making process was an effective response to these issues. The concept of customers making tradeoff decisions about the product attributes was easily understandable and a Sawtooth Software demonstration diskette helped us demonstrate the interview process. The resulting utility scores were presented as a benefit which would help focus planning efforts. Only the more technically inclined (R&D and technical support staff) had additional questions.

For a few individuals, the price tag of another division's conjoint project was another stumbling block. Although this fee included services beyond the conjoint research, some remembered only that it was conjoint project with a significant price tag. Assurances had to be made that costs would be minimized.

Political forces were also a determining factor as to whether conjoint analysis would be embraced. Success hinged upon identifying and understanding these forces and any individual's hidden agenda. In this particular case, the credibility of market research as a function and the allocation of resources (including human resources) were key hurdles to overcome.

Therefore the conjoint approach was successfully sold through getting the internal customers to voice their objections and by using feature — benefit presentations. Our next task was to obtain the resources necessary to conduct the project.

COMMANDING ORGANIZATIONAL RESOURCES TO SUPPORT THE PROJECT

The importance of executive support cannot be overlooked. This was particularly helpful because resources were scarce. It assured that at least a certain level of resources would be made available, both in terms of budget approval as well as granting human resources. Existing staff were asked/ required to devote time to this project to: 1) participate with questionnaire and sample preparation, 2) facilitate the interviews, and 3) participate with short- and long-range planning. This affected many areas of the business, including sales, marketing, technical support, R&D and manufacturing. Such an internal level of participation not only enhanced the quality of the project but also helped ensure buy-in of the results.

In the name of saving company dollars, two Goodyear employees were chosen to facilitate the in-person, laptop interviews. These individuals were part of an internal training program in which new hires rotate among different product and functional areas. Management agreed to pull these two employees from their current assignments to save external interviewing costs. The rationale recognized that their involvement would be mutually beneficial and rewarding. It gave the trainees additional insight into market research, the industry and the product. It also provided them with the opportunity to work directly with customers. In return, we obtained cost-effective interviewers.

This methodology decision was not made lightly. Many factions were involved in weighing the pros and cons of this innovative approach to such a costly aspect of the methodology. The dangers of significant interviewer bias and a higher than normal level of interviewer error were major considerations. The logistics, intensity and content of the training effort were carefully considered. The management of the data collection effort was also scrutinized.

The interviewer training consisted of many phases. Approximately one-half day was spent on the areas of professionalism, bias and how to handle unique questions/issues. Interviewer instruction sheets were developed to guide the interviewers from start to finish. An interviewing checklist was also prepared to ensure details were handled properly (see next page). The list ranged from being sure that the laptop batteries were charged to when to document unique situations.

Management of the daily data collection was accomplished via the interviewers' maintenance of master interview schedules and the routine faxing of lists of completed interviews and interviewer notes. Initially, the interviewers made daily phone check-ins.

Interviewing Checklist

A. Before you leave to conduct the interview:

- Be sure the battery has been charged.
- Allow enough travel time and have directions and phone numbers.
- Take extra interview diskettes as well as the DOS diskette.
- Take laptop, including extension cord, adaptor, battery pak, and power cord.
- Take attribute definition sheet, master schedule, tablet and pens, and a map.

B. Computer setup:

- Check the room for location of outlets, least glare, and seating arrangements.
- If there is a problem:

1. Check the power cord/power supply.
2. Did you boot up with the interview diskette in? Remove the diskette and reboot without the diskette in.
3. Try a different interview diskette (remember to maintain/change master schedule).
4. Turn the system off and start over with the floppy removed.
5. When all else fails, call Katie at _____.

C. Interview

- Fill in respondent number on master schedule
- Take notes when:

1. a respondent refuses to answer a question (be sure to note why)
2. a respondent has questions and/or complaints regarding Goodyear. Remember to listen closely and tell the respondents that <u>their</u> questions/concerns are very important. Let them know you are taking notes to make sure that you communicate back to the plant accurately.
3. if anything is ambiguous or confusing to the respondent during the interview. If the respondent looks confused, jump in and ask "how's it going?" to help the respondent through the problem.

PRESENTATION OF RESULTS

To complete the project, we made three separate presentations, each with a different purpose. The first presentation was made to the cross-functional business team at the plant. The purpose of this presentation was to give those most closely linked with the product and its future "the first crack" at the data and their meaning. This presentation was fairly informal, and debate about the meaning of the findings was encouraged.

The second presentation was at the corporate headquarters level. The group in attendance included three people from the plant and representatives from other product lines within the division. The primary purpose of this presentation was to expose other marketing and business managers to the conjoint technique and show how one group planned to utilize the information. The questions we received at this presentation were more related to methodology than results, because the individuals wanted to understand the process and how it might apply to their business and the decisions they needed to make.

After the HQ presentation, an action planning meeting was held with the managers most involved with the product. The majority of the 10 participants were from headquarters and provided the division with research and development services. (This corporate department serves many divisions, which must compete for this valuable service.) The purpose of this meeting was to translate the research findings into priorities and action steps for the product team. Participants were asked to "shift gears" and think creatively about the business in both the short as well as long term and to consider strategic issues. Robert Miller (from MarketVision) functioned as group facilitator, guiding the discussion and charting the issues. The discussion outline used in the meeting is shown on the following page. The first step was to outline external threats to the business. Even "longshot" threats, such as product obsolescence, were considered and discussed by the group. Any internal threats, dealing with the way the business was being managed on a day-to-day basis, were considered next. In light of the current competitive marketplace, opportunities available to improve the position of Goodyear and its product were listed. Internal resources and capabilities available to combat the threats and take advantage of the opportunities were also listed. Taking into account these threats, resources, opportunities, and the research findings, general goals for the business were outlined. From this list of general goals, an extensive list of specific goals was created. The meeting participants then individually voted on their top seven specific goals. Through this voting process, a priorities list was developed and responsibilities were then assigned.

This process allowed those with a direct influence on product development to internalize the research results. It also led them to consider how the findings could be applied today and in the future, through the company's long range planning process. The presentation was designed to move people — instead of simply saying "the customer wants 'x' and I don't know how we can do that," participants were brought together in a setting where ideas could flow freely. The customer needs identified through the research were considered as a starting point, and developing a plan for action based on the findings was the primary focus.

Planning Meeting Outline

I. External Threats
II. Internal Threats
III. Opportunities
IV. Internal Resources and Capabilities
V. General Goals
VI. Specific Goals
VII. Priorities
VIII. Resource Allocation
IX. Timing

The entire research process was managed with the implementation stage in mind. Gaining understanding and buy-in from all parties early in the process assured us of a group eager to implement the findings. We also defined the responsibilities and proper roles for internal marketing research analyst and supplier at the outset of the project to take full advantage of each others' strengths.

PRICE-SENSITIVITY MEASUREMENT OF MULTI-ATTRIBUTE PRODUCTS

Dirk Huisman
SKIM Market and Policy Research (The Netherlands)

SUMMARY

Due to accelerating technological developments, products can be extended with a range of new features, functions and product claims. In view of these opportunities the strategic pricing question discussed in this paper reads as follows: "Which combination of products has to be offered when, where, and at what price range, to realize the full profit potential?"

Conjoint analysis is a very appropriate method to measure price sensitivities of multi-attribute products. This paper discusses the merits and demerits of various tradeoff techniques for the measurement of price sensitivity, namely:

A. The price(s) of the products are specified as an attribute of the product.
   1. Using absolute prices (for instance $ 2000; $ 2500; and $ 3000).
   2. Using respondent specific prices (such as base price + $ 500 or + 5%).

B. The attributes are priced separately.
   1. Each attribute is priced separately at various levels; priced attributes are traded off.
   2. Each attribute is priced separately at various levels and the products traded off are priced as the sum total of the prices of the attributes.

C. The two-step approach: price is an attribute which is related to a group of attributes. The individual utilities of these attributes are measured in a separate tradeoff process.

Depending on the number of attributes and on the stage of the strategic pricing process, the researcher can use one of the conjoint analysis techniques outlined above. Within their limitations all models generate data to measure price sensitivity. Applied on their own, but especially in combination with other research techniques, they are tools that help the marketer analyze which product differentiations will be worthwhile and will help realize better prices.

INTRODUCTION

In this paper I will discuss various conjoint analysis methods to measure price-sensitivity. The application of these methods is described and discussed primarily with regard to multi-attribute products that have a technological content.

As a consequence of accelerating technological developments, the development and restructuring of markets is becoming increasingly technology-driven, finding expression in:

- shorter production time
- changing cost structure and break-even points
- improved and highly differentiated products tailored to many market niches
- the introduction of ranges of new products
  shortening product life cycles.

In view of this development, the strategic pricing question should now read as follows:

Which combination of products has to be offered when, where, and at what price range, to realize the full profit potential?

PRICE SENSITIVITY MEASUREMENT

Although very interesting and relevant, in this paper we will not pay attention to the dynamic pricing element, the "when" of the strategic pricing question. We will focus on the question: "Which products have to be sold at what price range to which buyers?"

To sell a product, or to realize a transaction, the price of that product has to be less or equal to the product's value as perceived by the buyer. This perception is influenced by a number of factors. In conjoint analysis we try to measure the value and to evaluate the individual's choice process. In interpreting the value measured in the conjoint analysis we have to take into account the factors that influence the perception. Nagle (1987) has described nine factors influencing the perception of a product's value or its price sensitivity (summarized in Box 1).

As these factors may be of influence, it is not only in the interpretation that one has to take these factors into account, but also in the conjoint design.

> Instance:
> Measuring the value of a drug for patient A, a patient who does not suffer very much, there may be a great many alternatives. For patient B, though, there are hardly any substitute drugs available, so for patient B the drug has a much higher value. The choice of the drug is often made by a physician. So, when measuring the value of a drug it is very important to define the patient for the physician. For patient A the physician will be aware of many substitute drugs, but not so for patient B. Consequently, the value of the drug when measured with patient B in mind will be higher (factor 2 in Nagle's list).
>
> It is not only the patient that has to be specified, but the reimbursement situation will have to be dealt with as well. Some drugs are reimbursed while others are only partially, or not at all. The choice process will be different when the drug is not reimbursed. When the drug is reimbursed the physicians are generally less sensitive to price.

Apart from the situational factors, it is also the nature and the phasing of the actual choice process which has to be taken into account when designing the tradeoff process. With conjoint analysis we measure one tradeoff process. But, the buying decision is often made stepwise, certainly in the

industrial environment. At each step price may play a role and this role may differ per step. So, the way we measure the sensitivity to price, or the impact of price during the choice process, must differ as well.

Box 1: Nine factors that commonly influence a buyer's price sensitivity, as identified by Nagle.

1. The UNIQUE VALUE EFFECT
   - Does the product have any unique attributes that differentiate it from competing products?
   - How much do buyers value those unique, differentiating attributes?

2. The SUBSTITUTE AWARENESS EFFECT
   - What alternatives do buyers have (brands and products)?
   - Are buyers aware of the alternative suppliers or substitute products?

3. The DIFFICULT COMPARISON EFFECT
   - How difficult is it for the buyers to recognize the attributes that make the product "different"?
   - Can buyers make easy comparisons of price-offers, or are there many sizes and combinations which make a comparison difficult?

4. The TOTAL EXPENDITURE EFFECT
   - What percentage of their incomes do buyers spend on the product (for consumer products)?
   - How much is this in dollar terms?

5. The END-BENEFIT EFFECT
   - What benefit is important to the buyers, and how price-sensitive are they to this benefit?
   - What portion of the benefit does the price account for?

6. The SHARED COST EFFECT
   - Do buyers have to pay the full price or do they have a possibility to reduce the cost by reimbursements or tax deductibility?

7. The SUNK INVESTMENT EFFECT
   - Must buyers of the product make complementary expenditures in anticipation of its continued use?
   - For how long are buyers locked in by those expenditures?

8. The PRICE-QUALITY EFFECT
   - How important is a prestige image and can this be established by using price?
   - Is it possible to ascertain the price-quality ratio before purchase?

9. The INVENTORY EFFECT
   - What is the size of the inventories of customers?
   - What are the expectations of the buyers regarding price developments?

MEASURING PRICE SENSITIVITY WITH THE HELP OF CONJOINT ANALYSIS

Conjoint analysis can be applied in various ways to measure the price sensitivities of multi-attribute products.

A. The price(s) of the products are specified as an attribute of the product.
   1. Using absolute prices (for instance $ 2000; $ 2500; and $3000)
   2. Using respondent specific prices (such as base price + $ 500, or + 5%)

B. The attributes are priced separately.
   1. Each attribute is priced separately at various levels; priced attributes are traded off.
   2. Each attribute is priced separately at various levels and the products traded off are priced as the sum total of the prices of the attributes.

C. Groups of attributes are related to price. The role played by the individual attributes is part of a separate tradeoff process.

A 1. Price is an attribute of the product, specified in money terms.

In most tradeoff studies in which price is included, price is specified as one of the attributes. As the buyer of a product normally perceives and evaluates a product as a whole and then weighs the price of that product, this way of treating price is often perceived to approach reality best. A typical tradeoff question with price as attribute is shown in box 2.

Box 2: A typical tradeoff question with price as attribute.

Tradeoff with Price as Attribute

Please specify your preference for camera A or camera B, assuming that all other features of these two cameras are the same.

Camera A	Camera B
• Canon • US$ 305 • autofocus • high-speed flash	• Minolta • US$ 290 • autofocus with zoom • standard flash

Strong Preference for A	1	2	3	4	indifferent 5	6	7	8	9	Strong Preference for B

The demand for a product can be estimated by "weighting" one of the preference models for each respondent. Using the Share of Preference model, Smallwood (1991) describes the demand function in the market as:

$$D(k) = \sum_i w_i \left\langle \frac{\text{Exp}\,[\,u_i(p_k) + u_{ik}\,]}{\sum_j \text{Exp}\,[\,u_i(p_j) + u_{ij}\,]} \right\rangle$$

D(k) = total demand product k
Exp [ ] = Exponential function as described in note 7
$u_i(p)_k$ = utility for the price of product k where i = resp.i
u_{ij} = utility of all other attributes for each product in the market
w_i = weight reflecting the importance or number of customers represented by respondent j

From this equation it is clear that the utilities measured can be used to form a demand function. By changing the price of product k the utility for the price of product k will change, and so will the demand estimated with the specified demand function. Relating the percentage change in demand to the percentage change in price gives the "price-elasticity" for the kth product, which is specified by Smallwood as:

$$E(k) = p_k \frac{dD(k)/dp_k}{D(k)}$$

It will be clear that when all the data are available cross elasticities can be calculated as well.

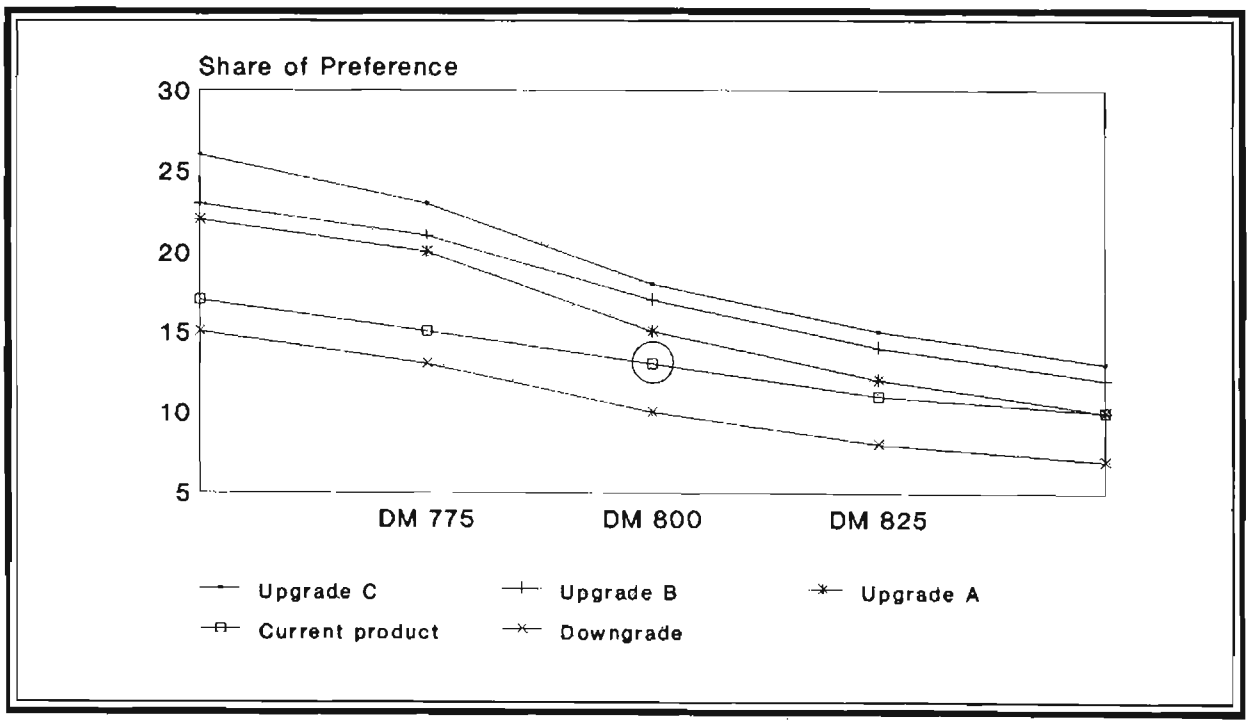
A 2. Price, specified in relative terms or using respondent specific prices, as an attribute of the product.

There are many transactions where price is not a fixed entity but depends on factors such as transport costs, quantity purchased, relative power of the buyer, and so on. This is especially true in some industrial markets. In the industrial market one has to specify a broad price range and in this range the relevant sub-range of price levels has to be isolated for each respondent. If prices diverge a lot, it is advisable to work with a base price (which is individual- specific) and a range which is defined as "base price plus or minus 2%; 4 %; 6%", or defined as "base price plus or minus DM 50.-; DM 100; DM 150."

Instance:
In a study on behalf of a producer of raw materials used in the refractory industries we had to use this technique because the prices of the relevant raw material from the various suppliers ranged ex-factory from DM 650 to DM 1300 per ton. This difference was partly caused by the differences in specification of the raw materials supplied by the various suppliers. In addition to these ex-factory prices came the transportation and handling costs, differing per country and supplier. But it was not only the differences in cost which caused the variation in price. It was also the size and the importance of the buying organization that made the perception of the "reasonable price level and price range" more or less buyer specific. As the tradeoffs have to be realistic, we had to use a respondent-specific specification of the price range (base price plus or minus DM 25; DM 50).

In addition to price, five product specifications were included in the tradeoff (for example, crystal size and density). Knowing all the specifications of the products on the market it was possible to include all these products in the simulation model and to calculate the shares of preference for all these products. These shares could be compared with actual market shares to check if the model reflected the actual market. Differences had to be explained and could be explained by the intangible factors not included in the tradeoff, such as "preference for manufacturer" and "reliable delivery." Based on the "tested model" it was possible to estimate the share of preference and the demand at various price levels. It now became also possible to estimate the effects of a change in the product specification at various price levels.

Figure 1: Shares of preference at different price levels for the current product and for upgrades and a downgrade of the product (using Share of Preference model with correction for product similarity).



Limitations:

Though very appealing, there are limitations when using the technique which treats price as one of the product's attributes.

As was stated in the introduction, due to technological progress, products can be extended with a number of new features, new functions and product claims to fulfill all kinds of additional needs. As a consequence products have to be specified on a great number of attributes. Each attribute is part of the product and may contribute to the value of the product as perceived by the buyer. Combining the attribute "price" with 10 to 20 or even more attributes will bias the price-sensitivity measurement. In case the respondent is trading off only a few attributes at a time (combinations based on tradeoff tables, or on the preference data already collected [adaptive approach]) these attributes have to justify the price difference of the total product. As in reality, the price difference is based not on a few but on an extensive range of attributes, price sensitivity may be overestimated.

Another limitation we sometimes experienced is that respondents tend to get exasperated: they feel that they are not being taken seriously when asked to tradeoff the price difference of the total product with only one or two changes in the specifications.

The limitations related to price as one of the attributes are also relevant for the full-profile approach. When a respondent is trading off products specified on 15 attributes at a time (which is near to full profile) the respondent's evaluation will include only a limited number of attributes. During the tradeoff process the respondent will perceive selectively. The selected attributes are evaluated in combination with price. It is, however, not clear on which attributes the respondent's evaluation is based and, as a consequence, the price sensitivity data may be biased.

> Instance:
> A tradeoff interview for cars included 18 attributes. Observation of respondent Z showed that Z was primarily sensitive to "ABS," "German make," "diesel engine" and "station wagon." So as soon as the profile included a station wagon of German manufacture with ABS and diesel engine, high prices were accepted. If only a few of these primary attributes were included, other attributes were evaluated at a glance and seemed to be included coincidentally, yielding reactions like "this one has ABS and station wagon and that one is a German made diesel; the difference in price is only DM 800, let's see ...look the ABS one has power steering as well, yes I'll take this one." Trading off other pairs of cars, Z hardly noticed "power steering."

In reality the respondent Z of our instance will list all the German station wagons equipped with a diesel engine and ABS first (he might include a Swedish or French Diesel as well) and will only then start to evaluate. The buying process is often stepwise. There is an upper and a lower price limit, which may float a little, and there are certain criteria to be met (attributes that have to be included). Those cars that meet all or most of these criteria are included in the "real life" tradeoff. If the attributes do not change over time and the combinations of attributes in this stepwise process are fixed for groups of buyers, the tradeoff interviews had better be limited to step two, targeted at certain groups of buyers. Due to new technologies, new attributes are likely to influence the evaluation process; combinations of attributes may also affect the composition of the groups of buyers that included the same group of attributes/cars in their evaluation. In addition, flexible production automation enables producers to create a great variety of products (variations of one product) at a cost which is only marginally higher compared with mass production of a standard product.

So the variety of products will increase and it will become increasingly difficult to approach real life tradeoffs, to measure price sensitivity and limit the survey to step two.

An alternative may be to price each attribute separately at various levels. This makes the tradeoffs more realistic, because respondents are shown the consequences of their preferences.

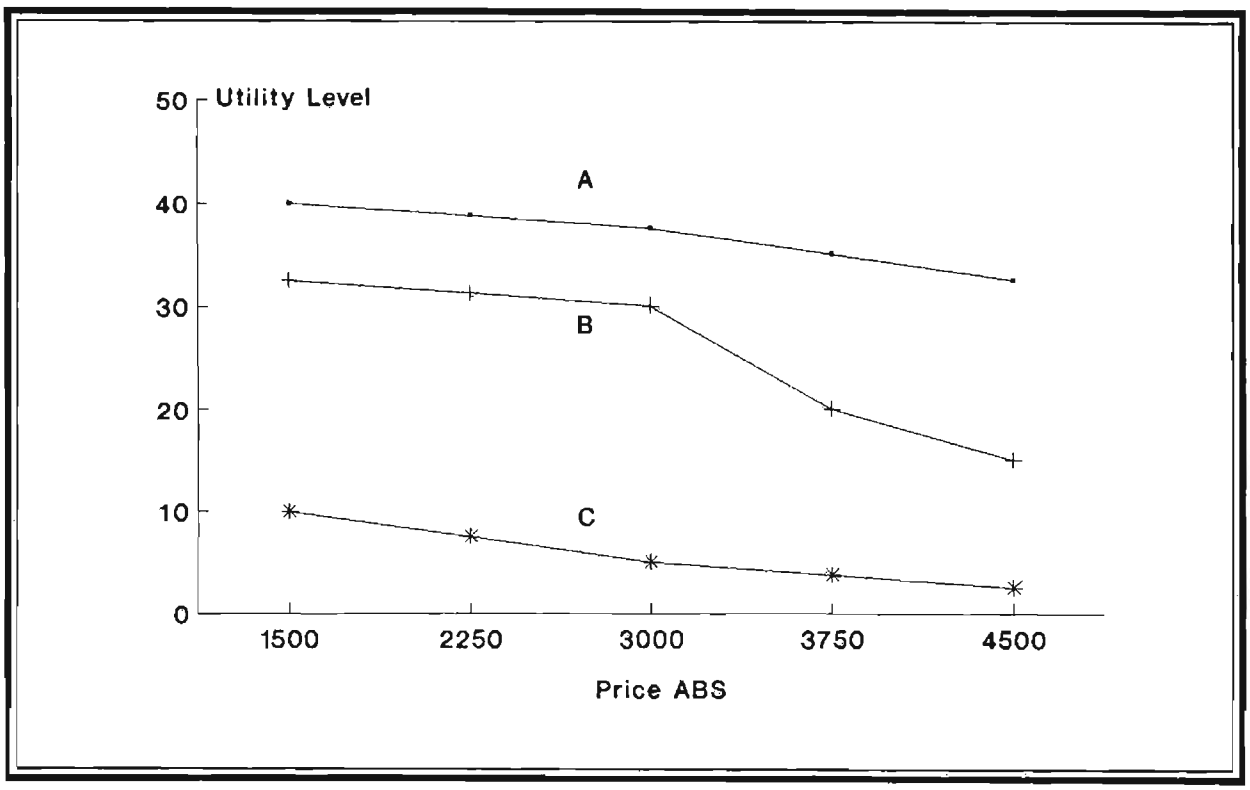
B 1. Each attribute priced separately at various levels.

In this situation, with many attributes, an alternative may be to specify the price the buyer has to pay for each feature separately. To get an idea of the total price one has to specify the base price for a "stripped version" of the product as well and the respondent has to add the price for the extra features to the base price.

Using the adaptive approach and trading off only combinations of a few "priced" attributes at a time after the tradeoff analysis, one learns how price sensitive each respondent is to price changes of each of the attributes.

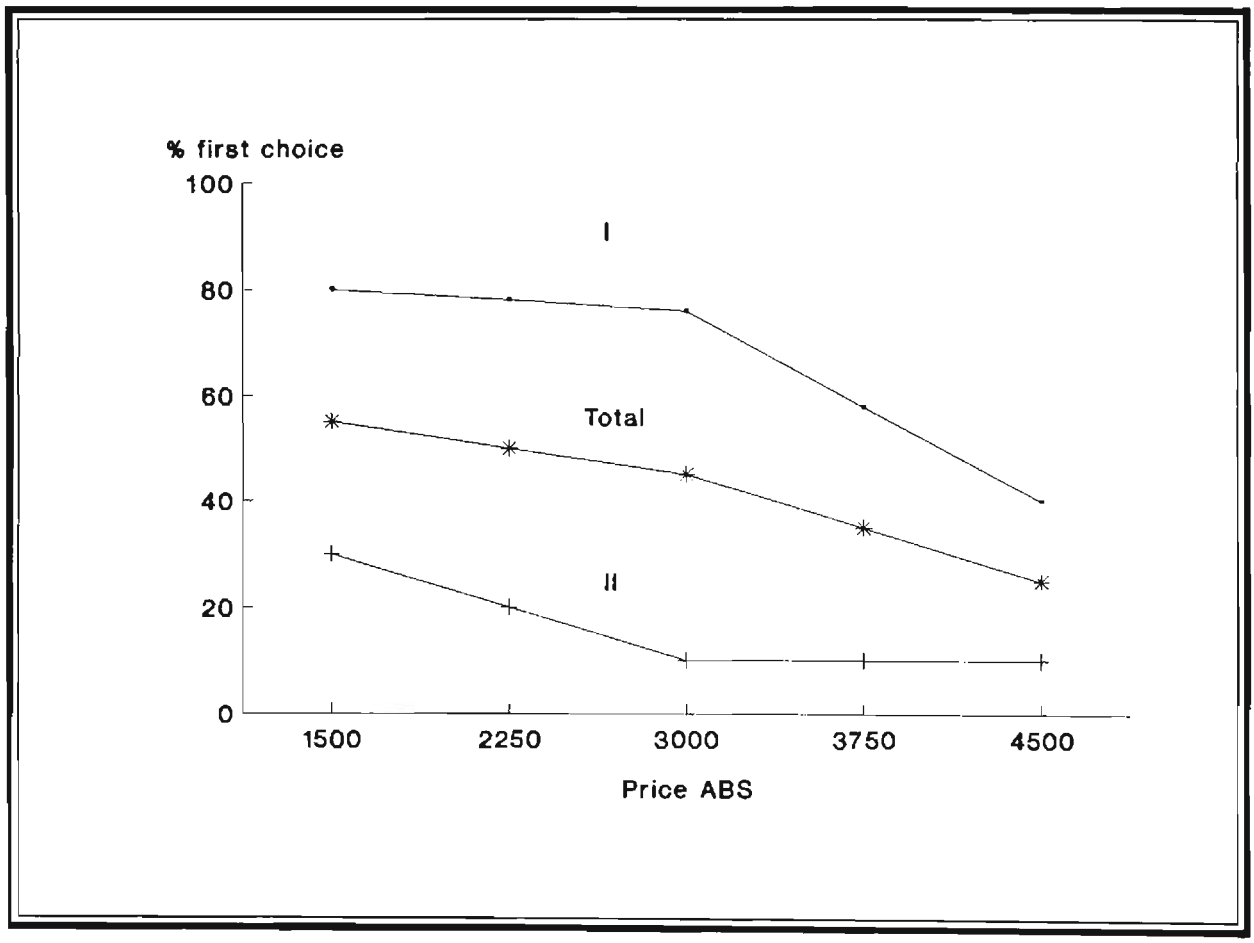
Figures 2a and 2b show the sensitivity to a price change of ABS on an individual level as well as for two groups of respondents. From figure 2a it is easy to conclude that respondent A wants ABS on his car even if he has to pay an additional DM 4500.- for it; respondent B wants ABS on her car only if it costs less than DM 3000,-; respondent C does not want to pay for ABS.

Figure 2a: Utility levels of three respondents for the attribute ABS at three price levels.



For figure 2b we used the "First Choice" model and specified two cars as similar, excepting "ABS." The lines in the graph represent the percentage of respondents that prefer a car with ABS at the price specified. From this chart one can conclude that buyers in segment I are ABS-sensitive and are willing to pay up to DM 3000,-. If they have to pay more, a large proportion drops out. In segment II only a limited proportion is sensitive to ABS but most buyers in this segment do not want to pay for ABS.

Figure 2b: Percentage of "buyers" in total and per segment who are willing to pay the specified price for ABS.



When the utilities for the stripped car (base price) and the utilities for the attributes at a given price are available, the market may be simulated by specifying a range of cars. The problem is that it is possible to make several combinations that result in the same price for the car.

Instance:
Say BMW was included in the tradeoff at a base price of DM 38 000.-. Suppose a BMW can be purchased with ABS, power steering, electronically adjustable seats, electronically adjustable and heated mirrors, and central door lock at DM 43 000,-. The difference of DM 5000 equals the

sum of the least prices specified for each of the five attributes. However it also equals the sum of three attributes specified at a higher price level, which is of interest to the buyers who were willing to pay these higher prices but are not interested in the other two attributes. Other combinations are possible as well.

This example illustrates the complexity of the simulation and the problems one comes across when measuring the price sensitivity for products based on priced attributes only. But technically and theoretically it is possible.

This method also has a number of drawbacks:

1. Normally people investigate within a price range which is acceptable to them. In this case the tradeoffs include only a few priced attributes at a time. These attributes may fit the budget, but in combination with other attributes they can exceed the budget restrictions. So it is possible that respondents accept higher prices sooner or accept prices for less important attributes sooner than they would if the consequences for the total budget should be displayed when answering these tradeoffs.
2. Specifying various prices for an attribute may suggest a "higher price / better quality" relation and as such have an influence on the utilities.
3. To be able to react and give a meaningful response the respondent must be able to perceive the attribute as a separate entity, which can be added on to the product but can be left out as well. In case of attributes like "the side effects of a new drug," it is not possible to "price" the attribute.

Despite these limitations this method, compared with the price sensitivity measured according to method A (price as attribute), generates, if applicable, a more realistic picture of the contributions of the attributes to the acceptance of a higher price of the product. This is due to the fact that the respondent is partly confronted with the consequences of his preference.

B 2. Each attribute is priced separately at various levels and the products traded off are priced as the sum total of the prices of the attributes.

One of the limitations described above can be overcome by cumulating the prices of the attributes, which happens when you use CVA (CVA System by Sawtooth Software). In this case the sum total reflects the price of the product presented to the respondent. The tradeoff questions will be as in box 3. The advantage of this method is that the respondent learns directly if the price fits the budget; this way the respondent is really able to trade off price with the attributes included within the restraints of the budget. The series of tradeoff questions has to include the same attributes, specified at different price levels. As the respondent is only able to take in a limited number of attributes at a time, this method of measuring price sensitivity is, in practice, limited to products with a limited number of attributes.

Box 3: A typical tradeoff question specifying prices per attribute and the total price of the product.

Tradeoff with Price as Sum Total

Please specify your preference for camera A or camera B.

Camera A	
• Canon	$200
• autofocus	25
• high-speed flash	80
	$305

Camera B	
• Minolta	$175
• autofocus with zoom	75
• standard flash	40
	$290

Strong Preference for A	1	2	3	4	indifferent 5	6	7	8	9	Strong Preference for B

The results of this method, in which the prices of the attributes are cumulated, are comparable with the results as shown in figures 2a and 2b. Theoretically this way of measuring the price sensitivity of the product is better than the method described under B.1.

Instance:
In a survey on behalf of a pharmaceutical industry we conducted two conjoint studies. In stage one we conducted a study including 12 attributes including price (method A1), to find out which product improvements would be worthwhile. Based on this study it was possible to isolate the five most important attributes excluding price. These attributes were priced individually and were included in a tradeoff in which all five attributes and the total price were specified simultaneously. The primary aim of this second tradeoff was price optimization. As a part of the samples overlapped we were able to compare the results and especially the price sensitivity data. Both studies generated the same groups of respondents (clusters). One of these groups appeared to be sensitive to a specific product specification, while another group was sensitive to price.

The price sensitivity of the "product specification sensitive" group was virtually the same in both studies, but the results for the "price sensitive group" differed. When we used the first method (A.1) the "price sensitive group" was willing to accept a higher price and was willing to trade off combinations of product characteristics that went beyond their budgets.

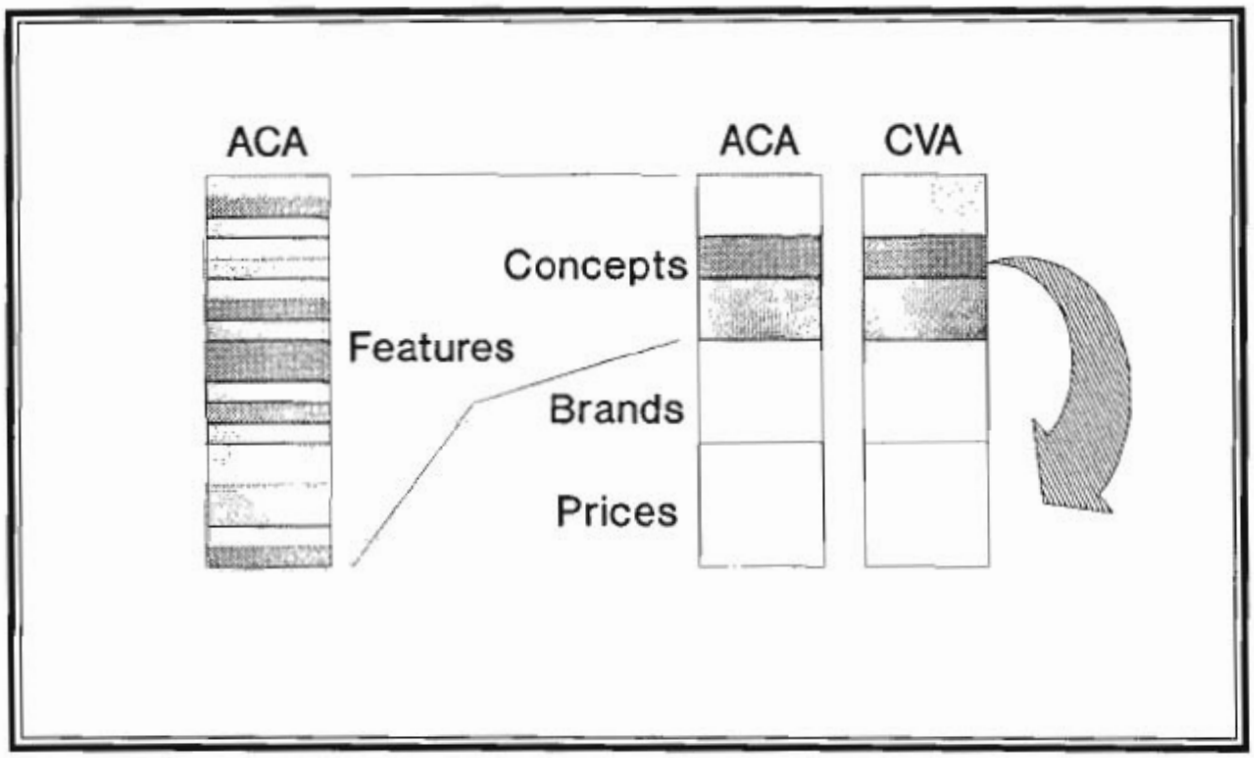
Using the other method (B.2), price dominated the tradeoff for this "price sensitive group." Only beneath a certain price level were they willing to trade off the attributes. In this area a number of attributes could not be traded off because these attributes generated the high prices. Further analysis learned that this "crucial" price level equalled the reimbursement level of the National Health institute.

As it is already difficult to compare five or six attributes or product features at a time, it is even more difficult to compare five priced attributes and the total price of the product simultaneously. In CVA it is possible, and definitely advisable, not to show the prices per attribute, but still to include them in the analyses. This approach resembles the method described under A: the price of the product, though, is not an attribute but is the sum total of the prices of the attributes (prices of the attributes are specified but not shown on the screen; respondents only see the sum total of the prices of the attributes). We have experienced that when the levels of an attribute have identical descriptions and only differ in price, the variation of the utilities measured is very high. Yet, if the descriptions of the attribute levels differ, this way of measuring price sensitivity yields very satisfactory results.

C. The stepwise approach in which groups of attributes are related to price

The limitation of the CVA method as described above is that it is limited to products with a limited number of attributes. In case of many attributes, as in the car example, one has to create a rather complex survey design, or one has to measure the decision process stepwise. At the first step all relevant attributes with price (method B.1.) or without price are included in a tradeoff process. For step 2 groups or combinations of attributes are constructed, which are to be traded off against price.

Box 4: Schematic representation of the two-step approach to measure price sensitivity of multi-attribute products.



Experimental studies by Azalbert of McKinsey London (1992) show that this approach results in a greater importance of price in the choice process. Azalbert's studies also show, though, that the predictive value of the dual and triple ACA (ACA System by Sawtooth Software) approach is higher. If step 2 is based upon CVA and one uses the selection opportunities of Ci2 or Ci3 Systems by Sawtooth Software, it is possible to fine-tune this second tradeoff to the individual situation by, for instance, only showing brands which are part of the evoked set or by showing only those concepts that fit the budget.

PRICE SENSITIVITY MEASUREMENT IN A BROADER PERSPECTIVE

Conjoint analysis is a very important tool for the estimation of price sensitivity. It generates the basic data for further analysis and it helps us understand the buyers. Further analysis with the utilities will help even more in understanding the buyers and in anchoring the pricing strategy.

A further step is clustering on the utilities measured or on the importance of the attributes. This generates groups of respondents which are sensitive to the same attributes or to the price changes caused by the same attributes. Clustering on utilities or on the importance of the attributes generates groups of respondents which are interested in the same benefits and can be labelled as benefit segments. If a company is able to produce a product or to adapt and market a product which fulfills the specific needs of the buyers in this segment of the market, better or higher prices will be realized. Instead of analyzing price sensitivity for the total market it is better to analyze per cluster or market segment, as can be concluded from figure 2b. Analysis of the price sensitivity per segment does not have to be limited to the segments generated by the cluster analysis. Analysis within already-defined segments may yield interesting results as well and will offer opportunities to generate higher revenues.

Value and price sensitivity of a product are not only influenced by tangible attributes. There will always be intangible attributes which cannot be part of the rational tradeoffs. Some of these intangible attributes may be covered by specification of the brand name as an attribute level. The price sensitivity of a brand can be measured, but if brand is included as attribute in the tradeoff it is advisable to measure the brand image and the attributes underlying this image as well, in the same interview. This will help explain the value and price sensitivity of a brand name.

CONCLUSION

One of the consequences of technological developments is that it is now possible to create a range of products or to rig a product with various product claims, product specific features, or functions. It is the marketer's task to specify at what price this range of products has to be put on the market to realize the full profit potential.

Conjoint analysis is a tool with which the researcher can measure price sensitivities of multi-attribute products. Depending on the number of attributes and on the stage of the strategic pricing process, the researcher can use one of the four conjoint analysis techniques outlined above. Within their

limitations, all models generate data to measure price sensitivity. Applied on their own, but especially in combination with other research techniques, they are tools that help the marketer analyze which product differentiations will be worthwhile and help realize better prices.

This paper was also presented at the Second SKIM Seminar in Rotterdam, the Netherlands, in May 1992 and is included in the Proceedings of that seminar: Marketing Opportunities with Advanced Research Techniques, Rotterdam, 1992, SKIM Market and Policy Research.

REFERENCES

Azalbert, Xavier (1992). "About Micro and Macro Attributes: Suggestions for Advanced Market Modelling." Rotterdam, 2nd SKIM Seminar: 'Marketing Opportunities with Advanced Research Techniques'.

Nagle, Thomas, T. (1987). *The Strategy & Tactics of Pricing — A Guide to Profitable Decision Making*. Englewood Cliffs, NJ, Prentice Hall.

Smallwood, Richard D. (1991)."Using Conjoint Measurement for Price Optimization." *Sawtooth Software Conference Proceedings*, 157-162.

Share of Preference Model

$$ShP_{ik} = \frac{e^{b U_{ik}}}{\sum_{j} e^{b U_{ij}}}$$

ShP_{ik} = Share of preference for product k
i = respondent i
j = indication for all other products included
b = constant determined for each respondent in calibration
U_{ik} = total utility of product k for respondent i

Comment on Huisman

Jon Pinnell
IntelliQuest, Inc.

The measurement of price sensitivity is a very important topic within marketing research today. While buyers continue to be focused on price and promotions, manufacturers must be able to measure what value the market places on products, product features, and brand names. Only by accurately measuring the price/value and the value/cost relationships are manufacturers able to make informed decisions on product design and pricing.

Huisman's paper outlines five approaches to measuring price sensitivity with conjoint analysis. I would like to add a few points and, in an attempt to make the discussion complete, present some limitations frequently encountered measuring the price/demand relationship with conjoint analysis.

To begin with, it would probably be easier to think of the five approaches as fitting into three categories. They are:

- Including price as an attribute (as would easily be done using Sawtooth Software's ACA System)
- Including a price for each level of an attribute (as is done in Sawtooth's CVA System)
- Utilizing a dual conjoint design

I believe that these are presented in reverse rank order of their effectiveness in measuring price sensitivity.

IntelliQuest has used the dual design with good success in the past. This design works best when there are several product attributes included as well as price and brand. The product attributes can be measured using ACA (for example) in the first conjoint, and price and brand can be used in the second conjoint along with a composite or bundle of attributes from the first. This bundle is then used to bridge the two studies together. Unfortunately, the respondent task can become burdensome with many attributes. Possibly a bigger concern is the "bridge" between the two designs. The more complex the design, the more room there is for error in analysis, though this can be overcome with careful planning and thorough pre-tests.

The CVA design (price for each level) can also be burdensome, but with few attributes can provide very rich data. I agree with the author that it is preferred not to present each level's price, but rather a price of the configured product.

The ACA approach provides very good data for most product design and segmentation uses. For strategic pricing applications, however, including price as just another attribute will, as pointed out, likely produce biased results.

In general, I have been disappointed with the reference or base price approach to these problems. I believe that the differences in base price are a function of some observable phenomenon, such as

familiarity with the category, distribution source, or product quality. As such, they should be included in the conjoint design:

- As additional attributes,
- As a means to expose different respondents to different conjoint tasks, or
- As a special cut of the data to separately analyze subgroups.

I don't have anything in particular to quibble with about the content of the paper. I would, however, like to outline some limitations of conjoint analysis when used to formulate strategic pricing decisions. It is important to note that I am not criticizing conjoint analysis. Rather, my goal is to help make conjoint a better used and better understood research tool.

To that end, I would like to outline three limitations of conjoint analysis that are more likely to be issues in pricing research than in other types of conjoint research. They are:

- The number of levels effect
- The main effects only design (inadequate treatment of interactions)
- The measurement of share and not volume

All three topics have been discussed at this conference, and at other conferences.

Later in these proceedings, Dick Wittink discusses the number of levels effect more thoroughly than I can here. (The interested reader is referred to Wittink *et al.* (1989) and Wittink (1990).)

The second point is that conjoint generally involves a main effects only design. Basically, this assumes that the effect of an attribute towards a product's preference is a function of only its level, and is not at all affected by the levels of other attributes. This topic was covered well by Finkbeiner and Lim (1991) at the previous Sawtooth conference. Joel Huber, in commenting on the paper, outlines that interactions are generally not a problem. In pricing research, however, they can be — especially to the extent that brand name is evaluated also.

In my mind, the third limitation is the most critical. Conjoint is very powerful at measuring preference between configured products. It is not as powerful at measuring whether respondents are willing to choose any of the products, or choose to delay purchase. Because of this, we are measuring a price elasticity of share instead of a price elasticity of demand. The information can still be very powerful, but can also lead to inaccurate pricing decisions.

Several options exists to address this concern including:

- Creating a purchase likelihood score outside of conjoint
- Utilizing alternative techniques for pricing decisions — especially in conjunction with conjoint analysis

Discrete choice modeling (referred to by some as choice based conjoint), is an alternative methodological approach supported largely by Jordan Louviere, and can better address the volume and interaction concerns just raised. This technique was discussed by Johnson and Olberts (1991) at the ART Forum, by Louviere and Gaeth (1988) at the Sawtooth conference, and by Louviere (1988) in his conjoint publication from Sage.

In summary, the author's paper fairly presents three alternatives to measuring price sensitivity in conjoint. This is an important topic and I hope to see much work done on this in the future.

REFERENCES

Finkbeiner, Carl and Pilar Lim (1991). "Including Interactions in Conjoint Models." *Sawtooth Software Conference Proceedings.*

Johnson, Richard M. and Kathleen A. Olberts (1991). "Using Conjoint Analysis in Pricing Studies: Is One Price Variable Enough?" *Advanced Research Techniques Forum Proceedings,* American Marketing Association.

Louviere, Jordan (1988). *Analyzing Decision Making: Metric Conjoint Analysis*, Sage Publications series on Quantitative Applications in the Social Sciences, 07-067. Beverly Hills, CA: Sage Publications.

Louviere, Jordan and Gary Gaeth (1988). "A Comparison of Rating and Choice Responses in Conjoint Tasks." *Sawtooth Software Conference Proceedings.*

Wittink, Dick R. (1990). "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions." *Advanced Research Techniques Forum Proceedings,* American Marketing Association.

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.*

CONJOINT ANALYSIS IN JAPAN

Shota Hattori
Kozo Keikaku Engineering (Japan)

Three years ago, Kozo Keikaku Engineering first introduced Sawtooth Software's Adaptive Conjoint Analysis System (ACA) and Ci3 System to Japanese users. We have found that Japanese firms pay little attention to marketing research, especially conjoint analysis. Since some Japanese industrial products such as automobiles and consumer electronics have dominated the world market, many Americans and Europeans think that Japanese firms are well established in all functions of business activities. However, Japanese firms have very little ability to conduct marketing research. In this article, we discuss actual Japanese marketing research capability among academics, research firms, and business. Then, we introduce several of the few Japanese conjoint analysis studies. Finally, we present a comparison of ACA with computerized Full Profile methods and demonstrate how ACA is superior.

I. MARKETING RESEARCH IN JAPAN

A. Size of Marketing Research in Japan

According to the Japan Marketing Research Association, the total Japanese research business is about 60 billion yen ($480US million). This number is very small compared to the total Japanese advertising business, which is $30US billion.

B. Japanese Marketing Science Education

In Japan, there are only two business schools which are regarded as having the same quality of graduate education as organizations in the U.S. However, these Japanese schools focus mainly on "management strategy," rather than "marketing strategy."

The quality of faculty is also not up to U.S. standards. For example, although the University of Tokyo is regarded as the best university in Japan, there is only one marketing science professor in the faculty of Economics — among 60 professors. The Marketing Science Association in Japan has only 150 members. Fewer than ten of them have submitted articles to international marketing-related journals.

C. Marketing Research Practices in Japanese Firms

Japanese firms have been copying products originally developed in U.S. and European countries. With the high quality of production and with well-educated employees, these lower-priced, high-quality products prevail in the world marketplace. Since Japanese companies have been copying existing products, it has not been thought necessary to conduct detailed marketing research.

Japanese Public Relations companies such as Dentsu and Hakuhodo do use marketing research to validate their public relations' strategies. These companies are oriented toward product image research. So, their research uses factor analysis and principal component analysis. Conjoint analysis is a rare tool in the Japanese research industry.

Japanese engineers in product development and planning divisions are not familiar with marketing research methods. They are very good at improving and adjusting technologies to products within a limited time, but they do not conduct market evaluations. Up to now, they have paid no attention to the importance of marketing research activities, especially to product specifications among segments. In lieu of doing market research, their strategy has been to include every feature at the lowest price.

II. SOME RECENT DEVELOPMENTS IN JAPANESE CONJOINT ANALYSIS.

There are several trials based on computer interviewing using conjoint analysis in Japan. We will explain several of them.

A. Computer interviewing for conjoint analysis using minimum expected entropy principle.

A new method of data collection for conjoint analysis using a personalized computer interview is proposed. At each stage of questions, the computer asks the respondent to make a choice from a pair of profiles, where the total set of attributes is shown, as in the full profile method. A respondent is assumed to belong to a set of representative taste segments. The probabilities of membership are updated at each stage of questions, based on a Bayesian rule. The pair to be presented is selected to minimize the expected entropy of the membership probabilities after updating. The motivation of the proposed method is to collect as much information as possible from a limited number of questions, the underlying notion being that what questions are asked make a difference in terms of the accuracy of parameter estimation.

To validate the advantages of the proposed method, Monte Carlo simulations are run to compare the method with randomly selected pairs of profiles. In terms of the mean square error between the normalized estimates and parameters, it is shown that the proposed method, for a given number of questions, gives estimates as accurate as those from the random method with twice as many questions.

This study is an academic work and the approach has not been applied to an actual marketing research study yet. One problem is that each time the PC presents the pair-questions, it has to perform calculations, and it keeps the respondent waiting. But, in terms of algorithm, this study suggests a new direction in conjoint analysis.

B. Unique applications of conjoint analysis in Japan

1) *Using Sawtooth Software's Conjoint Value Analysis System (CVA) on a Macintosh*

An imported beer company wanted to conduct a pricing study, to investigate the Japanese market. They used CVA to focus on price, brand, and bottle type. We created graphical representations of the pairs created by CVA and displayed them on a Macintosh using HyperCard. Each respondent viewed a series of pairs and utilities were calculated individually. Respondents reported that they enjoyed participating in the survey, particularly because of the graphic representations of bottles and cans of each brand.

2) *Using ACA on a Macintosh*

We used a Macintosh in a study about computers. The four attributes were monitor type, price, CPU capability, and size. Each attribute level was represented graphically and pairs of attributes were combined to create an image of the set of attributes. Respondents used a mouse to proceed from question to question.

3) *Conjoint Analysis on the Train*

A regional Japanese railroad company wanted to compare its super-express train with highway bus service. Laptops, which are common in Japan, were used to interview train passengers. The study had seven attributes, including fare, on-time service, numbers of trains and buses per hour, and minutes to destination. The passengers, who otherwise sleep or read on the train, were delighted to participate.

III. A COMPARATIVE STUDY OF DATA GATHERING PROCEDURES IN CONJOINT MEASUREMENT

Data for this study were collected jointly with Professor Kosuke Ogawa of Housei University and Mr. Masahiko Yamanaka of Ajinomoto.

A. Study purpose

Conjoint analysis is used commercially to estimate acceptance of product concepts and determine appropriate price levels. In conjoint analysis, we collect data for profiles that are combinations of attributes to determine the "preference" weight for the attributes of products.

The most popular method of data collection for conjoint analysis in Japan is to rank profiles in order of preference. The profiles are described using 16-32 pictures or cards. This method is called the "Full Profile" (FP) method and is used in many studies because preparation is easy and presentation of profiles resembles the real situation of product selection.

Conjoint analysis has another data collection method, the "tradeoff" method, which became more popular with the introduction of computer interviewing. In the tradeoff method, pairs of profiles are shown one after another on the computer screen, and the respondents are asked which profile is preferred, and how much one is preferred to the other. The pairs are created automatically by the computer, according to a selection logic. The computer keeps track of the previous answers, so that the task becomes increasing difficult for the respondent. Utilities of the product's attributes are calculated.

The tradeoff method has an important advantage: respondents can answer questions easily because of the computer-interactive interview. However, the method also has a weak point, which might cause a bias in the estimated utilities, caused by the particular logic for the calculation of utilities. It was not possible to determine which method was better when we were comparing the difference between estimating one pair at a time and many profiles at the same time. So, we tried a comparative experiment on these two methods of data collection. We collected two types of data

from the same respondents, using full profile and tradeoff methods for the two product groups. We compared:

1) To what degree are the partworths estimated from the two methods consistent (the "reliability" of partworth)

2) When we predict consumers' purchases based on the estimated partworths, which method better predicts the real selection of products (the "validity" of estimation)

3) What effect the choice of consumers and product features have on the "reliability" and "validity" of the partworth estimations.

We used CVA (Conjoint Value Analysis) software for the full profile method and ACA (Adaptive Conjoint Analysis) software for the tradeoff method. CVA involves a full profile approach. We'll refer to the full profile method as FP and refer to the tradeoff method as ACA.

B. Summary of Survey Plan

1) *Selection of products, attributes, and levels.*

CHOCOLATES and SOFT DRINKS were selected as the products for this experiment, because they were well known to the respondents — who were university students and young employees — and because the budget for the experiment was limited. Both males and females had abundant knowledge about SOFT DRINKS. However, for CHOCOLATES, many male respondents weren't regular consumers and weren't very familiar with the brands.

For SOFT DRINKS, the attributes were price, brand, taste, and volume. For CHOCOLATES, the attributes were price, brand, taste, and style.

Each attribute's levels are described in Table 1. The cards used in the FP method were made by combining the attributes and levels in accordance with Adelman's Latin Square. There were 16 cards for CHOCOLATES and 25 for SOFT DRINKS.

Table 1

Attributes and Levels

SOFT DRINKS

Attribute 1 (Price)	Attribute 2 (Brand)	Attribute 3 (Taste)	Attribute 4 (Volume)
80 yen	Coca-Cola	Soda	125 ml
100 yen	Otsuka	Functional	250 ml
120 yen	Suntory	Tea-type	350 ml
	Kirin	Coffee	

CHOCOLATES

Attribute 1 (Price)	Attribute 2 (Brand)	Attribute 3 (Taste)	Attribute 4 (Style)
100 yen	Lotte	Milk	Bar-type
150 yen	Meiji	White	Bit-type
200 yen	Morinaga	Bitter	
	m&m's	Nuts	

2) *Respondent profiles and survey schedules*

The respondents were 115 students who took the lecture "Management Science" in the Management Department of Housei University, and 106 young employees of Kozo Keikaku Engineering, Inc. The final sample size was 206 (133 male, 73 female).

The interviewing was done in a two week period, June 3 -17, 1991 at the Tokyo and Kumamoto offices of Kozo Keikaku Engineering.

3) *Features of Computer-Interviewing*

We developed Macintosh software especially for this survey. For ranking cards in the full profile method, respondents used a mouse to drag and rank cards. At the end of the interview, respondents were requested to pick one real drink and one chocolate, to help measure how well our results predicted the real purchase behavior.

The experimental design included four combinations, because two types of data for each category were collected from each respondent. To avoid order effects, the 206 respondents were divided into four groups after the introductory questions were asked:

Design 1 (n = 50)	Design 2 (n = 51)	Design 3 (n = 54)	Design 4 (n = 51)
FP CHOCOLATE	FP SOFT DRINK	ACA CHOCOLATE	ACA SOFT DRINK
FP SOFT DRINK	FP CHOCOLATE	ACA SOFT DRINK	ACA CHOCOLATE
ACA CHOCOLATE	ACA SOFT DRINK	FP CHOCOLATE	FP SOFT DRINK
ACA SOFT DRINK	ACA CHOCOLATE	FP SOFT DRINK	FP CHOCOLATE

C. SURVEY RESULTS

1. Interviewing time

On average, the total interview took 40 minutes. One hundred seventy three respondents took 35-55 minutes. The ACA method was faster than the FP method. Excluding the demographic questions, the average times were:

FP SOFT DRINK	12.5 minutes
FP CHOCOLATES	9.0
ACA SOFT DRINK	8.0
ACA CHOCOLATE	7.0

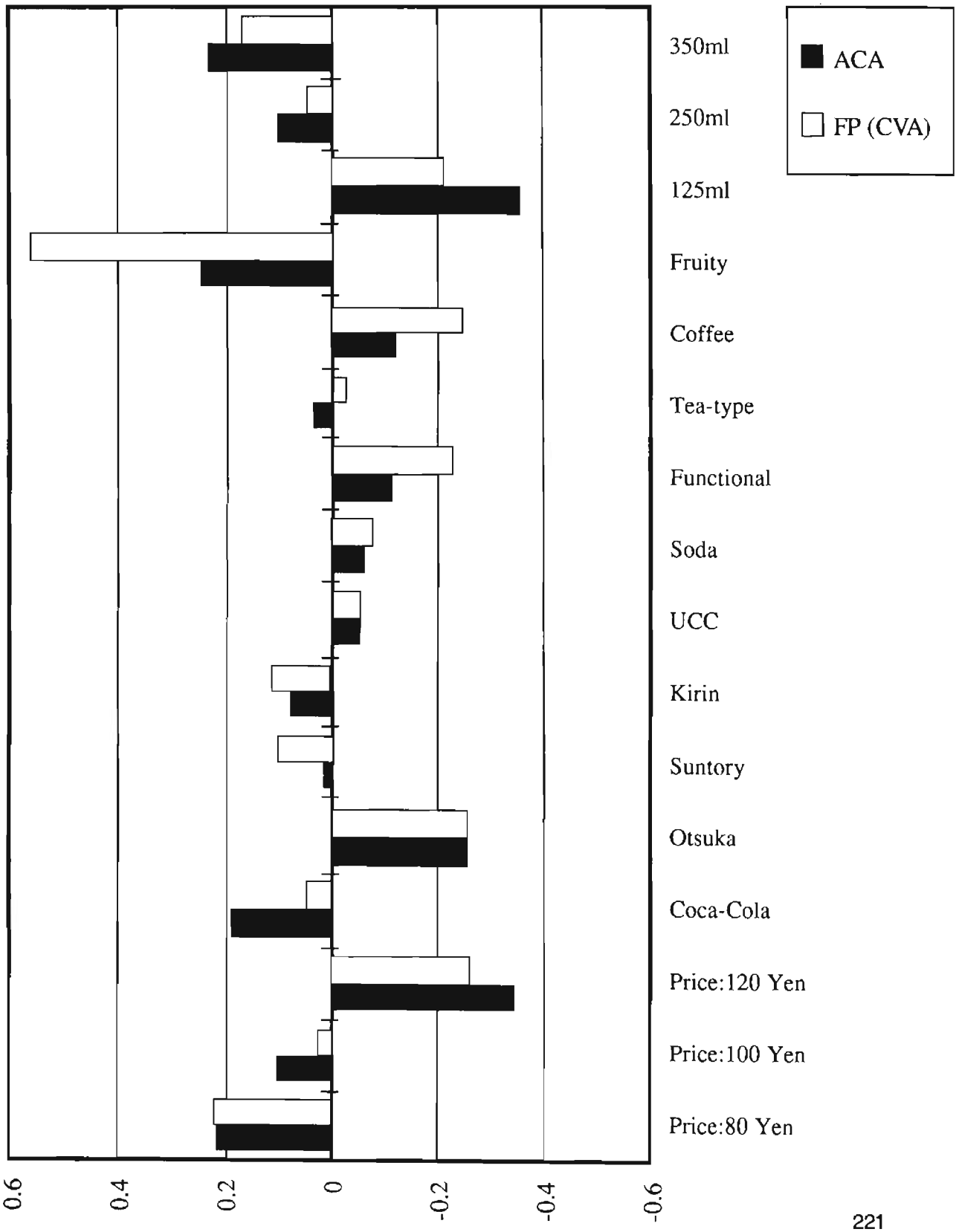
2. Partworth correlation between methods

Partworths calculated by ACA and FP were similar. For both SOFT DRINKS and CHOCOLATES the average correlation of individual partworths was 0.53. (See Figures 1 and 2 on the following pages.)

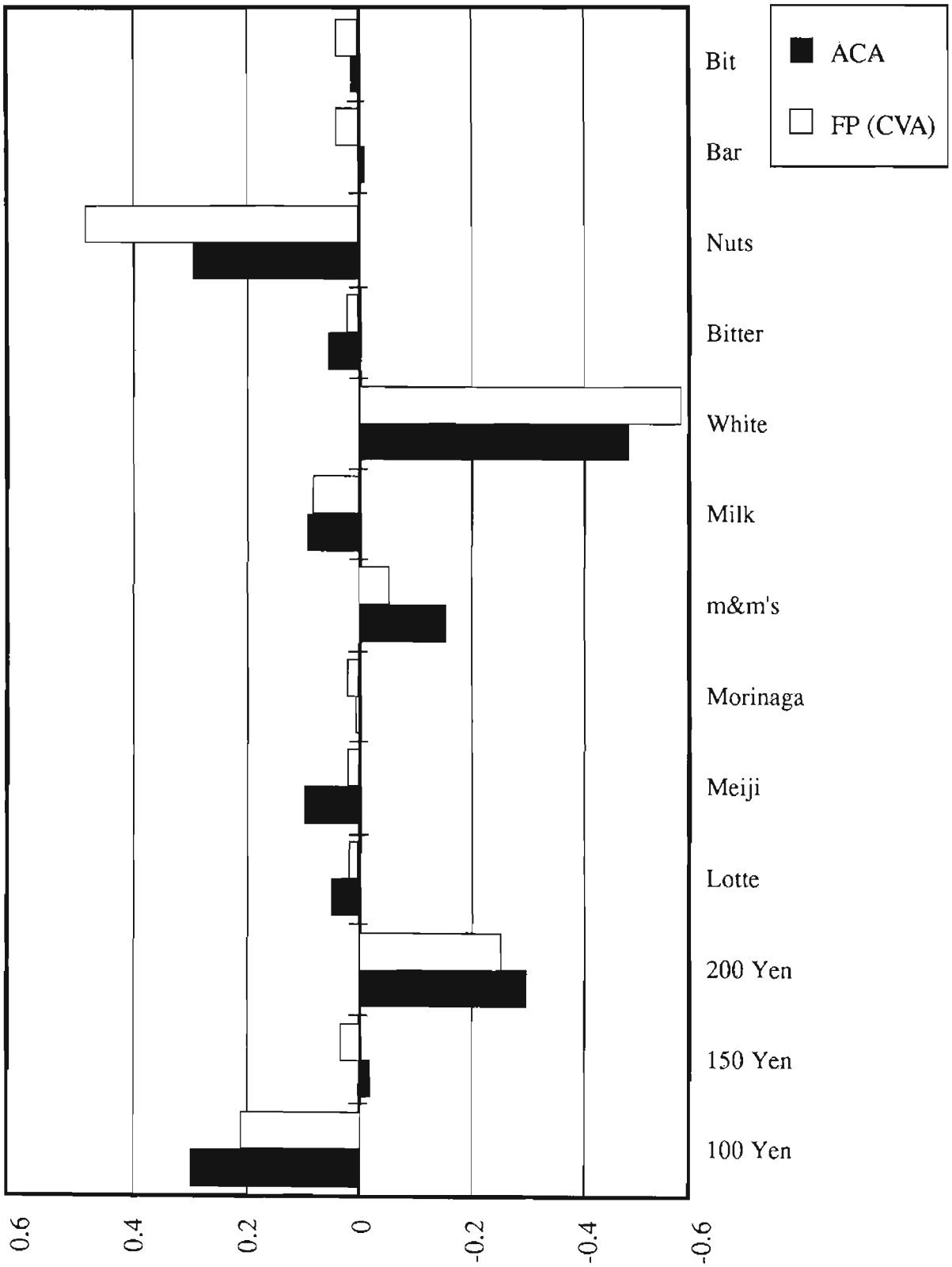
3. Conclusions

We report the following conclusions: First, in estimating partworths of each level of attributes, ACA and FP are well correlated. Second, in increasing the number of attributes and levels, ACA's estimating of partworths is better than that of FP's. Finally, the ACA method takes less time than the FP method.

Average Partworth (soft drink) N-206



Average Partworth (chocolate) N-206



Comment on Hattori

Ray Poynter
Sandpiper Computer Centre (England)

Mr. Hattori has made a number of useful and valid observations. He has given, however, the conventional view of Japan and there is, I believe, an alternative perspective. This alternative perspective applies to both the size of the market research industry and the nature of the Japanese marketplace.

SIZE OF THE MARKET RESEARCH SECTOR

Mr. Hattori has given the published figures, plus a sensible adjustment for under-reporting. There are however other factors that could be taken into account. The first is the large amount of media research undertaken by Dentsu and Hakahodo. In many markets all of this would be included in the market research sector; in Japan it is usually not.

The second feature of the industry is the massive scale of visits from senior people in the supplying company to clients, shops, factories, and outlets. This provides a massive body of market information — which in part is a substitute for market research expenditure.

Another factor of the Japanese market is the number of antenna stores in fashionable areas such as Shibuya. These stores provide leading edge ideas, prototypes and new products. Through these stores manufacturers get a good idea of what is wanted.

Finally there are a number of anthropological institutes, often associated with the advertising agencies, who research basic trends, including market related ones.

The conclusion I would draw is that the level of consumer research activity in Japan is considerably higher than the market research figures would suggest.

The next question I would like to address is whether Japan needs the same type of research as that practised in, say, the USA or the UK.

Mr. Hattori describes Japan as essentially a "copying" economy. Whilst this may have been true in the 50s and to some extent the 60s, it is far from true now. What is true is that Japan's companies often prefer a small innovation to a large scale revolution — although the Walkman was an example of a large scale step. The small step innovation is very often design-led rather than research-led. A product is improved because it *can* be, not necessarily because there is consumer demand for that change.

Another feature of Japan is the size and homogeneity of the consumer world. Whilst Japan is becoming more segmented it still remains far more homogenous than Europe or the USA. This leads to simpler research designs and simpler sampling, as indicated by Mr. Hattori.

The vertical integration of the market leads to dramatic levels of competition. A good example is the competition that has taken place in the fax market where the lead players have had a cutthroat innovation-led battle for dramatic supremacy. A side affect of this domestic battle has been Japanese world domination of the fax market.

Another feature of Japan is the number of retail outlets (more shops than in the USA with half the population) plus 5 million vending machines. This creates a dynamic atmosphere where change can be, and usually is, very rapid.

Manufacturers respond by being willing to innovate very fast. In the motorbike war a few years ago, the lead manufacturers introduced some 50 new lines each within one year. In the soft drinks arena some 1000 new products are launched each year. Often research is out of the question because of the delays it would cause.

In conclusion, I would support the vast majority of Mr. Hattori's observations, but with a note of caution. In some cases Japan is, in research terms, less developed because it simply has not yet reached the appropriate level. In these cases, there is a market advantage waiting for someone. In other cases, the technique is less developed because it is not suitable for direct transplant into the Japanese marketplace.

A COMPARISON OF TELEPHONE CONJOINT ANALYSIS WITH FULL PROFILE CONJOINT ANALYSES AND ADAPTIVE CONJOINT ANALYSIS

Keith Chrzan
Walker: Research & Analysis
Douglas B. Grisaffe, Ph.D.
Walker: CSM

INTRODUCTION

Telephone data collection allows fast turnaround, cost efficiency, flexibility to adjust questions and, perhaps, greater response rates and representativeness than do mail or in-person interviewing (Lockhart, in press). A conjoint methodology that can capitalize on these advantages obviously will be a valuable innovation for researchers. Terry Elrod (1991) has invented a very efficient methodology for pairwise choice-based Telephone Conjoint Analysis (TCA).

In Elrod's TCA, choice alternatives within pairs differ on only two attributes, so respondents can handle the task even in a telephone survey. TCA's experimental design ensures near orthogonality of attribute differences, but suffers from a serious flaw. Except under exceptional and lucky conditions (a significant non-linear transformation of a quantitative variable), the design matrix is singular: any one attribute difference is a perfect linear combination of differences in other attributes.

In conversation with Elrod, we became convinced that his original design strategy could be modified to prevent the singularity problem while retaining the advantages that would make TCA a viable strategy for performing a conjoint task by telephone. Our modification of Elrod's design involves the following:

1. choices between profiles differ on only three attributes; one devises these choices by a straightforward rotational design (see Appendix A);
2. each respondent evaluates each attribute equally often;
3. across respondents, attributes appear as often in the first position as in the second position of choice pairings, thus controlling for order effects;
4. one constructs the design in "blocks" so that any one respondent need only receive a portion of the experimental design. In fact, the task can be split such that one need only ask of any respondent as many choice questions as there are attributes under study. This gives TCA the potential to handle large numbers of attributes.
5. finally, one can use widely available binomial logit to decompose choice responses into aggregate or segment level utilities, and the multinomial logit choice rule to simulate shares from the utilities.

Beyond the attractive design characteristics, the task is brief and easy to administer in a telephone interview. Also, the response scale is choice, a more natural reaction to marketing stimuli than rating or ranking. All things considered, TCA appears promising.

Of the many issues of interest, we concentrate on TCA's experimental design and modeling capabilities as they affect predictive validity. One may view TCA as a design strategy for choice-based full-profile conjoint (CFP) analysis (Louviere and Woodworth, 1983; Louviere, 1988) that surrenders some statistical efficiency and design flexibility to gain the potential for telephone administration. In this case we want to compare the predictive validity of TCA to that of CFP to measure any loss in predictive validity caused by the loss in efficiency. From a practical standpoint, the loss may be small relative to the potential advantages.

In addition to the TCA/CFP comparison we compare TCA to other conjoint approaches. We will compare TCA's predictive validity to that of ratings-based full profile (RFP) conjoint analysis, arguably the current "standard" approach in conjoint analysis. Finally, Adaptive Conjoint Analysis (ACA — ACA System by Sawtooth Software) may be TCA's only rival for performing decompositional conjoint analysis entirely over the telephone, so we compare the predictive validity of TCA and ACA.

Note the limited scope of our analysis: even if TCA yields lower predictive validity than the full-profile conjoint approaches, we can at most quantify what the researcher must trade off to acquire the potential benefits of TCA.

METHODS

Survey Instruments

Separate instruments serve the TCA/CFP/RFP and TCA/CFP/ACA comparisons. For practical purposes we administered the TCA/CFP/RFP instrument *via* a paper-and-pencil survey. Telephone data collection is not a *necessary* feature of the TCA design. Administration by paper and pencil allows greater experimental control and reduces chances for method variance contamination. For similar reasons we collected the TCA/CFP/ACA instrument *via* self-administered PC interviewing with ACA and Ci2 (Ci2 System by Sawtooth Software) software.

We want to evaluate the importance of five incremental services that supermarkets in Indianapolis often provide. Both instruments instruct respondents to "imagine that you need to shop at a supermarket today." Both include the following five attributes of local supermarkets (two levels each):

double coupons (absent/present)
warehouse prices (absent/present)
open 24 hours (absent/present)
video tape rental (absent/present)
pharmacy (absent/present)

A TCA task with five attributes requires 10 questions which may or may not be split into two blocks.

We wanted about 100 responses per choice question and thought we would have trouble getting enough respondents if we increased the number of attributes beyond five. This decision also allowed us to keep the paper-and-pencil questionnaire to four pages and, potentially, to increase our response rate. Similar reasoning led us to use two level attributes: polytomous (multileveled) attributes multiply the number of blocks required by TCA's design. The selection of this relatively

small number of attributes is a fair comparison but shows neither TCA nor ACA at its strongest relative to the full-profile approaches: TCA and ACA can plausibly handle 20-30 attributes, while CFP and RFP cannot (Green and Srinivasan, 1990).

TCA/CFP/RFP

The first instrument supports comparisons of TCA, CFP and RFP. It contains the 10 choice questions required for a complete TCA design, the eight full-profile ratings questions for a standard main effects RFP task (Addelman, 1962), the eight full-profile choice sets necessary for an optimally efficient CFP design (Louviere and Woodworth) and a validation set of six multiple choice questions different in format from all of the above. Six versions of the instrument allowed all possible rotations of the three calibration tasks, thus correcting for a known task order effect (Huber, *et al.*, 1991). The six validation choices yield 20 choice shares held out from all of the above models with which to test predictive validity. The validation questions appear at the beginning of all versions. All questions use only verbal stimuli. Examples of the question types appear in Appendix B.

TCA/CFP/ACA

The second instrument contains the 10 choice questions for a TCA task, the eight full-profile choice sets for the CFP task and an ACA interview whose Interview Control Parameters appear in Appendix C. Again rotation controls for order effects in the calibration tasks and the six multiple choice validation questions begin each rotation. Examples of ACA question types appear in Appendix D.

Respondents

A total of 400 householders from two Indianapolis neighborhoods received the TCA/CFP/RFP instrument. Of these, 183 or 46% returned surveys, of which 176 (44%) were free of missing data. These 176 we selected for analysis. A convenience sample of 54 PC users completed the TCA/ CFP/ACA task. Completes break out as follows by version:

Table I

Task Order

Version	Administration	Task Order	Respondents
1	Paper and Pencil	CFP/RFP/TCA	29
2	Paper and Pencil	CFP/TCA/RFP	27
3	Paper and Pencil	RFP/CFP/TCA	29
4	Paper and Pencil	RFP/TCA/CFP	28
5	Paper and Pencil	TCA/CFP/RFP	31
6	Paper and Pencil	TCA/RFP/CFP	32
7	Computer	CFP/ACA/TCA	7
8	Computer	CFP/TCA/ACA	8
9	Computer	ACA/CFP/TCA	8
10	Computer	ACA/TCA/CFP	9
11	Computer	TCA/CFP/ACA	12
12	Computer	TCA/ACA/CFP	10

Analysis

We compare the various conjoint models with respect to predictive validity. Within each instrument we compare the predictive validity of the three component conjoint models relative to the 20 validation choice shares by way of a repeated measures ANOVA. In addition, we perform paired t tests of differences in correlations to compare the validity of alternate models two at a time using expanded sets of holdout choice shares.

Repeated Measures ANOVA

The six validation choices in each instrument provide 20 choice frequencies held out from estimation of all conjoint models. Each model in a given instrument produces a set of corresponding share predictions. Subtracting the vector of actual choice frequencies from each of the three vectors of predicted shares yields three vectors of raw errors. Squaring these errors produces three vectors of squared errors. Means of these squared errors are measures of the relative predictive validity of the respective conjoint models.

The 3 x 20 matrix of squared error terms may be subjected to repeated measures ANOVA. The critical F at p=.05 for the null hypothesis of no difference in mean squared error depends on the homogeneity of variance correction of the initial 2 and 38 degrees of freedom. If an ANOVA produces an F greater than the critical F, then we conclude that a significant difference in mean squared error exists among methods and we utilize a Newman-Keuls *post hoc* test to identify which group means differ from one another.

Pairwise t tests for Differences in Dependent Correlations

A stronger pairwise testing strategy goes as follows: in addition to the six validation questions, the eight CFP choices are holdouts relative to the TCA versus RFP and TCA versus ACA comparisons. Similarly, the RFP questions are holdouts relative to the TCA versus CFP comparison. Subjecting differences in correlations between each method's predicted and actual holdout choice frequencies to t tests will test the hypotheses that the correlation between actual and simulated choice shares for TCA is lower than for CFP, RFP and ACA. In each comparison the same respondents provide data for TCA and the other conjoint technique, so the t tests are for *dependent* correlations (Cohen and Cohen 1983). At p=.05 the critical t statistic depends upon the number of holdout choice shares for a given comparison. If a critical t is exceeded, we will conclude that the particular full-profile model or ACA has predictive validity significantly different from TCA. The increased number of observations add power to the test and the full-profile holdouts make for a more realistic test of validity. The pairwise comparisons, however, increase the chances for Type I error, so we use a Bonferroni correction procedure. (In addition, the 10 TCA questions are holdouts relative to CFP, RFP and ACA, so they can be combined with the validation questions in pairwise comparisons of ACA versus CFP and CFP versus RFP. These comparisons appear in Appendix E).

The TCA model was generated with the SPSS/PC+ logistic regression procedure for binomial logit. Each line in the TCA design from Appendix A appears twice in the SPSS program, once with a 0 code for respondents choosing the first choice alternative and once with a 1 code for respondents choosing the second choice alternative. Numbers of respondents choosing each alternative serve as weights for the binomial logit. We estimated the CFP model with the Systat LOGIT program for maximum likelihood multinomial logit and the individual level RFP models with the SPSS/PC+ conjoint procedure for ordinary least squares conjoint analysis. ACA produced the individual level ACA models. These programs typify software available to researchers. The multinomial logit choice rule simulates shares for TCA and CFP while the first choice rule governs share simulations for RFP and ACA (Green and Srinivasan, footnote 8).

RESULTS

Repeated Measures ANOVA

TCA/CFP/RFP

Table II shows the utilities derived from the CFP, RFP and TCA calibration models. The three conjoint models failed to produce significant coefficients for the Video Tape Rental and Pharmacy attributes.

Table II

Comparison of Utilities

Attribute	CFP	RFP	TCA
Warehouse Prices	.942	1.477	2.650
Double Coupons	.602	1.191	2.191
Video Tape Rental	-	-	-
Pharmacy	-	-	-
24 Hour Service	.271	.577	.773

Validation choice frequencies and their simulations by CFP, RFP and TCA appear in Table III.

Table III

Predicted and Actual Choice Frequencies for Validation Questions

Question	Choice	CFP	RFP	TCA	Validation
1	A	6.4	1.1	2.5	4.5
1	B	72.3	76.4	75.5	64.8
1	C	21.3	22.4	22.0	30.7
2	A	7.0	7.1	1.7	13.1
2	B	4.1	1.7	.8	1.7
2	C	89.0	91.2	97.6	85.2
3	A	56.6	56.3	56.0	65.3
3	B	14.8	11.1	8.6	5.1
3	C	28.7	32.7	35.4	29.5
4	A	49.4	38.9	41.0	44.9
4	B	43.1	59.7	56.1	50.0
4	C	7.5	1.4	2.9	5.1
5	A	56.6	56.3	56.0	61.4
5	B	14.8	11.1	8.6	9.7
5	C	28.7	32.7	35.4	29.0
6	A	7.3	.6	3.7	0.0
6	B	7.3	.6	3.7	3.4
6	C	12.6	10.2	7.9	6.3
6	D	48.3	56.1	51.9	64.2
6	E	24.5	32.5	32.8	26.1

The square roots of the mean squared errors are 6.5% for CFP, 6.0% for RFP and 6.9% for TCA. The computed F statistic of .52 is far below the critical $F_{1,8,34\,2}$ of 3.45, however, so no significant differences in mean squared error exist among CFP, RFP and TCA. The means of *absolute* errors were 5.5% for CFP, 5.2% for RFP and 5.7% for TCA, and they did not differ significantly.

Validation share and simulations are correlated as follows:

Table IV

Correlations of Validation Shares and Share Predictions

	CFP	RFP	TCA
RFP	.98		
TCA	.98	.99	
Validation	.97	.98	.97

TCA/CFP/ACA

Table V shows the attribute utilities for the CFP, ACA and TCA calibration models.

Table V

Comparison of Utilities

Attribute	CFP	ACA	TCA
Warehouse Prices	.736	1.27	2.044
Double Coupons	.446	.98	1.511
Video Tape Rental	-	.35	-
Pharmacy	-	.40	-
24 Hour Service	.360	.96	1.354

Table VI displays validation choice frequencies and their simulations by CFP, ACA and TCA.

Table VI

Predicted and Actual Choice Frequencies for Validation Questions

Question	Choice	CFP	ACA	TCA	Validation
1	A	8.1	1.9	2.8	5.6
1	B	72.2	81.5	84.4	70.4
1	C	19.7	16.7	12.8	24.1
2	A	15.0	24.1	9.7	27.8
2	B	7.3	0.0	2.5	3.7
2	C	77.7	75.9	87.8	68.5
3	A	49.2	40.7	47.9	51.9
3	B	23.2	24.1	24.0	20.4
3	C	27.6	35.2	28.1	27.8
4	A	42.1	42.6	29.4	50.0
4	B	48.4	55.6	66.8	48.1
4	C	9.5	1.9	3.8	1.9
5	A	49.2	57.4	47.9	50.0
5	B	23.2	16.7	24.0	20.4
5	C	27.6	25.9	28.1	29.6
6	A	9.1	0.0	5.5	3.7
6	B	9.1	3.7	5.5	1.9
6	C	19.0	20.4	21.4	24.1
6	D	40.3	51.9	42.6	48.1
6	E	22.5	24.1	25.0	22.2

The square roots of the mean squared error are 5.5% for CFP, 6.0% for ACA and 9.8% for TCA. The computed F statistic of 3.91 is slightly lower than the critical $F_{1,3,24,7}$ of 3.98 at p=.05. Again, differences significant at p=.05 do not exist among mean squared errors for the three conjoint models' predictions. Mean absolute errors are 4.3 for CFP, 5.2 for ACA and 6.2 for TCA. These are not significantly different.

Share simulations are correlated as follows:

Table VII

Correlations of Validation Shares and Share Predictions

	CFP	ACA	TCA
ACA	.97		
TCA	.98	.96	
Validation	.97	.97	.93

Pairwise t tests for Differences in Dependent Correlations

Cohen and Cohen (1983) describe a t test for differences in dependent correlations. All of the following are within-instrument comparisons based on autologous data, so the dependent test is appropriate. Degrees of freedom for each test equals the number of paired observations minus three (n-3). We employ a one-tailed test to examine the hypothesis that TCA's correlation with holdout choice frequencies is lower than that of CFP (as a relatively inefficient version of CFP, we expect it will fare less well). Not having any expectations about TCA's performance relative to RFP and ACA, we employ two-tailed tests for these comparisons.

TCA versus CFP

Relative to the TCA versus CFP comparison, the six validation choices and the 8 RFP questions are held out. The eight RFP profiles can be paired in 28 possible ways. We make "pseudo choices" by assuming that a given respondent chooses the more highly rated profile in a possible pairing (Johnson, 1989). Tied ratings are distributed 50% to each profile. There is some "noise" in this creation of pseudo choices, but the handicap is common to both CFP and TCA, so it adds no identifiable bias in favor of one method or the other. This adds 28 choice shares to the 20 validation choice shares from the validation questions (not 56, because the choices are binary and the second choice share from a pair is just a complement of the first). Thus the t test for difference in dependent correlations has 45 degrees of freedom and therefore a critical t of 1.68 (2.42 after Bonferroni correction for multiple comparisons). The correlations that produce the t statistic are:

$r_{TCA,\ Holdout} = .9672$
$r_{CFP,Holdout} = .9738$
$r_{TCA,CFP} = .9796.$

The computed t statistic of 1.01 does not exceed the critical t. Correlations of TCA and CFP simulations with holdout choices and holdout pseudo choices are not significantly different.

TCA versus RFP

Similarly, the eight CFP choices add eight holdout choice shares to the 20 from the validation questions for a total of 28 observations, or 25 degrees of freedom for the t test, yielding critical ts of ±2.06 (±2.79 corrected). The correlations relevant to the t statistic are:

$r_{TCA,\ Holdout} = .9570$
$r_{RFP,Holdout} = .9776$
$r_{TCA,RFP} = .9794.$

The computed t of 2.41 is not significant after the Bonferroni correction, so we conclude that correlations of TCA and RFP simulations with actual holdout shares do not differ significantly.

TCA versus ACA

Again the eight CFP questions are held out relative to TCA and ACA, so the critical t for the test of difference in dependent correlations is ±2.06 (±2.79 after correction). The correlations between simulated and actual holdout shares are:

$r_{TCA,\ Holdout} = .9484$
$r_{ACA,Holdout} = .9594$
$r_{TCA,ACA} = .9743.$

These produce a non-significant t statistic of .88. TCA and ACA do not differ significantly in predictive validity.

DISCUSSION

Study Summary

We set out to test Elrod's proposed telephone conjoint methodology. The method interested us because of the attractive features it offers for conjoint studies: telephone data collection, potentially for large numbers of attributes, with choice decisions rather than ratings or rankings. To explore the method, we focused on its predictive validity compared to other, standard, approaches to conjoint modeling. We collected data that allowed us to compare the predictive validity of Elrod's method, ratings-based full profile conjoint, choice-based full profile conjoint, and Sawtooth Software's Adaptive Conjoint Analysis.

We used two basic approaches to compare predictive validities. First, we compared the mean squared error of prediction across the various conjoint methods. This involved generating squared differences between predicted and actual choice frequencies, and then testing for significant differences among the means of that quantity across the various conjoint approaches. Secondly, we compared the conjoint tasks on the degree of association between predicted and holdout choice frequencies. Within a given conjoint method, we generated the correlation between predicted and actual choice frequencies. We then tested for differences in these correlations between different conjoint methods.

In our study, TCA did not work significantly less well than the standard methods. Thus our findings provide initial evidence supporting the method's soundness. We believe one can feel good about "giving up" a little in the orthogonality of the design matrix and the completeness of the choice profiles, in return for "getting" good conjoint information without significant decrement in predictive validity (however, we will point out some qualifications to consider before making any generalizations).

With that general summary in hand, the rest of this discussion section will occur in four parts. First we offer some commentary about our choice of data collection method. After discussing that, we address generalizability issues. Next, we discuss some contributions we feel this study has made. Finally, we offer a few ideas for future research.

Choice of Data Collection Method

With the exception of ACA, it would not have been possible to do a telephone-only data collection approach to test TCA against the other methods. It would have been possible to take a phone-mail-phone approach, but not a telephone-only approach. Because full-profile methods are generally impractical for such data collection, we needed a different data collection. Further, we did not want to introduce method bias by collecting TCA data by telephone, but RFP and CFP data on paper. Thus we collected our TCA data on RFP, CFP, and ACA terms, using paper-and-pencil and computer interview methodologies. (We recontacted a few respondents, by telephone, approximately a month and a half after initial paper-and-pencil data collection. We repeated the TCA questions over the phone. Of 80 opportunities for agreement, respondents repeated their previous answers 69 times (86% agreement). This suggests the feasibility of telephone administration of TCA.)

We reasoned that if TCA did well in traditional conjoint circumstances when compared to standard approaches, one would feel more confident in using it where RFP and CFP cannot go, namely telephone-only studies. Further, given the inordinant complexity of a full-profile telephone-only data collection, we might expect an inherent bias in favor of TCA in a telephone-only study. Our choice of data collection method also posed no conflicts or inconsistencies for our research question. We were not generating predictive validity data to explore TCA as an alternative to RFP or CFP, but rather as a viable option when RFP and CFP were not available (that is, telephone-only conjoint studies).

Issues Concerning Generalizability

First, we note that our findings come from a study of supermarket features. We have no reason to believe that findings would have differed had we measured responses to pizza restaurant features, long distance telephone service features, or any other area of interest. However, we note the possibility of subject matter effects that might have led to different findings in different "contexts."

Second, we note that two of our five attributes dominated as the most important features. It is possible that findings might have differed given a study where importance was more evenly distributed among the attributes of interest. In other words, these two attributes might be so dominant that almost any approach would give the "right" answers.

Third, our study had five attributes, each with two levels (present or absent). We do not know if our results would have been different had we measured 8 or 10 attributes rather than 5. However, if there were differences, we might expect TCA to be more viable under such conditions. Recall that the ability to handle large numbers of attributes is one of the stated benefits of the TCA methodology. The practicality of the full profile methods is suspect when the number of attributes gets very large (Green and Srinivasan).

Fourth, we recognize that our samples were ones of convenience. Again, we have no reason to suspect that findings would differ for other samples, but we note it as a possibility. Our interest was methodological rather than substantive, making the representativeness of the sample a less salient issue. Further, it is even less of a concern because each conjoint approach would presumably be equally affected by whatever inherent sample biases existed. Thus at a minimum, a sample-specific comparison of the conjoint methods would still be a "fair" test.

Fifth, our personal computer sample was small (n = 54). There are two possible problems with this. One is the general fact that small samples are a concern anytime statistical estimation is involved. Beyond that issue, a second concern is that the conjoint approaches we compared use different statistical estimation procedures. If the different estimation methods are *differentially* affected by small samples, then there is the possibility that true comparability is lost. For example, the maximum likelihood estimation in TCA's logit may produce utilities that are more affected by the small sample than ACA's utilities (that is, more error-laden because maximum likelihood estimators are only asymptotically unbiased).

Contributions of the Study

We feel at least a few meaningful contributions have been made by our study. First, a serious design issue in Elrod's original formulation has been identified and resolved. The presence of the problem was an important discovery in the earliest work we did on this study. In fact, working on the design problem resulted in the second contribution of this study. A new general approach was created for generating the appropriate design matrix for TCA studies (see Appendix A).

Third, to our knowledge, we have done something novel in using repeated-measures ANOVA to compare the predictive validity of the methods. Beyond our application for this study, we feel the approach is not limited to the simple use of a one-way design. In fact, one could imagine complicated factorial experiments where the "treatments" involve different conjoint methods, subject matter contexts, numbers of attributes and levels, different methods of administration, different sample characteristics and so forth. Such factorial designs would allow one to control for many potential biasing influences simultaneously, thus providing better generalizability of findings.

Finally, we feel we have produced supporting evidence allowing practitioners to "walk away" with a new research tool in their chest of approaches. We feel we have provided findings that allow one to comfortably consider Elrod's telephone conjoint approach as a viable method for administering a conjoint study *via* telephone.

Future Research

Here we simply offer a few ideas for possible follow-up studies concerning TCA. One rather obvious comparison would be to test TCA and ACA using a telephone-only data collection. In fact, it would seem that ACA is the only other viable alternative to telephone-only conjoint studies (stated differently, TCA is the only alternative to ACA). To really utilize the capabilities of the approaches, it would make sense to include a large number of attributes in such a study.

Secondly, one might carry out phone-mail-phone studies with RFP and CFP materials. An interesting study would involve TCA, ACA, RFP, and CFP, where the responses to the conjoint task were always taken over the telephone. It would be possible to compare utilities, choice shares, and to use ANOVA with the four treatment conditions to compare errors in predictive validity.

Expanding on the ANOVA idea, one might create a more complex factorial experiment to discover the circumstances under which the various telephone collections did or did not work well. Such a design might experimentally vary the number of attributes, numbers of levels, subject matter of the conjoint tasks, sample compositions, and so on.

Finally, it would be of interest to compare predictive validities not simply for holdout questions, but for "real-world" behaviors. In fact, our holdout questions may be predictable in part because they share much methodologically with the items used to predict them. But how would the various methods compare in predicting choices in the marketplace, separated from the conjoint task in both time and context? That kind of study would be a more powerful test of how the different conjoint approaches compare. For example, the variety of conjoint tasks tested in our investigation could be carried out with respondents participating in a panel study. In that case, comparisons of predictive validities would involve the actual behaviors that one ultimately would like to predict.

Conclusion

In the final analysis, we have tested Elrod's telephone conjoint design against current standards and conclude that, at least in this study, it has performed quite well. It did not perform significantly worse that RFP, CFP, or ACA. Thus, one "gets" a method that appears to offer much promise in its ability to extract correct information, for potentially large numbers of attributes, and to be able to do so with telephone-only data collection.

The "price" paid for these benefits is a slight non-orthogonality in the design matrix, possible loss in realism due to the partial profile nature of the task, loss of ability to model interaction effects, and no

integrated software for design and analysis. Nevertheless, given the straightforward design algorithm (Appendix A), available software for logit modeling (for example, SAS, SPSS, and SYSTAT), and the reasonability of main effects models, it would seem Elrod's telephone conjoint is a viable and valuable addition to the marketing researcher's available "tool kit."

The authors thank Terry Elrod for his guidance in modifying the experimental design for telephone conjoint analysis, and Dick R. Wittink and Daniel C. Lockhart for their helpful criticisms of earlier drafts of this paper.

REFERENCES

Addelman, Sidney (1962). "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments." *Technometrics*, 4 (February) pp. 21-46.

Cohen, Jacob and Patricia Cohen (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd ed. (Hillsdale: Lawrence Erlbaum).

Elrod, Terry (1991). "Conjoint Analysis by Telephone." revision of a paper presented at the 1990 TIMS Marketing Science Conference.

Elrod, Terry, Jordan J. Louviere and Krishnakumar S. Davey (1992). "An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models." *Journal of Marketing Research* (August) forthcoming.

Green, Paul E., Abba M. Krieger and Manoj K. Argawal (1991). "Adaptive Conjoint Analysis: Some Caveats and Suggestions." *Journal of Marketing Research* 28 (May) 215-22.

Green, Paul E. and V. Srinivasan (1990). "Conjoint Analysis in Marketing Research: New Developments and Directions." *Journal of Marketing*, 54 (October) pp. 3-19.

Huber, Joel C., Dick R. Wittink, John A. Fiedler and Richard L. Miller (1991). "An Empirical Comparison of ACA and Full Profile Judgments." *Sawtooth Software Conference Proceedings*, pp. 189-202.

Johnson, Richard M. (1989). "Assessing the Validity of Conjoint Analysis." *Sawtooth Software Conference Proceedings*, pp. 273-80.

Lockhart, Daniel C. (In Press). "Mail and Telephone Surveys." *Handbook of Marketing Research* (Richard p. Bagozzi, ed.)

Louviere, Jordan J. (1988). "Analyzing Decision Making: Metric Conjoint Analysis." Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-067. Beverly Hills: Sage.

Louviere, Jordan J. and George Woodworth (1983). "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data." *Journal of Marketing Research*, 20 (November) pp. 350-67.

Oliphant, Karen, Tom Eagle, Jordan Louviere and Don Anderson (1992). "Cross-task Comparison of Ratings-Based and Choice-Based Conjoint." Paper presented at the AMA's Third Annual Advanced Research Techniques Forum, and *Sawtooth Software Conference Proceedings*, (this volume).

Appendix A

Constructing TCA Experimental Designs

One may best represent the TCA experimental design as a matrix with one row per choice question and one column per attribute. Cells in the design matrix denote attribute differences, that is, a row vector is choice alternative A's design vector minus choice alternative B's design vector. Thus a 0 represents an attribute not differentially possessed by the two choice alternatives (and hence not mentioned in the choice question); a -1 denotes an attribute present on the first alternative but not on the second; finally, a 1 implies an attribute absent from the first choice alternative but present on the second. The design for the TCA in this study, for example, is:

```
 1 -1  0  0  1
 1  1 -1  0  0
 0  1  1 -1  0
 0  0  1  1 -1
-1  0  0  1  1

 1  0 -1  0 -1
-1  1  0 -1  0
 0 -1  1  0 -1
-1  0 -1  1  0
 0 -1  0 -1  1
```

This design is nearly orthogonal. Like other forms of choice-based conjoint analysis, the design matrix is a matrix of attribute differences, and is not necessarily orthogonal. The condition number of this matrix is 3.5, implying near orthogonality. Each attribute appears equally often in the first and second choice alternative and the number of attributes present in both the first and second alternatives are as often one as two. Finally, all choices have one alternative with one attribute present and one alternative with two other attributes present, so no alternative is dominated by its counterpart.

Note that the design above could be split into two blocks, so that a given respondent need only answer as many choice questions as there are attributes.

The design can be generalized for any number of attributes (A) as follows:

If A is odd:

Number of blocks (T) = (A-1)/2.

1.Design the first row of Cth block:

If C is odd: 1 [(C-1) zeros] -1 [(A-C-2) zeros] 1

If C is even: 1 [(C-1) zeros] -1 [(A-C-2) zeros] -1

For the remainder of each block, shift cells one place to the right (rightmost cell wraps around) for the next A-1 lines.

Special case if T=odd: make two versions of the Tth block, one as above as if C were odd, and one as above as if C were even.

If A is even:

Number of blocks (T) = A/2.
Same as above for first T-1 blocks;
Tth block has two parts to be concatenated: T_1 and T_2.
If T is odd:
First row of T_1: 1 [(C-1) zeros] -1 [(A-C-2) zeros] 1
First row of T_2: -1 [(C-1) zeros] 1 [(A-C-2) zeros] 1
For both, rotate cells one place right for each of the next (A/2)-1 lines.
If T is even:
First row of T_1: -1 [(C-1) zeros] 1 [(A-C-2) zeros] -1
First row of T_2: 1 [(C-1) zeros] -1 [(A-C-2) zeros] -1
For both, rotate cells one place right for each of the next (A/2)-1 lines.

Appendix B

Examples of CFP, TCA, RFP and Validation Question Types

Sample TCA question:

If two supermarkets were alike in all other ways, would you rather shop at . . .
[ ] One with a pharmacy or
[ ] One with video tape rental and double coupons

Sample CFP question:

If two supermarkets were alike in all other ways, would you rather shop at . . .
[ ] One with a pharmacy, double coupons and 24 hour service or
[ ] One with video tape rental and warehouse prices
[ ] Neither of the above

Sample RFP question:

Described below are eight supermarkets. Please read through the descriptions and assume that services not specifically listed are not offered. In the space to the left of each supermarket, please rate the supermarket on the scale shown below, where 1 means 'probably would not shop here' and 10 means 'probably would shop here.'

Probably **would not** shop ... Probably **would** shop
1 2 3 4 5 6 7 8 9 10

____ One with a pharmacy and 24 hour service

Sample Validation Question:

If three supermarkets were alike in all other ways, would you rather shop at . . .
[ ] One with a pharmacy and 24 hour service,
[ ] One with double coupons and warehouse prices or
[ ] One with video tape rental

Appendix C

ACA Interview Control Parameters

Interview limit: 30 minutes

Ask unacceptibles or most likelys: no

Maximum attributes in pairs section: 5

Maximum number of pairs questions: 3

Rating scale maximum for pair questions: 9

Number of attributes in first pair: 2

Number of pairs at each stage: 1

Number of attributes in last pair: 3

Prohibit any paired attributes: no

* ACA automatically limits interview to 3 pairs for a problem of this size

Appendix D

Examples of ACA Question Types

Preference Ranking:

Type the number of your first choice, assuming everything else to be equal:

1 Has a pharmacy
2 Does not have a pharmacy

Attribute Importance Rating:

If two supermarkets were alike in all other ways, how important would this difference be?

A = Has a pharmacy
versus
B = Does not have a pharmacy

4 Extremely Important
3 Very Important
2 Somewhat Important
1 Not Important at All

Graded Paired Comparison:

Which would prefer? Please type a number from the scale below to indicate your preference.

Has a pharmacy	Does not have a pharmacy
Does not offer double coupons	Offers double coupons

Strongly Prefer Left 1 2 3 4 5 6 7 8 9 Strongly Prefer Right

Appendix E

Further Comparisons

The predictive validity of RFP and ACA have been compared, most recently by Huber *et al.* (1991) and Green *et al.* (1991). This study allows two further pairwise comparisons of predictive validity: that of CFP compared to RFP and that of CFP compared to ACA. The former is the subject of a pair of very recent studies, Elrod, *et al.* (1992) and Oliphant, *et al.* (1992), but to our knowledge the latter has never appeared in print.

CFP versus RFP

The TCA questions add 10 unique holdout choice shares to the 20 from the validation questions. With 27 degrees of freedom, the critical t for a two-tailed test of the difference in correlations between $r_{RFP,\,Holdout}$ and $r_{CFP,Holdout}$ is ±2.05 (±2.78 correcting for multiple comparisons). The three correlations that go into the computation of t are:

$r_{CFP,\,Holdout} = .9712$
$r_{RFP,Holdout} = .9832$
$r_{CFP,RFP} = .9648.$

These lead to a t statistic of 1.45, not significant at p=.05. This confirms the finding of Elrod *et al.* (1992) that CFP and RFP do not have significantly different ability to predict holdout choices (partial profiles in this case).

CFP versus ACA

As above, the critical t is ±2.05 (±2.78). The three correlations are:

$r_{CFP,\,Holdout} = .9726$
$r_{ACA,Holdout} = .9412$
$r_{CFP,ACA} = .9416$

which yield a t statistic of 2.15, not significant at p=.05 after correcting for the fact that we perform five such comparisons.

Comment on Chrzan and Grisaffe

Steven Struhl
Total Research Corp.

The authors have done the hardest work required in adapting conjoint analysis for use in telephone interviewing. Much careful thought obviously went into this paper, especially concerning complex mathematical problems involved in this approach. The authors appear to have resolved many of the shortcomings in Elrod's original formulation.

However, we cannot yet say that this paper proves this approach will work. The reason for this is simple: the authors have not yet tested telephone conjoint analysis by its proposed method of administration, that is, the telephone.

This is not an incidental issue that we can presume to have no effects because the procedure seems to work using pencil and paper. Methods of administration and the suitability of proposed approaches to these methods matter as much as basic mathematical problems.

All researchers quickly learn that decision tasks, conjoint or otherwise, can easily overtax respondents, and that telephone interviews are most likely to strain respondents' patience. Two factors seem to contribute to problems in telephone interviewing. First, in common with mail surveys, there are interviewer (or more accurately, non-interviewer) effects. Respondents seem to try hardest face-to-face with an interviewer. Second, the telephone limits respondents to auditory information. Most people require visual stimuli to structure and process complex tasks. As a result, it becomes all too easy for respondents to disengage themselves in telephone interviews, refusing to continue, or worse, going on but providing thoughtless and worthless answers.

All of us have seen surveys that asked a few too many — or many too many — questions, or that included too complex a task somewhere along the way. The results are not pretty, at best.

The authors stated that they did not test telephone conjoint by telephone because it would introduce a "method bias." As they designed this study, this could pose some problems, but it is not clear how these would differ in magnitude from problems posed by a study that already uses two methods of administration, pencil and paper and Sawtooth Software's ACA System.

In any event, the method bias problem could be resolved by shifting the design of the study to use two groups of respondents, one doing the exercise with pencil and paper, and the other doing precisely the same task over the telephone.

A secondary set of concerns revolve around the limited complexity of the task. Using five 2-level attributes produced a less complex design than those usually addressed by conjoint analysis. In fact, this task involves only 16 alternative store configurations in total. You probably could get most respondents to rate all 16 of these over the telephone without much difficulty.

Also, the complexity of the task likely was diminished further by including two apparently vague attributes that did not affect overall ratings. These attributes are presence/absence of a pharmacy

and presence/absence of video rentals. A pharmacy apparently does not have clear benefits unless respondents know other facts about it, such as hours of operation, pricing, turnaround time on prescriptions, acceptance of health plan discounts, and so on. Similarly, availability of video rentals probably does not mean much without information about range of titles, pricing, and rental policies. The near-zero or zero utilities of these two attributes seem to reflect respondents' inability to rate the basic descriptions provided.

The effect of this may well have been to make the decision effectively into one based on three attributes: warehouse prices, double coupons, and 24-hour-a-day service. If so, this would hardly provide a taxing test of any choice measurement method, conjoint or otherwise.

Finally, the authors stated that their design sacrifices strict orthogonality, and that this did not seem an unreasonable sacrifice to make for advantages in ease of administration. As they pointed out, many choice designs are not strictly orthogonal, in any event, and they retained much of the "power" of an orthogonal design. However, this method could only be strengthened by a more thorough investigation of the practical effects of non-orthogonality. Perhaps the authors or others may wish to compare predictions based on this approach with predictions based on standard full-profile conjoint, using some representative designs, to see if any practical differences emerge.

In conclusion, the authors have made a careful and valuable contribution toward solving the mathematical problems that existed in telephone conjoint. However, they have yet to demonstrate that, in real life, respondents can or will complete this type of task over the telephone. I would like to feel confident that this approach can work, but this will have to wait for another test, such as the one suggested earlier. In this, the task devised for the telephone would get administered both using pencil and paper and (yes) the telephone, and perhaps to two separate groups of respondents. Perhaps also these two alternatives could get contrasted simultaneously with the same problem approached by standard full-profile and ACA methods. A more definitive test would also involve a design of average or larger size, for instance something requiring 16 to 25 cards in a full-profile exercise. I hope that the authors of this paper, or others, will undertake this critical next step.

DOING CONJOINT ANALYSIS ON THE TELEPHONE

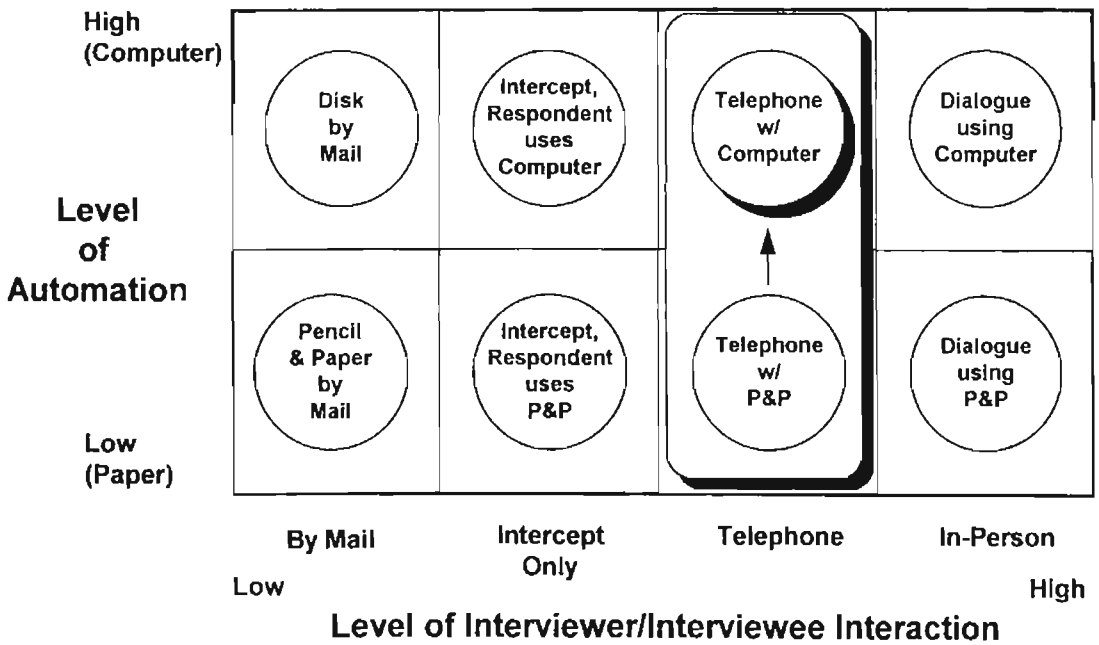
Roger Moore
Sawtooth Software

One goal of market research is to "answer a question" at minimal cost. The methodologies and media used to perform this research are extremely diverse. In this paper, I assume that the researcher has already determined that conjoint is the appropriate methodology and therefore my attention will focus on differentiating conjoint on the telephone from other possible media.

DIFFERENTIATING THE MEDIA

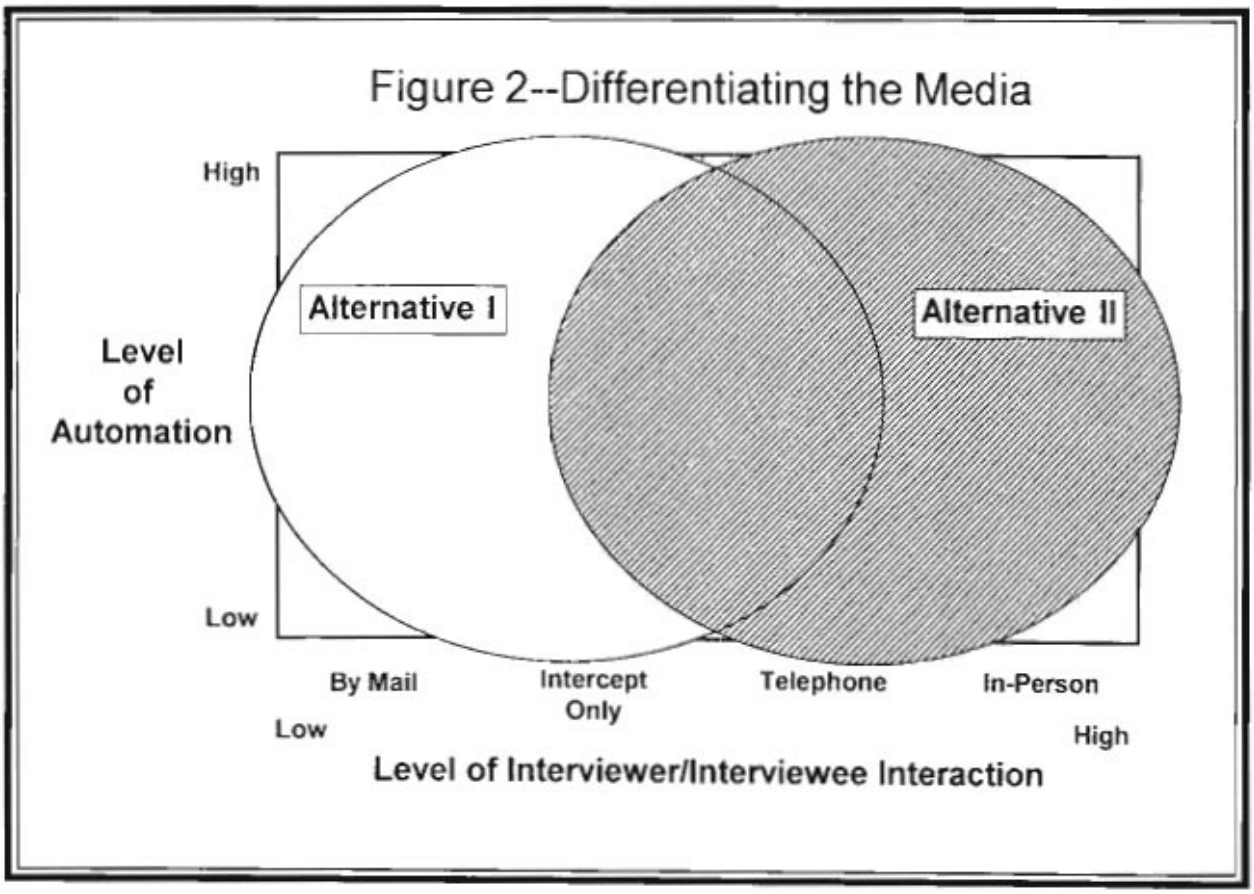
Analyzing the media can be accomplished using a simple matrix (See Figure 1). Although this matrix does not represent all possible types of interviews, it does provide a cross section of those most frequently used.

Figure 1--Differentiating the Media



The first dimension, level of automation, contrasts using a computer versus pencil and paper. From the interviewee perspective, the interview process requires computer and paper options for each level of interaction, except for the telephone. On the telephone, the target population is isolated from the data entry and therefore the decision to use a computer is essentially up to the researcher. Disk-by-mail makes sense only if the target population has access to computers. Pencil-and-paper interviews also have their purpose — some interviewees may not have access to computers or may not be comfortable with computers. In an intercept situation, some interviewees may be uncomfortable with computers and will shy away from the interview, biasing the sample — here again pencil and paper makes sense. In an in-person "dialogue" there may be instances when computers would be inappropriate to use and therefore paper and pencil may be required — taking notes on paper is routine in business meetings, however typing notes on a computer during a meeting is still considered taboo.

The second dimension quantifies the amount of interaction required, from low to high. Several factors drive the need for the different levels of interviewer/interviewee interaction: population size, product complexity, target population sophistication, geographic dispersion of population, and value of "margin notes." By establishing two alternative groups we can compare some characteristics along the interaction dimension (See Figure 2).



Characteristics of Alternative I	Characteristics of Alternative II
Larger Sample Population	Smaller Sample Population
Disperse Geographic Population	Concentrated Population
Simple Product (well known)	Complex Product (requires explanation)
Typical Consumer Product	"High Ticket" Consumer Product
	Business-to-Business Product

Conjoint by telephone falls in the high level on the automation dimension and at the intersection of the two alternatives described on the interaction dimension. To understand when to use telephone conjoint requires analyzing the advantages it provides in both of these dimensions along with other key factors. There are five areas in which a telephone conjoint offers advantages over alternative media, although each advantage is not unique to a telephone conjoint, their combination suggests that a telephone conjoint would be the best choice. These areas include: error minimization, cycle time improvement, dynamic segmentation, cost reduction, and moderate interviewer/interviewee interaction.

Error Minimization. Having the same person interviewing the respondent and entering the data limits both interpretation and transcription errors. Minimal interpretation errors occur because both the interviewee and interviewer can clear up questions they have as the survey and data entry progress. The interviewee will be able to clarify the meaning of a question; the interviewer, the meaning of an answer. Furthermore, both speed and accuracy constrain data entry. By removing the speed constraint, the accuracy is likely to improve and thus lead to fewer data entry errors. Minimizing the number of errors reduces the related costs of fixing all of the errors.

Cycle Time Improvement. Using the telephone can limit the amount of time required to get "the answer" by allowing parts of the survey process to be completed in parallel. With ACA (ACA System for Adaptive Conjoint Analysis by Sawtooth Software), data entry occurs while the survey is being administered. This approach provides the researcher with data at the end of each day. With these cumulating data, the researcher can get a head start on data analysis. The researcher can test initial hypotheses on the limited data and investigate different analytical techniques to determine the approach to take once all of the data are available. This early analysis provides an opportunity to examine formats for the final analysis and presentation. At the end of each day, the researcher can generate and examine data for accuracy, completeness, and presentation impact. The researcher can also review the results with the appropriate individuals to determine if the survey is "answering the appropriate question" and answering it at an acceptable level of detail. At the completion of data collection, analysis routines will have been fine-tuned and final presentation format established. You can often utilize automation by batches, scripts, programs, and linked documents to limit the time required for each round of analysis. For example, you can feed the survey data into a spreadsheet; perform desired calculations; link the results into charts; and link both the results and

charts into a presentation package. The presentation package can generate the final slide format utilizing the preliminary data. Once the surveys have been completed, the final data can be fed into the spreadsheet with the final numbers "linking" through to the presentation.

Dynamic Segmentation. With the data coming in at the end of each day and preliminary analysis being performed, the initial segmentation assumptions can be verified and adjusted if necessary. Furthermore, segmentation quotas can be achieved and adjusted accordingly. In pencil-and-paper or disk-by-mail interviews, if the return rate of a specific segment is lower than expected, then the researcher has to decide whether to go with a smaller sample size or send out another round of questionnaires. If the initial segmentation assumptions made for the paper-and-pencil or disk-by-mail interview prove to be incorrect then the researcher faces even more drastic options: sticking with the initial segmentation — possibly providing the wrong answer; sending out more surveys to match the current segmentation — which is costly in terms of both time and money; or answering the question with a new segmentation and current data — which is likely to be inaccurate in under-represented segments. By using a telephone, adjustments can be made on a day-to-day basis as to the segmentation scheme and the quota for each of these segments, thus limiting both time and money required to "answer the question" correctly.

Cost Reduction. If the survey involves a geographically dispersed population and an unbiased call list is available, conjoint on the telephone can be the most cost efficient method of completing interviews. With geographic dispersion, in-person interviews would be too costly. Although surveys could be mailed, hit rates are likely to be higher if respondents are contacted via telephone — a mail survey can end up on the "to-do" stacks of respondents' desks.

Moderate Interviewer/Interviewee Interaction. The use of the telephone lets the interviewer explain the product or service in terms the interviewee will understand. The interviewer can query the interviewee to ensure that the information is understood. The interaction also provides opportunities for the interviewer to get in-depth opinions on the issues, which go beyond pat answers. Interviewees will rarely put opinions down on paper, but typically will offer them up in a conversation. For a business-to-business product or service where the interviewee is an "expert," information in the margin notes can prove invaluable.

TELEPHONE CONJOINT CONSIDERATIONS AND CONSTRAINTS

In making the decision to use the telephone for the survey, several areas need to be addressed. They include: the target population, the structure of the conjoint, preparation of the interviewer, and administration of the conjoint.

Target Population

In general it is better to use a telephone for conjoint when the target population is well educated about the product and/or service being researched. The better the target population understands the product, the less likely there is to be confusion about the importance of product attributes. Furthermore, product familiarity typically means that the interviewee is less likely to have problems understanding more complex tradeoff concepts — they are already intimately familiar with the product or service and its features. In my experience, conjoints utilized to improve existing products, especially business-to-business products, work better than conjoints being used to launch new products into the general consumer market. For example, a conjoint on telephone systems targeted

toward telecommunication managers is more likely to succeed than a conjoint on a new computer technology targeted at general consumers. The telecommunication managers are likely to have an understanding of what attributes they consider important in making their business decisions about phone systems, whereas the consumers might not even understand how the new technology works, let alone how they can use it. For example, early research on cellular phones targeted at the general consumer market suggested that they would fail miserably, yet they have become an important part of today's business communications.

Structure of the Conjoint

The key to structuring a conjoint for the telephone is to keep it as simple and as short as possible. A rule of thumb is to keep the survey under 30 minutes; 15 minutes is preferable. The areas that can be controlled by the researcher include: number of attributes, the levels of the attributes, the number of attributes per concept, the scale for concept comparison, and the number of demographic/ segmentation questions.

Number of Attributes. A good rule of thumb is to keep the number of attributes to fewer than ten. If you have substantially more than that, then it's time to start examining each attribute — is it relevant to the question that needs to be answered? Use initial research, testing and in-person interviews to limit the number of attributes to those relevant to the question at hand. Customers rarely make decisions about products on more than a few attributes. Segments rarely differ on all attributes, but rather tend toward different levels of the same attributes.

Levels of Attributes. The key here is to keep a balance both in terms of the number and the length of the description of the levels of the attributes. Three is reasonable for the number of levels, which should not exceed one line for the length of the description. If the number of levels differ greatly from attribute to attribute then there is the possibility of sending audio cues to the interviewee, who might think, for example, "I keep hearing stuff about warranties, does that mean they're supposed to be important?" Limit the length of the description to keep the concept simple and balance the lengths to avoid audio cues as described for the number of levels. Remember that the interviewees cannot easily go back and reread the descriptions of each of the levels; they must ask the interviewer to repeat them, which takes time (and costs money).

Number of Attributes per Concept. Simplify this task by limiting the number of attributes per concept to just two. If more attributes are included it becomes difficult for the interviewee to remember all of the differences and they are likely to compare on just a single attribute rather than on the concept as a whole. With two attributes in a concept the interviewee must remember four descriptions of levels — about the limit for a person to remember. Increasing to three attributes raises the descriptions to remember to six — not impossible, but definitely not trivial.

Scale for Concept Comparison. ACA typically uses a nine-point scale for comparison of concepts. In my experience, this scale used on telephone conjoints causes interviewee fatigue. Simplify the scale to either two or three points to limit fatigue (and the likelihood of the interviewee terminating the interview). In a two-point scale, the interviewee simply prefers Concept A or Concept B, whereas the three-point scale offers the interviewee an opportunity for neutrality. An approach that may prove useful is to set the conjoint internally to a three-point scale, but to present it as a two-point scale. In the instances where the interviewee cannot make a choice between the two concepts, the researcher has an available option.

Demographic/Segmentation Questions. Adding extraneous questions causes interviewee fatigue and makes an interviewee less likely to focus on answering the entire survey accurately. In determining which demographic questions to ask during the survey, the researcher should concentrate on those that provide meaningful segments that are useful in "answering the question." The researcher should start with an idea of the number of segments that exist and the demographic make-up of each segment. More than five segments for a product typically provide little usable business information. These five segments should require about five, and at the very most ten, demographic questions to differentiate each interviewee.

In some instances there are demographic questions that would be nice to know but are not critical to the segmentation. Include such questions at the end of the survey — if the interviewee terminates the interview, critical information is not lost. It probably would be prudent to actually offer the interviewee the option of exiting after collecting the critical information — in small populations maintaining a good relationship with each individual is key in being able to maintain a high response rate when conducting future interviews.

Interviewer Preparation

The interviewer's understanding of the product or service is a critical success factor for getting a correct answer to "the question." If the interviewer unintentionally misleads the interviewee or loses credibility when unable to answer a question posed by the interviewee, then that survey is likely to provide useless information. There are three actions that can help better prepare the interviewer for the task that lies ahead: extensive product training, questionnaire testing, and visual cues during the survey.

Extensive Product Training. Provide the interviewer with a packet of information about the product which they can read and refer to during the interviews. If possible bring the actual product or demonstrate the service that is being tested in the survey. Try to get the interviewers involved and interested in the product, ask them to try it, ask them what they think about it, ask them what they would change to make it better. If it is unclear that they have a full understanding, give them a test on the product and don't allow them to attempt interviews until they complete the test to your satisfactio — their knowledge demonstrates credibility to the interviewee and ultimately translates into higher hit rates and more accurate data.

Two-Way Survey Testing. Once the interviewer understands the product or service then it is helpful to have that interviewer on the receiving end of the survey. This provides a final test of the interviewer's knowledge and final check of question, attribute, and level clarity. The interviewer should then run through the interview on his or her own several times. Finally, the interviewer should administer the interview to someone else to get a feel for timing and interviewee interaction.

Visual Cues Included in the Survey. Don't forget that your interviewers can see the computer screen during the interview and that the interviewee can't — use the screen to provide visual cues to the interviewer. Doing so can limit problems in complex sections of the survey by explaining exactly what is expected of the interviewer. Make sure that these cues are not confused with information to be read to the interviewees by making it a special color or putting the text all in caps or some combination thereof — but apply the cue consistently throughout the interview.

Administering the Conjoint

When it comes to actually running the surveys, there are three actions that can increase the probability of success on the telephone. These actions include: pre-testing, pre-screening, and interviewee incentives.

Pre-Testing. Design an initial set of questions, attributes and levels. Test these out in-house with people familiar with the product. Debrief each individual and determine if there are areas of the survey that were unclear, wordings that could be improved or terminology that is incorrect in the given context. Next, try the survey on experts outside the firm who are familiar with or are using the product or service to be tested. Go through the same debriefing as before. Finally test the survey on people who do not know the product. Debrief them and determine if there are ways to improve the "simplicity" of the survey. This can also be accomplished as above in training the interviewers.

Pre-Screen and Provide Information Packets. Although it slows down the cycle time, calling interviewees and sending them an information packet can ensure that they understand the product or service being tested. Include a set of pages with the levels of attributes listed on both the left and right side and a scale at the bottom to help the interviewee understand the concepts. During the interview the interviewer can have the interviewee circle the levels in each concept and then circle his or her selection on the scale. This provides the interviewee an opportunity to think about and reread each of the concepts without having to ask the interviewer to repeat it. Including the printed scale simplifies your task if using the nine-point scale cannot be avoided.

Interviewer Incentives. Providing some sort of incentive for the interviewee to complete the interview will sometimes prove helpful. However, great care must be taken not to insult the individual by suggesting his or her time is only worth some token amount. It is usually much better to convince the interviewees that by completing the interview they are actually helping provide the marketplace with a better product or service that they themselves will be able to use to their advantage. Furthermore, unless a pre-screen packet is being used, there would be no easy way of getting the tangible incentive to the interviewee without promising "the check is in the mail."

CONCLUSION

Conjoint on the telephone clearly offers advantages over alternative media. Researchers can exploit these advantages, but they must also be careful to recognize the limitations of the media. This paper has provided a framework for deciding when the use of a telephone is advantageous for conjoint and details areas that should be considered prior to executing the survey.

Comment on Moore

TELEPHONE CONJOINT: IT'S FAST, IT'S CHEAP AND YOU GET WHAT YOU PAY FOR

Marshall G. Greenberg
National Analysts, Inc.

I'd like to begin by amending Moore's opening statement to suggest that a goal of market research is not merely "to answer a question at minimal cost," but to determine the correct answer at minimal cost. This suggestion is offered neither to be flippant nor to belabor the obvious, but rather to cut quickly to the heart of what I have always believed to be the major shortcomings of any efforts to conduct conjoint analysis by telephone. We at National Analysts have sought for twenty years to develop approaches that would enable us to capitalize upon the cost and timing efficiencies of doing telephone conjoint studies, but have been unable to solve what we consider to be serious problems with the quality of the data that can be collected that way. Consequently, I believe that in all but the most trivial of tradeoff analysis problems, the conjoint model should not be administered by a telephone survey. (The comments in this discussion do not necessarily apply to data collection methodologies that combine telephone with mail administration.)

My discussion is deliberately intended to be provocative in an effort to stimulate a dialogue. It is especially designed to try to create a feeling of uneasy discomfort among users of the telephone conjoint methodology — I know you're out there — and to make explicit some of the sacrifices required to achieve its benefits.

As background, it is important to remember the primary rationale for using conjoint analysis in marketing research. It is because in certain situations conjoint analysis offers numerous advantages over self-explicated importance ratings.

Let's first review the situations or conditions under which the conjoint analysis model is most appropriately employed. They are as follows:

- Rational (non-impulse) purchases — The conjoint model works best in modeling purchase decisions in which comparisons among alternative products or services are made in a thoughtful, rational manner by the purchase decision-maker.

- Relatively high involvement decisions — The model is most appropriate when the purchase decision involves high stakes or is infrequently made, requiring that the purchaser live for a significant time with his or her choice.

- Clearly defined and clearly communicated attributes and levels — Obtaining good data requires that the survey respondent understands as precisely as possible the alternative product descriptions in terms of their attribute levels.

- Simple additive model — While not absolutely necessary, most conjoint analyses assume a simple additive model among the attributes.

When these conditions apply, the primary advantages of using a conjoint model, as opposed to self-explicated importance ratings, include the following:

- Reduces respondent's tendency to make socially desirable responses — As a decompositional model, conjoint analysis examines preferences, not stated values, and derives inferences as to what is important to a respondent.

- Simulates "real world" decision making — The data collection task in a properly conducted conjoint analysis requires respondents to make decisions under conditions as similar as possible to those under which they would normally do so. As a minimum, they should be offered:

  - Realistic product descriptions
  - Attributes evaluated in the context of other attributes
  - Time to review and reflect upon their choices

- Employs an efficient experimental design

In his paper, Moore recommends that a number of constraints be imposed on the survey and conjoint design in order to utilize a telephone methodology effectively. I agree that these constraints are both appropriate and, indeed, necessary ones. However, I submit that these — and other constraints imposed on a telephone conjoint study — exact far too great a price upon the design flexibility and upon the resulting quality and integrity of the data to justify a telephone methodology.

Table 1 below lists a series of constraints that either are, or should be, imposed on a conjoint study if telephone interviewing is employed as the sole method of data collection. Along with each constraint is shown one or more effects that diminish the value of the approach.

Table 1

Constraints Imposed by Telephone Methodology and Their Effects

Constraint	Effect
• Lack of visual material - Logos - Product styling - Prototypes	• Severely limits ability to communicate effectively many types of attributes and levels - No opportunity for respondent to reread and reflect upon descriptions
• Two attributes per concept	• Eliminates the all-important context effect — a major reason for using conjoint approach
• Number of attributes: fewer than 10	• Some products cannot be described adequately
• Number of levels: 3 or fewer	• Price and brand often require more levels to model the relevant market
• Two- or three-point scale for concept comparisons	• Sharply reduces information in data collected — we know people can make finer discriminations
• Interview length limited (15-30 minutes)	• Places restrictions on ways to segment respondents

Given the limitations outlined in Table 1, I cannot conceive of employing a telephone methodology exclusively for any but the most trivial of conjoint analysis problems (for example, a price elasticity study in which only one or two attributes of two levels each were varied.)

Many of the problems imposed by using only the telephone as a data collection instrument can be overcome by using it in combination with other methodologies, such as mail or disk-by-mail. Yes, it will take longer to complete the project. It will probably cost more as well. But I believe that your increased faith in the quality of the data collected and, therefore, in the conclusions drawn from them, should more than justify the tradeoff to you and to your client or your management.

AN EMPIRICAL INVESTIGATION OF LEARNING EFFECTS IN CONJOINT RESEARCH

Gordon Lewin
Elrick & Lavidge, Inc.
Abel Jeuland
University of Chicago
Steven Struhl
Total Research Corp.

INTRODUCTION

Background

Conjoint analysis has become one of the principal methodologies used to assess the value of alternative product features and the relative merits of alternative product formulations. While the method has enjoyed wide acceptance among practitioners and academicians, the question of the method's validity is still being debated. As stated by McLauchlan (1991) "...the ultimate value of any conjoint design resides in its ability to correctly predict choice behavior...." Naturally, it is consumer behavior in absence of a conjoint task that we want to be able to predict.

A common method used to test the internal validity of a conjoint design is the ability of the utilities generated from the conjoint test to predict holdout choices. It is possible, however, that testing effects occur in the conjoint task itself. A respondent's choices may depend in some way on the conjoint process itself. If significant testing effects do occur, they should be accounted for and corrected in the administration or analysis of a conjoint study. McLauchlan goes on to state:

> "We may be in danger of losing sight of a more fundamental challenge: developing a better understanding of the mitigating effects of the techniques themselves on individual value systems and, therefore, on results. Do respondents `learn' in full-profile ranking or rating tasks? ...Are value systems impacted [by these methods]? ...Definitive answers to these questions will evolve only from further empirical research..."

We conducted this study to measure testing effects empirically. We attempted to choose a "typical" consumer product category, one to which conjoint analysis might commonly be applied.

Through this study, we attempt to measure:

- The extent to which learning occurs in a typical conjoint task,
- The effect of conjoint versus non-conjoint tasks on respondents' holdout card choices,
- The ability of two alternative conjoint methods to predict holdout card choices.

Methodology

This study was conducted among 200 female heads of households who are category users in both the test and control categories. The test category is room air fresheners and the control category is hard-surface spray cleaners. Although the test category is not optimal for the application of conjoint analysis, we were willing to accept a suboptimal test category to achieve "realism" in the research. We felt that conjoint analysis might actually be applied to this type of category in a commercial setting. In fact, conjoint analysis was being considered by the client for this test category at the time the study was conducted. We chose a "low involvement" product category to contrast our results with a similar study in a "high involvement" category currently in process.

All consumer interviews were conducted by mall intercept in two major Midwestern cities. Interviews were completed between April 24 and May 9, 1992. Respondents were randomly assigned to one of four test or control groups:

Test Category: Air Fresheners	**Control Category: Spray Cleaners**
Cell 1: - Tradeoff conjoint task	Cell 3: - Full-profile conjoint task
Cell 2: - Full-profile conjoint task	Cell 4: - Non-conjoint task

Fifty respondents were assigned to each cell, 25 per city, for a total of 200 completed interviews - 100 per city.

All respondents completed "before" and "after" rankings of four holdout cards. The holdout card ranking task was completed separately from the conjoint task itself. The holdout cards consisted of alternative air freshener product concepts (see Exhibit 1). All respondents ranked the same set of four holdouts. Respondents were explicitly not allowed to see their previous conjoint responses or holdout card rankings at any time during the interview.

All interviewing was conducted in-person using a paper-and-pencil interview. The full-profile conjoint tasks (both test and control) were conducted by having respondents sort sets of 25 product concept cards according to their likelihood to buy each concept. A double-sort technique was employed.

The tradeoff conjoint task was self-administered and consisted of completing a set of six tradeoff matrices. Each matrix required the respondent to rank from six to sixteen pairs of attributes. Tradeoff matrices were rotated to control for order bias. Tradeoff matrices were used since they require the respondent to decompose the preference ranking task to a higher degree than does the full profile task. By evaluating two methods which depend on different degrees of decomposition of the preference ranking task, we felt that these treatments might better isolate differences in respondent learning from the two alternative conjoint tasks.

Respondents in the non-conjoint cell completed a survey relating to spray cleaners. This survey was designed to be relatively complex and take about the same length of time to complete as the three conjoint tasks required of the other respondents.

EXHIBIT 1

HOLDOUT CARDS

Holdout Card A

JAR SHAPE:
- Shape 3

FRAGRANCE:
- Fragrance 3

COLOR:
- White

DECORATION:
- Design 2

PRICE:
- $2.49

REFILLS:
- Refills Available

Holdout Card B

JAR SHAPE:
- Shape 2

FRAGRANCE:
- Fragrance 4

COLOR:
- White

DECORATION:
- Design 3

PRICE:
- $2.19

REFILLS:
- Refills Not Available

Holdout Card C

JAR SHAPE:
- Shape 1

FRAGRANCE:
- Fragrance 1

COLOR:
- Blue

DECORATION:
- Design 1

PRICE:
- $1.89

REFILLS:
- Refills Available

Holdout Card D

JAR SHAPE:
- Shape 4

FRAGRANCE:
- Fragrance 2

COLOR:
- Pink

DECORATION:
- Design 3

PRICE:
- $2.19

REFILLS:
- Refills Not Available

EXHIBIT 2

ATTRIBUTES AND LEVELS: TEST PRODUCT

JAR SHAPE:
- Shape 1
- Shape 2
- Shape 3
- Shape 4

FRAGRANCE:
- Fragrance 1
- Fragrance 2
- Fragrance 3
- Fragrance 4

DECORATION:
- Design 1
- Design 2
- Design 3
- Design 4

COLOR:
- Pink
- Green
- Blue
- White

PRICE:
- $1.89
- $2.19
- $2.49

AVAILABILITY OF REFILLS:
- Refills available
- Refills not available

EXHIBIT 3

ATTRIBUTES AND LEVELS: CONTROL PRODUCT

GREASE CUTTING ABILITY:
- Cuts through the toughest household stains and soils
- Cuts through common household stains and soils
- Cuts through light household stains and soils

ELIMINATES ODORS:
- Eliminates the toughest household cleaning odors
- Eliminates moderate household cleaning odors
- Eliminates mild household cleaning odors

STREAKS:
- Does not leave streaks on the cleaned surface
- May occasionally leave streaks on the cleaned surface
- May leave behind a slight residue on the cleaned surface

LEAVES A SHINE:
- Always leaves a bright shine
- Leaves a shine on certain surfaces

TYPE OF SCENT:
- Scent 1
- Scent 2
- Scent 3
- Scent 4

PRICE:
- $1.89 for 22 fluid oz.
- $2.19 for 22 fluid oz.
- $2.49 for 22 fluid oz.

A total of six attributes, each with two to four levels, were used for both the test and control categories. The attributes and levels for both product categories, which have been altered to protect client confidentiality, appear in Exhibits 2 and 3.

Prior to completing the interview, including ranking the holdout cards, respondents were exposed to the set of attributes included in the test. Respondents were shown cards which depicted the alternative decorations and colors, and photographs of the various jar shapes. Respondents also were exposed to each of the four fragrances in a controlled "sniff test."

The research design, then, is outlined below:

Treatment		
Obs 1 —>	- Cell 1	—> Obs 2
Obs 3 —>	- Cell 2	—> Obs 4
Control		
Obs 5 —>	- Cell 3	—> Obs 6
Obs 7 —>	- Cell 4	—> Obs 8

where observations (obs) 1-8 refer to the pre- and post-holdout card rankings. We would expect some percentage of the respondents to be inconsistent in their pre- and post-holdout rankings. If learning takes place through the process of participating in the test category conjoint exercise, then we would expect to observe a higher level of inconsistency in the pre- and post-holdout rankings in the treatment cells (Cells 1 and 2) than in the control cells (Cells 3 and 4). Furthermore, we would expect the utility scores generated from the two treatment cell responses to better predict post-holdout rankings than pre-holdout rankings.

RESULTS

Treatment Effects on Holdout Rankings:

A fairly high level of inconsistency was found between pre- and post-holdout rankings across all cells. In all but one cell, fewer than half of the respondents had identical pre/post rankings. Cell 4 (non-conjoint outside the test category) had the highest level of consistency with 30 respondents (60%) ranking the pre- and post-holdouts identically. (See Exhibit 4.) The least consistency was seen in Cell 3 (full profile conjoint outside the test category), with only 17 respondents (34%) having identical pre- and post-rankings.

Exhibit 4

PRE/POST HOLDOUT MATCHES

No. of Pre/Post Matches	Cell 1	Cell 2	Cell 3	Cell 4
0	3	5	6	1
1	7	6	9	5
2	17	19	18	14
4	23	20	17	30
	50	50	50	50

A simple comparison can be made across cells between respondents with identical pre/post rankings (that is, 4 matches) and those with differences in pre/post rankings (that is, 0, 1, or 2 matches). The calculated chi-square value for this comparison is 7.515 which has $.05 < p < .10$. Thus there is a significant difference between consistency in the four cells at the $\alpha = .10$ level. Comparing only Cells 3 and 4 for consistent versus inconsistent respondents, the calculated chi-square value is 6.988 which has $.005 < p < .01$.

The calculated chi-square value for three matching levels (0 and 1 combined, 2, and 4 matches) across all cells is 8.8616 which has a p-value >.10. Hence for the simple consistent versus inconsistent comparison a significant difference can be seen between cells. However, when comparing three levels of matches, differences between cells are not significant.

Exhibit 5 displays the results of a one-way ANOVA which was run on the mean number of matches within each cell. The computed F ratio of 3.16 has $p = .0252$. The Sheffe test (Exhibit 6) indicates that a significant difference (at $\alpha = .10$) is found between Cells 3 and 4.

We also conducted a Kruskal-Wallis one-way ANOVA on the squared differences between pre and post holdout card rankings across the four cells. Results were consistent with those of the parametric one-way ANOVA calculated on the mean number of matches. The calculated chi-square statistic generated by the Kruskal-Wallis test (corrected for ties) was 9.23, with a p-value of .03.

Thus we see that there is a significant difference in the level of consistency across the four groups both in terms of the mean number of matches and the squared differences between pre- and post-holdout rankings. The greatest difference in between-cell consistency occurs with Cells 3 and 4, the two control cells. Cell 4 respondents (non-conjoint outside the category) display the highest level of pre/post consistency while Cell 3 respondents (full-profile conjoint outside the category) show the lowest.

Respondents were generally consistent in terms of their pre- and post-first choice ranks. Across all cells, an average of 77 percent of respondents were consistent in their first choice ranks between the pre-sort and post-sort tasks. However, between-cell differences are not significant. The calculated chi-square value is 1.2420, with a p-value much higher than .50. The counts for first choice pre/post consistency are shown below.

1st Choice Pre/Post	Cell 1	Cell 2	Cell 3	Cell 4
Consistent	38	40	36	40
Inconsistent	12	10	14	10
	50	50	50	50

Exhibit 5

ONE-WAY ANOVA: MEAN NUMBER OF MATCHES FOR CELLS 1-4

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	17.2150	5.7383	3.1759	.0252
Within Groups	196	354.1400	1.8068		
Total	199	371.3550			

Group	Count	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
Cell 1	50	2.6600	1.3494	.1908	2.2765	to	3.0435
Cell 2	50	2.4800	1.3886	.1964	2.0854	to	2.8746
Cell 3	50	2.2600	1.4115	.1996	1.8589	to	2.6611
Cell 4	50	3.0600	1.2191	.1724	2.7135	to	3.4065
Total	200	2.6150	1.3661	.0966	2.4245	to	2.8055
Fixed Effects Model			1.3442	.0950	2.4276	to	2.8024
Random Effects Model				.1694	2.0759	to	3.1541

Random Effects Model — Estimate of Between Component Variance .0786

Exhibit 6

MULTIPLE RANGE TEST FOR ONE-WAY ANOVA:
MEAN NUMBER OF MATCHES FOR CELLS 1-4

Scheffe Procedure
Ranges for the .100 Level -

3.56 3.56 3.56

The actual range used is the listed range* .1901

(*) Denotes pairs of groups significantly different at the .100 level.

Mean	Group	Cell 3 2 1 4
2.2600	Cell 3	
2.4800	Cell 2	
2.6600	Cell 1	
3.0600	Cell 4	*

Homogeneous Subsets (Subsets of groups, whose highest and lowest means do no differ by more than the shortest significant range for a subset of that size.)

SUBSET 1

Group	Cell 3	Cell 2	Cell 1
Mean	2.2600	2.4800	2.6600

SUBSET 2

Group	Cell 2	Cell 1	Cell 4
Mean	2.4800	2.6600	3.0600

Comparison of Predicted and Actual Holdout Card Rankings:
Actual rankings of the holdout cards (pre- and post-) were compared with the predicted rankings based on both the full-profile and tradeoff conjoint analyses. Respondents could have 0, 1, 2 or 4 matches between predicted rankings and each holdout card sort. High levels of pre/post inconsistency were observed in both the tradeoff conjoint and full-profile conjoint cells (Cells 1 and 2, respectively). For the full-profile cell, significantly more matches were found between predicted and post-holdout rankings than between predicted and pre-holdout rankings. This was not true of the tradeoff conjoint cell.

Full Profile Conjoint:

The counts of matches (predicted versus actual ranks for the four cards) pre-sort and post-sort, ran as follows:

	None	One	Two	Four
Pre-sort	15	12	14	9
Post-sort	10	7	21	12

Testing the difference in the mean number of matches (predicted versus actual) pre-sort to post-sort with a dependent t test produced a t-value of -2.77 (p=.008), reflecting a higher mean number of matches between the predicted and actual post rankings than between predicted and actual pre rankings.

The crosstabulation below (Exhibit 7) shows these results in more detail.

Exhibit 7

PRE-SORT BY POST-SORT MATCHES
BASED ON FULL-PROFILE CONJOINT ANALYSIS

	Cells Show Counts	Pre-sort Matches 0	1	2	4	Row Total
Post-sort Matches	0	7	3	0	0	10 20.0
	1	2	3	2	0	7 14.0
	2	6	6	7	2	21 42.0
	4	0	0	5	7	12 24.0
	Column Total	15 30.0	12 24.0	14 28.0	9 18.0	50 100.0

We also looked at first choice hits for pre- and post-holdout rankings. For the full-profile conjoint data, the predicted first choice holdout card matched the actual 23 times, or in 46 percent of the cases, for both pre- and post-holdout card rankings. Therefore, while the mean number of matches of predicted to actual ranks was higher for the post rankings, the number of first choice hits pre- and post- was the same.

Tradeoff Conjoint:

The number of matches between predicted and actual holdout card rankings do not improve from pre to post for the tradeoff conjoint cell. Counts of predicted and actual matches are shown below.

	None	One	Two	Four
Pre-sort	12	22	13	3
Post-sort	16	19	11	4

The dependent t test on the mean number of matches produces a t-value of 0.00 which shows no difference between the pre and post matches.

Exhibit 8 shows the pre-sort by post-sort matches for the tradeoff conjoint cell. There is little consistency between pre and post rankings. This may represent randomness in the responses

Exhibit 8

PRE-SORT BY POST-SORT MATCHES
BASED ON TRADEOFF CONJOINT ANALYSIS

	Cells Show Counts	Pre-sort Matches 0	1	2	4	Row Total
Post-sort Matches	0	9	6	1	0	16 32.0
	1	3	10	4	2	19 38.0
	2	0	3	8	0	11 22.0
	4	0	3	0	1	4 8.0
	Column Total	12 24.0	22 44.0	13 26.0	3 6.0	50 100.0

or could imply that respondents changed their post rankings due to participating in the tradeoff conjoint exercise.

Based on the utilities generated from the tradeoff conjoint analysis, the predicted first choice holdout cards matched the actual only 15 times (30%) for both pre-sort holdout card rankings and 16 times (32%) for the post-sort rankings.

CONCLUSIONS

More than half of the respondents displayed at least some level of inconsistency in their pre/post holdout ranks. Pre/post consistency, measured in terms of the number of pre/post matches, was

relatively poor in all of the test cells. First choice consistency pre and post was fairly good, however. This may be attributable to the product category in which the test was conducted. Several of the attributes which were tested are very qualitative — color, fragrance, and design. Respondents may have been able to identify one of the holdout concepts that appealed to them, but were not consistency able to determine a second, third, or fourth choice.

Treatment effects were evident in two of the cells under study. Respondents who participated in a non-conjoint task outside of the test category (the category for which they ranked holdout cards) had more pre/post matches than respondents who participated in a conjoint task outside of the test category. Respondents who completed the non-conjoint task also had more pre/post matches than those who completed either of the two conjoint tasks in the test category, although the difference was not statistically significant. We would expect the non-conjoint task outside of the category to introduce the least interference with respondents' pre/post holdout card rankings. Conjoint exercises, either in or out of the category, appear to have a confounding influence on pre/post holdout rankings.

Utilities generated from either the full-profile rankings or the tradeoff matrices were not able to predict holdout choices very accurately. Therefore, it is difficult to assess the level of learning which may have taken place by respondents who participated in the in-category conjoint tasks. However, while the full-profile conjoint utilities were not strongly predictive of holdout behavior, they did predict post-conjoint holdout choices better than pre-conjoint holdout choices. Learning effects appeared to occur among respondents who completed the full-profile conjoint task.

Respondents completing the tradeoff conjoint exercise were not consistent in ranking pre- and post-holdouts. The utilities calculated from these data do not predict post-holdout choices better than pre-holdout choices. Therefore, while learning may have taken place in this cell we were not able to isolate such effects.

Due to the small base size of each cell, some effects may have been too subtle to identify in this study. These data should be considered a first step to further investigations of learning effects in conjoint research. The authors are in the process of investigating such effects in another product category.

REFERENCES

Chrzan, Keith (1991). "Unreliable Respondents in Conjoint Analysis: Their Impact and Identification." *Sawtooth Software Conference Proceedings*, 205-227.

McLauchlan, William G. (1991). "Scaling Prior Utilities in Sawtooth Software's Adaptive Conjoint Analysis." *Sawtooth Software Conference Proceedings*, 251-268.

COMMENT ON LEWIN, JEULAND, AND STRUHL

Dick R. Wittink
Cornell University

INTRODUCTION

Lewin, Jeuland, and Struhl (LJS) have collected interesting data to investigate learning effects in conjoint analysis. Their work is innovative and it should stimulate others to build on their findings. Such possible effects are important to understand. For example, it is possible that the predictive validity of conjoint results is overstated, if the holdout data used for this purpose are influenced by respondents' completion of a conjoint task. The conjoint task may alter what respondents rely on for preference (or choice) judgments. That is, the roles different attributes play in a holdout task following conjoint may be different from the roles those same attributes play in the same holdout task prior to or in the absence of conjoint. In this comment, I review the objective of the LJS study, the experimental design, and the statistical analyses. I also give suggestions for further research on this topic.

OBJECTIVE OF THE STUDY

LJS propose to determine the existence of learning effects in conjoint. Learning takes place if the manner in which respondents make decisions about products is influenced by a conjoint task. For example, the manner in which attribute information is processed may change. Also, given a preference or choice model, the importance or weight of attributes may change (see Huber *et al.* in this volume). Or, the predictive validity of judgments or choices may be influenced. LJS focus essentially on the last aspect.

To the extent that conjoint analysts include a holdout task, the existence of learning effects may bias the predictive validity of conjoint results. That is, the conjoint exercise may get respondents into a specific mode of integrating data. And they may continue to use this mode in the holdout task. If this mode is not representative of marketplace choice behavior, it is easy to see that the predictive validity results based on the holdout task are biased upward. This can be viewed as a different version of overstatement or bias in predictive validity resulting from (excessive) similarity in conjoint and holdout tasks.

The study's objective appears to be to determine if there is a difference in predictive validity between holdout data collected prior to and after a conjoint task. Given that the two holdout tasks are identical to each other, it is also possible to see how consistent the holdout data are. And, in a between-subject design in which different exercises are inserted between the two holdout tasks, it is possible to see whether the consistency depends on: a) the use of a conjoint versus a non-conjoint exercise; and b) the type of conjoint exercise.

THE EXPERIMENTAL DESIGN

LJS have a four-cell experimental design. The four cells involve: i) a tradeoff conjoint exercise for the same product category as the holdout tasks; ii) a full-profile conjoint exercise for the same product category; iii) a full-profile conjoint exercise for a different product category; and iv) a non-conjoint exercise for a different product category. However, the reason for including cell iii in the design is unclear. Having different but somewhat related product categories for the conjoint and the holdout tasks can confuse respondents, and this confusion can increase the inconsistency in the two sets of holdout data. By including cell iii, LJS may have used an experimental design that exaggerates the effect of a conjoint task in the same product category relative to not having respondents complete a conjoint task.

In the non-conjoint exercise, respondents provided primarily information on product usage. Thus, the data for cells i, ii, and iv can provide information about:

- whether the consistency in holdout judgments depends on whether a conjoint task for the same product category or a non-conjoint task (for a different product category) is completed;
- whether the specific conjoint method (full profile versus tradeoff matrix) influences the consistency in holdout judgments;
- whether the predictive validity to the post-conjoint holdout task data is greater than to the pre-conjoint holdout task data; and
- whether the conjoint methods differ in predictive validity.

Apart from the problem due to the inclusion of cell iii in the design, it is also unclear why the tradeoff matrix approach is used. This method has been said to be virtually obsolete. Indeed, in a commercial survey of conjoint (Wittink and Cattin, 1989), the tradeoff matrix approach was used in only six percent of the projects in the United States during the 1981-85 period (versus 61 percent for full profile). This period largely preceded the availability of Sawtooth Software's ACA System. In Europe, during 1986-90 (Wittink *et al.* 1992) tradeoff matrices were used in 15 percent of the projects, but ACA had 42 percent usage (versus 24 percent for full profile). Thus, ACA would have been a more appropriate method to use for the study.

For the determination of predictive validities it is important that the methods are implemented appropriately. LJS obtained a rank order, according to likelihood of purchase, for 25 full profiles (although rating scales dominate commercial conjoint applications). For the tradeoff matrix approach, six out of fifteen possible matrices were used. With six attributes in the study, this means that each attribute was used twice. Unfortunately that makes it likely that the tradeoff parameter estimates depend greatly on which attributes were paired, thereby affecting the predictive validity for this method.

THE STATISTICAL ANALYSES

LJS, using a chi-square test, obtain statistically significant differences in consistency in the distribution of holdout choice matches between the four cells in the experimental design. The largest

difference occurs between cells iii and iv. However, cell iii should not have been included in the design. If cell iii is eliminated, a chi-square test produces nonsignificant differences. LJS also use ANOVA to compare the mean number of matches. However, given that the holdout task consists of a ranking of four profiles, the possible number of matches is restricted to 0, 1, 2, or 4. It is inappropriate to assume that the difference in consistency between 4 and 2 matches is twice the difference between 2 and 1 matches. This assumption is made if the mean number of matches is computed. These difficulties compromise the evidence LJS report that the completion of a conjoint task (versus a non-conjoint task) influences the consistency in holdout judgments.

For cell i LJS find a significant difference in pre- and post-conjoint holdout predictive validity (although the analysis of the mean number of matches again has to be viewed with skepticism, as indicated above). However, if only first-choice predictions are considered, no difference is obtained. Even if the evidence in this paper is weak, it seems reasonable to assume that in general the predictive validity in a post-conjoint task is at least as high as the predictive validity in a pre-conjoint holdout task. Still, an appropriate question to ask is which task most closely resembles marketplace choices. The pre-conjoint task is preferred if the conjoint task does influence the choices made in the post-conjoint task. For example, one could imagine that respondents attempt to provide ranks for four holdout profiles that are consistent with the ranks for 25 conjoint profiles. On the other hand, it is unlikely that the ranking of 25 profiles is influenced by an earlier ranking (pre-conjoint) of four profiles.

At the same time, one could also imagine that the conjoint task itself may resemble aspects of consumers' activities in the marketplace. That is, consumers may reflect on the attributes, compare alternatives, and so on, prior to making a choice. If the pre-conjoint task does not allow for such reflection and introspection, the holdout choices or judgments made may lack some of these systematic considerations and thereby underestimate the predictive validity.

For cell ii there is no significant difference in pre- and post-conjoint holdout choice predictions. Thus, dependent upon the measure, full profile may produce learning while the tradeoff method apparently does not. Of course, it is good to keep in mind that the holdout task duplicates the full profile method. That is, the holdout profiles are constructed in the same manner as the conjoint profiles. For example, the order of the attributes shown on the cards is identical for the conjoint and holdout tasks. Maintaining the same order of the attributes for the description of the profiles is also likely to inflate estimates of the predictive validity of conjoint (see Huber *et al.* 1993).

SUGGESTIONS FOR FURTHER RESEARCH

Despite the lack of convincing evidence in the results of the study conducted by LJS, it is likely that learning effects exist in conjoint analysis. Learning may affect: 1) the manner in which information is combined; 2) the weights or importances associated with attributes, given a model such as the partworth model; and 3) the predictive validity of the estimated conjoint results. We still know very little about the extent to which the impact of major changes in the set of alternatives available can be predicted accurately. In the meantime, there is considerable need for further research on the topic of learning, and I hope that the LJS study will stimulate others to examine the nature of learning effects in conjoint analysis.

REFERENCES

Huber, Joel C., Dick R. Wittink, John A. Fieldler, and Richard L. Miller (1993). "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice." *Journal of Marketing Research*, February, forthcoming.

Huber, Joel C., Dick R. Wittink, Richard M. Johnson, and Richard L. Miller (1992). "Learning Effects in Preference Tasks: Choice-Based versus Standard Conjoint." *Sawtooth Software Conference Proceedings*, (this volume).

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

Wittink, Dick R., Marco Vriens, and Wim Burhenne (1992). "Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections." working paper, revised July.

LEARNING EFFECTS IN PREFERENCE TASKS: CHOICE-BASED VERSUS STANDARD CONJOINT

Joel Huber
Duke University
Dick R. Wittink
Cornell University
Richard M. Johnson
Sawtooth Software
Richard Miller
Consumer Pulse

INTRODUCTION

A current "hot topic" among conjoint users is the status of choice-based conjoint. Choice-based conjoint derives utilities directly from a series of choices among profiles. That is, instead of making judgments on the likelihood of choosing an alternative (as in full profile conjoint), or the degree of preference between a pair (as in ACA (ACA System for adaptive conjoint analysis, by Sawtooth Software)), respondents choose the best out of successive choice sets.

A major advantage of choice-based over standard conjoint is that its task more directly represents market behavior. After all, consumers do not generally rate alternatives in terms of preferences; they simply choose. Thus, one would reasonably expect choice-based conjoint to better reflect current demand. The disadvantage of choice-based conjoint is that it cannot estimate utilities for each respondent. Instead, choice-based conjoint requires the analyst to aggregate across respondents in order to derive stable coefficients, making it less effective than conventional conjoint for uncovering and defining segments.

In this study respondents complete a short choice-based task before and after ACA, enabling us to answer three questions. First, does simply taking ACA change attribute values? Second, are the results from ACA significantly different than from choice-based conjoint? Finally, can ACA be modified to approximate the choice-based results, thereby permitting us to achieve the "validity" of choices with the precision and individual-level analysis of ACA? The answer to all three questions is yes, but, perhaps the most interesting result is that the differences between ACA and choice-based conjoint follow a predictable pattern. ACA uncovers the microstructure of preferences: how a person would choose given sufficient information and time. By contrast, the choice results portray customers who are primarily motivated by brand name and price — an appropriate strategy assuming they have little time to make a decision. Thus the decision between the standard and choice-based conjoint depends on the kind of choice process that the company finds in the market. Elaborate searches are more likely to be reflected in the depth of processing captured by ACA, whereas choice-based conjoint is most appropriate in reflecting more immediate decision making.

Impact of Conjoint on Attribute Importance

Consider first how partworths can be changed by the environment of a conjoint task. A conjoint task, generally, and ACA in particular, separates attributes so that they can be evaluated independently

from the levels of other attributes. This separation occurs through two mechanisms. Respondents are often asked to assume that levels of other attributes are constant, and to focus on the particular ones displayed. Further, the attributes that are displayed are uncorrelated; the level of one attribute in a profile is not associated with the level of other attributes. Indeed, a necessary condition for an orthogonal array is that the probability of finding one attribute level remains unchanged regardless of other attribute levels in the profile. This independence enables the orthogonal array to efficiently derive partworths since the estimate for one attribute is not biased or contaminated by the others.

In contrast, attributes in the marketplace are anything but independent. Indeed it is their dependence that enables people to make rational decisions with just a few attributes. Perhaps the best example of an attribute used to make a decision is that of brand name. Brand name provides substantial information about likely performance, expected benefits and problems in use, as well as the relative price of the item. For example, a person considering an IBM PC might justifiably expect very solid quality and service, good but not outstanding performance, and a price slightly higher than could be found through a knowledgeable search.

In the conjoint environment, such inferences are soon shown to be unreliable. There one finds IBM models whose prices and performance levels range unpredictably from very high to very low. In short, the brand name, which is a reasonable criterion in the marketplace, becomes less so in the orthoganalized world of a conjoint exercise. This lessened value of brand as an indicator of quality leads to the prediction that names, such as IBM, will come to have less importance following a conjoint exercise.

Next consider price. High price conveys two pieces of information. A high price conveys negative information about the sacrifice of having to pay more, while at the same time, conveying positive information about high quality. Thus there is the sacrifice aspect of high price which is almost always aversive, and the inferential aspect which is almost always desired. Conjoint, however, tends to drive out this inferential mechanism since prices are, by design, unassociated with quality levels found. To the extent that the positive inferential mechanism attached to price is minimized, the importance of the sacrifice aspect of price should increase. Thus the expected effect of conjoint on price is opposite to that predicted for brand name. Price should increase in importance (become more aversive) as the association from high price to high quality is diminished in a conjoint task.

Summarizing, we hypothesize that the process of conjoint decouples attribute associations. That is, by showing attributes mixed with other attributes in unexpected combinations, respondents learn that inferences from one attribute to others are ineffective. The conjoint process thus teaches respondents to evaluate attributes individually and not by their inferred levels on other attributes. Thus a "name" brand will become less important if one cannot infer other good qualities from it, and a good (low) price should become more important because one does not infer other negative qualities on the basis of that price.

Most other attributes, like weight, size, and performance can be expected to act like price. That is, they are negatively correlated with each other in the environment. People will generally expect that a better level in one requires a sacrifice in the others. For example, decreasing the weight of a laptop will generally exact a penalty in terms of size, performance, or price. To the extent that a conjoint task decouples associations among negatively correlated attributes, then these attributes should become more important. This thinking leads to the hypothesis that a conjoint exercise makes brand name less important and most other attributes more important.

Below we describe a study that evaluates the impact of taking a conjoint task on attribute importance. Choice-based conjoint is used to measure changes in attribute weights and consistency due to an intervening ACA task. Then we consider the issue of the degree to which ACA matches these choice tasks and how it might be possible to convert from one to the other.

A Study to Examine the Impact of a Conjoint Task on Attribute Values

To evaluate the effect of conjoint on attribute values, we need a way to measure these values that is minimally invasive. Choice-based conjoint provides a fine way to measure relative tradeoffs at the group level and can be accomplished in a very short period of time. The technique we use is randomized choice-based conjoint. With this procedure, each respondent makes choices from a small number of sets whose attributes are randomly drawn from the set of all manipulated attribute levels. A multinomial logit model then derives partworth utility scores that predict these choices. The choice-based attribute partworths are analogous to standard partworths except that they predict the probability of choice rather than a preference rating.

The study (Wittink *et al.*, in this volume) involves the selection of laptop computers conducted by IntelliQuest. ACA is used to evaluate six attributes of laptop computers: brand name, price, relative speed, battery time, total weight and exterior size. Each of these attributes has four levels, with ranges shown in the first column of Table 1. The conjoint choices, given just before and after the ACA exercise, involve two pairs and two triples. For each respondent, profiles are randomly generated from the 6^4 possible profile combinations. These same choices are then repeated after ACA. Repeating the choice sets gives us a measure of consistency within respondents, while making them differ randomly across respondents permits us to make stable estimates of aggregate utilities and to test for any possible interactions.

For ease of exposition here, the four-level attributes are linearized to have proportional differences but to range between 0 and 1. Thus, for example, the four weights (5, 6, 7, 8) are normalized (to 0, .33, .66, 1). This normalization means that a coefficient for an attribute measures the value of its range. Brand name could not be linearized so its four levels are represented by the contrast between that brand and Librex.

The utilities shown in Table 1 reflect the raw utilities from the logit analysis divided by the consistency measures in the bottom row. We modified the raw utilities so that the various attribute coefficients would be comparable across methods. Understanding the reason for this transformation requires a brief digression into the nature of multinomial logit.

Multinomial logit defines the error level as a constant, so that greater fit is represented by larger coefficients. Those who use logit may have seen evidence of this property in the disquieting finding that, as one increases fit by adding more parameters, the absolute values of the original coefficients tend to rise. Logit can be usefully contrasted with linear regression in this regard since regression coefficients remain unbiased as one adds noise. Thus, if one adds random error to a dependent variable in regression, the coefficients remain quite constant, although the standard errors around those coefficients increase. With logit, since the error term is fixed, adding noise to the choices makes the coefficients smaller, while keeping the standard errors the same.

This property of logit in which the coefficients change, rather than the error term, hinders comparisons of coefficients across different data sets. For example, if the coefficient for price is different before and after ACA , it is unclear whether that is due to greater noise after conjoint or to

structural differences in relative utility value. Table 1 separates the consistency (or noise) component from the relative utility component. The partworths are scaled by a linear dilation through the consistency factor so that the utility coefficients for the second choices are as close as possible to those for the first choices.

An examination of the rescaled utilities for the choice set before and after conjoint supports the hypothesis given earlier. Brand name becomes less important as indicated by the lower adjusted utility values for the three brand dummy variables. By contrast, after ACA, price, performance and battery time become more important. The changes in calculator weight and exterior size are not statistically significant. Overall, these results show that after a conjoint exercise, brand name becomes relatively less important, and other attributes more important, as predicted by the idea that conjoint reduces associational links among attributes.

A second result is that the respondents are approximately 40% less consistent after ACA than before. This evidence of fatigue is unexpected, but may be due to the fact that the second choice set came at the end of a relatively long survey that contained a substantial section about computer usage in addition to the conjoint exercise.

Table 1
Utilities of Laptop Attributes For First and Second Choice Sets

Attribute:	First choices: Before ACA	Second choices: After ACA	ACA predicted choices
IBM vs Librex	1.23 (.11)	1.02 (.16)	0.97* (.09)
Toshiba vs Librex	1.13 (.11)	0.74* (.16)	0.77* (.09)
Dell vs Librex	0.92 (.11)	0.73 (.16)	0.77 (.09)
Price $2300 to $3600	1.61 (.08)	2.08* (.12)	1.48 (.08)
Performance: 20% better to 10% worse	1.19 (.10)	1.41 (.15)	1.35 (.09)
Battery time 4 to 2 hours	0.63 (.07)	0.83* (.11)	1.01* (.07)
Total Weight 5 to 8 lbs	0.38 (.08)	0.21 (.12)	0.68* (.07)
Exterior size 8"x10" to 10"x12"	0.07 (.07)	-.05 (.12)	0.39* (.06)
Consistency	1.0	0.6*	1.4*

* Difference between weight given and first choice weight is significant at $p<.05$

How Well Does ACA Reflect Choice-based Conjoint Results?

We have shown that simply participating in ACA affects subsequent choices. Another important question to ask is the extent to which ACA's predicted choices approximate those found in the choice-based survey. We use the ACA utilities to predict each respondent's choice using the maximum utility rule: always choose the object with the greatest utility. A subsequent logit analysis then predicts these simulated choices as a function of the attributes. This analysis, shown in the third column of Table 1, reveals two interesting results.

First, ACA has much less variability than either the first or second choice, being 40% more consistent than the first choice battery and more than twice as consistent as the second choices. This makes sense; ACA's choices come from an analysis of a logically consistent model based on about 20 judgments. The average of these is likely to be more consistent than four individual choices. Second, we see that brand name and price are less important in ACA relative to the functional attributes. This decreased importance of brand name may be due to the impact of ACA in reducing the association between the brand name and the attributes -- the same account as given for why brand name became less important after ACA. The lowering of price importance found in the ACA compared with choices requires a different explanation, since limiting price associations should increase the importance of price, rather than decrease it, as found. It may be that the focus on trading off the benefits of the alternatives, as found in ACA, underplays price. Consider the clever salesman who is able to steer customers to a higher priced product by focusing on the performance attributes. However, when one goes to choose, the critical attributes become brand name and price. In other words, focus is put on a few global attributes, rather than trying to balance a large number of attributes.

Which Measure is Correct?

The previous results raise an intriguing and important question about the usefulness of standard versus choice-based conjoint. We show that error levels and relative weights change depending on the method used to collect the data. Two results are clear. First, the fact that attribute associations are decoupled by standard conjoint leads to less reliance on brand name and to more reliance on functional attributes. Second, choice appears to additionally differ from ACA in putting more weight on price. But which measure is correct? Which measure provides a better estimate of what will happen in the market? Clearly, this study cannot answer that question, since there is no ultimate criterion by which to test the utilities found. We can however, usefully speculate on contexts in which one measure or the other may be more useful.

The big difference between ACA and choice-based conjoint is in the depth of processing. The four choices take the average respondent about two minutes, contrasting with ACA's time of 15-20 minutes. One must question how representative those choices are to those made in the marketplace. Consider first the category under study, laptops. Purchases of laptops are generally not made in anything like 30 seconds; people spend significant time discussing a wide range of features. Thus, the lack of attention to the "less important attributes" such as unit weight and exterior size may not be replicated in market choice. Further, market decisions are often made on the basis of recommendations from magazines or from people who have studied the options closely. Thus, a company following the advice of choice-based conjoint might well underestimate the importance of minor attributes and the depth of processing that occurs in market decisions. In contrast, ACA's depth may approximate these decisions better.

For consumer goods, a different argument leads the same recommendation to use standard over choice-based conjoint. On any given purchase consumers are typically very insensitive to anything other than brand name and price. Consumers learn about package goods through usage and through interacting with other users. A product that does not work well will eventually lose share in the marketplace. In a sense, the lower weight for brand names in ACA is consistent with the idea that ultimately it is features that people buy, not brand names. A brand name comes to have value over time because it is associated with good features. Choice-based conjoint appears to be representing this immediate response to the market offering and thus should better reflect current market share. However, it would be very risky, if not foolhardy, for a company to reduce features because short-run profits are maximized. Eventually, customers will learn, and that pattern of learning is likely to be more closely reflected in ACA's utilities than in choice-based ones.

Another way to express the foregoing is to say that ACA provides a reasonable normative model for decisions. It approximates customers' needs as expressed through careful tradeoffs. Thus, if a company really believes it is in its long run best interest to give customers what they need, ACA appears to give a better characterization of those needs than choice-based conjoint.

Can ACA be Adjusted to Predict Choice?

Suppose one's goal is to approximate the choice-based results, but desires the additional accuracy of ACA. That is, Table 1 indicates that choice-based conjoint puts greater weight behind brand names and price relative to the other attributes. A logit analysis predicting choice on ACA's part-utilities can be used to determine this optimal reweighting.

Table 2
Weights for ACA's Part-Utilities to Predict Choices in the First Choice Set

Attribute:	Coefficient	(Standard error)
Brand Name	1.45*	(.09)
Price	1.30*	(.07)
Performance	0.95	(.07)
Battery time	0.75*	(.08)
Weight	0.64*	(.10)
Size	0.37*	(.11)

* Coefficient is significantly different from equal weight (all 1.0), $p<0.05$.

The part-utility of each attribute is the utility assigned to the particular level of each attribute. For example, the part-utility for IBM might be .5 while the part-utility for 20% greater performance might be .7, with similar kinds of numbers for the other four attributes. Table 2 shows how the weighting assumed by ACA should be modified to best predict choice.

It turns out that the revised model predicts choices very well. The replication consistency of the first and the second choices is 71%. That is, if one uses the replicated choices to predict the first choices, one would be right in 71% of the time. However, with unadjusted ACA, the hit rate increases to 74%. Further, the reweighted ACA from Table 2 increases the hit rate to 78%. All of these differences in hit rates are statistically significant.

Thus, the ACA model of choices predicts choices better than choices themselves! Some of this result comes from the use of the erratic second choices as the basis for prediction, but much comes from the stability in the ACA utilities, compared to the noisiness of actual choice. The result does imply that one can augment the benefit of choice-based conjoint by combining it with ACA's output.

Discussion

These results tell us more about choice-based conjoint than ACA, although they are relevant to both. First, we find evidence that choice-based conjoint is susceptible to an intervening task. If respondents are tired, it may appear in the consistency of the choice-based conjoint. Second, the act of taking ACA also has the effect of increasing the relative importance of functional attributes, while decreasing the importance of brand name as a surrogate for those attributes. These results are consistent with the idea that a conjoint task generally lessens the impact of associations among attributes. Third, choice-based conjoint reveals a more simplified decision rule than does ACA, with greater emphasis on brand name and prices. Finally, we show that choice-based conjoint can be predicted quite well by modifying the relative weights given utility components for each attribute to better reflect choice behavior.

Choice-based conjoint has become increasingly popular lately, largely because its task appears closer to actual choice in the marketplace. Our results here are relevant to that perception. Choice-based conjoint appears to tap immediate decisions, where the attributes are clearly displayed for the decision maker. Most decisions for laptops, in particular, and durables in general, are not made quickly. Further, even if individual consumers may not trade off features, they read magazines and speak to people who do. Thus, there is good reason to believe that the high emphasis on brand and price found in choice-based conjoint may not accurately track long-term market response to feature improvements.

The implications for the design of conjoint studies is inescapable — collect both standard and choice-based conjoint. Standard conjoint provides a detailed image of the microstructure of customer needs, one that permits segmentation of each respondent and reveals tradeoffs for attributes that might be otherwise ignored. Choice-based conjoint, for its part, portrays a more accurate picture of short-term customer response to brand name and price. The relative weighting of choice-based and standard conjoint depends on the kinds of projections the marketing researcher needs to make. To the extent that the perceptions of features, brand names and prices are expected to remain stable in the future, then the choice-based analysis is quite appropriate. However, many strategies in the

marketplace involve altering features and relative prices — upsetting the correlational structure among the attributes. In that competitive environment, the effect of conjoint of breaking down the associations among attributes may be the best way to mimic the effects of competing companies in an active market.

Finally, the ability to predict choice-based conjoint with ACA provides a way to have the advantages of both methods, which we may call choice-adjusted ACA. In this procedure, Ci3 (Ci3 System for computer interviewing, by Sawtooth Software), can be used to build the same kind of short randomized choice task as employed in this study. Then ACA's partworth utilities are adjusted to predict these choices and provide input to the choice simulator. Choice-adjusted ACA permits aggregate choices to be predicted without losing information on each individual respondent. That is, one can segment respondents by their choice-adjusted utilities. By contrast, generally with choice-based conjoint there is not enough information from the 16 or so choices from each respondent to provide satisfactory stability for grouping individuals. Thus, ACA is needed even if one does not accept the argument that ACA's partworths are the appropriate ones for a customer-centered firm to use. It is needed to enable choice-based conjoint to better do its job.

Comment on Huber, Wittink, Miller, and Johnson

Keith Chrzan
Walker: Research and Analysis

In "Learning Effects in Preference Tasks: Choice-Based versus Standard Conjoint" Joel Huber *et al.:*

a) show the utilities of Sawtooth Software's ACA System and of "randomized" choice-based conjoint analysis tasks performed before and after the ACA task;

b) speculate on the relative merits of ACA and choice-based conjoint analysis; and

c) conclude that
   - ACA changes choice-based conjoint analysis utilities;
   - ACA and choice-based conjoint analysis produce different results; and
   - ACA utilities can be modified to approximate choice-based conjoint analysis utilities.

The attempt to calibrate ACA utilities to choice was interesting, as was the informative discussion of how ACA and choice-based conjoint analysis utilities could be compared at all.

My favorite thing about the paper is the invention of randomized choice-based conjoint analysis. This very interesting design strategy allows choice-based conjoint analysis to measure main and interaction effects by asking, potentially, a very small number of choice questions per respondent. Made viable by the extreme flexibility of Ci3 (Sawtooth Software's Ci3 System for computer-interactive interviewing), randomized choice-based conjoint analysis is an innovation that may, by itself, make Ci3 a "must have" piece of software for the choice modeler. Future research comparing the predictive validity of randomized and standard choice-based conjoint analysis would be especially interesting.

The flip side, of course, is that the authors compare a standard ACA task to a new and untested version of choice-based conjoint analysis. At best theirs is a comparison of ACA with randomized choice-based conjoint analysis. Generalizing to any conclusions about *standard* choice-based conjoint analysis is unwarranted. "It's okay to pet the neighbor's cat Fluffy" does not imply "It's okay to pet Siberian tigers."

Not that I think the data support a criticism even of randomized choice-based conjoint analysis. First, the authors argue that since choice-based utilities measured before and after an ACA task are different, ACA causes the change. Suppose every time I get a headache I take Tylenol, followed by Advil. Later my headaches go away. Did the Tylenol get rid of my headache or did the Advil? Perhaps it was the joint Tylenol-Advil effect, or perhaps either may have worked alone. Perhaps, too, my headaches are of short duration and they just go away by themselves. Similarly, learning during the first four choice questions may have caused the respondents to have answered the second four differently, or ACA may have effected the change. Perhaps the conjunction of the first four choices and ACA caused the change, or perhaps either alone may have done the job. Or maybe respondents just got tired. In the absence of an experimental design to control these confounded effects, one cannot credit ACA as the causal agent.

Nor is the suggestion plausible that ACA uniquely "decouples" attributes; that it teaches respondents not to infer other attribute levels from brand or price information. Orthogonal full-profile conjoint analysis would effect the same decoupling, perhaps even more efficiently. So too would a standard choice-based conjoint analysis built from an orthogonal main-effects plan.

Similarly, the design does not show that ACA and choice-based conjoint analysis produce different results. Knowing that learning effects occur in the course of taking multiple conjoint tasks, the most we can conclude is that a second of three conjoint tasks yields different results from a third of three (and that a first and third of three differ as well).

In summary, the authors present a novel design strategy for choice-based conjoint analysis. They also offer some interesting speculations about the relative merits of ACA and choice-based conjoint analysis, though these speculations are not supported by the data.

THE PREDICTIVE VALIDITY OF DERIVED VERSUS STATED IMPORTANCE

William G. McLauchlan, Ph.D.
McLauchlan & Associates, Inc.

INTRODUCTION

Background

Marketing researchers are often concerned with understanding the importance of various product and image attributes as related to choice behavior. If, for example, we know that consumers choose among products in a given category on the basis of length of warranty as opposed to manufacturers' return policies, we can leverage this information in the development of marketing strategies.

The determination of attribute importance is typically accomplished in one of two general ways: stated or derived measurement. Stated importance usually involves the direct rating or ranking of attributes. For ratings, respondents are shown attributes one-at-a-time and asked to indicate how important the attribute is in the product or service purchase decision. The scale points can be fully anchored:

5 Extremely Important
4 Very Important
3 Quite Important
2 Somewhat Important
1 Not At All Important

or anchored only at the extreme responses:

5 Extremely Important
4
3
2
1 Not At All Important

Another rating alternative is the Q-Sort technique which imposes a forced distribution on the scale with the intent of introducing variability into the data. The example below assumes a total of nine attributes. Only one attribute is allowed to be Extremely Important, two are allowed to receive "4" ratings, and so forth.

1 Not At All Important	2	3	4	5 Extremely Important

The ranking of attributes is really the non-parametric extension of the Q-Sort. With a reasonably limited number of attributes, respondents are asked to select the most important attribute, then the next most important, and so forth, until all attributes have been ranked.

For any of the stated importance data collection procedures, differences between mean importance ratings or rank-sums are tested using the appropriate parametric or non-parametric ANOVA and corresponding post hoc tests.

Derived Importance is more complicated and based on an interesting and yet paradoxical premise. The assumption is made that respondents will not reliably indicate what is truly important to them in a purchase decision. As such, the importance of attributes is estimated by analyzing brand performance ratings and the linkage of those ratings to some criterion behavior (such as Purchase Likelihood or Overall Satisfaction). The paradox is that while we choose not to believe respondents' importance ratings, we are perfectly happy to believe their brand ratings.

The modeling of Derived Importance can be carried out in many interrelated ways including the following:

- Correlation of brand-attribute ratings with a criterion measure
- Multiple regression of brand-attribute ratings with a criterion measure
- Multiple discriminant analysis of brand-attribute ratings with brand usage or ownership classifications
- Conjoint analysis (for hybrid models, conjoint represents a combination of stated and derived measurements)

There are numerous criterion measures that are used in derived importance modeling (for example, purchase likelihood, share of preference estimates, overall satisfaction, and brand usage). There are also a variety of data preparation alternatives for the attribute ratings. These include using raw data, respondent normalized (centered) ratings, double centered ratings, factor scores, among others.

In all of these approaches, the objective of the analysis is to account for the variance explained in the dependent measure by the brand-attribute performance ratings, under the supposition that attributes which explain more variance are more important than variables which explain less variance.

The question begged by the issue of stated versus derived importance modeling is straightforward: Which approach best identifies the attributes that are most important in a purchase decision? The answer to this question is anything but straightforward. It seems reasonable to hypothesize that the predictive validity of a given attribute importance measurement method is linked to the product category and to the manner in which products are advertised or marketed within the category.

More specifically, the importance of attributes for products that are heavily image advertised would seem better ascertained using derived importance measures. The reasoning here being that consumers are more likely to rationalize or disclaim their bases for purchases when the real reasons are largely image related. Consider two possible responses to a question such as "What was

important to you when you decided to buy a BMW?" Is a respondent more likely to say "I'm a YUPPIE and this is **the** car YUPPIEs drive" or "I bought this car because it's well engineered"?

Alternatively, the importance of attributes for products that are heavily feature advertised would seem better ascertained using stated importance measures. There is little social gain in disclaiming the real reasons for purchase. Consider two possible responses to the question "What was important to you when you decided to buy Minute Maid Orange Juice?" In this instance, is a respondent more likely to say "I'm a YUPPIE and this is the orange juice YUPPIEs drink" or "I like pulp in my orange juice"?

What is usually hypothesized is, then, a continuum of "fit" for importance measures as follows:

All Feature	<==========>	All Image
Stated Importance	<==========>	Derived Importance

The objective of the research reported here was to explore the similarities and differences in derived versus stated importance results for three product categories that fall in different places along the advertising/marketing continuum: Beer, Automobiles, Commodity Chemicals.

The research was conducted under the premise that Beer tends to be predominately image advertised, Automobiles are both feature and image advertised, and Commodity Chemicals are sold predominately on the basis of product and service features.

Methodology

Three separate studies were conducted. The Beer and Automobile data were collected specifically for the purposes of testing the hypotheses discussed above. The results for these two product categories are not disguised. On the other hand, the Commodity Chemical data were collected as part of a larger project for one of our clients. Because of the proprietary nature of this study, manufacturers and attributes have been disguised. The actual data are reported as observed or derived. Specific design elements for each project are presented separately below, followed by a summary of the questionnaire content.

Beer. In-person interviewing was used to collect the Beer data. Respondents were intercepted and screened in three geographically dispersed shopping centers (Chicago, Los Angeles, and Norfolk). Those individuals who qualified and agreed to participate were brought back to an enclosed room where the interviews were administered on PCs. Respondents keyed-in all answers to the computer-presented questions. The research employed APM (APM System for adaptive perceptual mapping by Sawtooth Software).

To participate in the study, respondents had to be between the ages of 21 and 65 and consumers of at least three beers in an average week. A total of 150 interviews was completed with respondents meeting these qualifications. Interviewing was conducted from March 4 through March 11, 1992.

Automobile. In-person interviewing was used to collect the Automobile data. Respondents were intercepted and screened in three geographically dispersed shopping centers (Minneapolis, Orlando, and San Antonio). Those individuals who qualified and agreed to participate were brought back to an

enclosed room where the interviews were administered on PCs. Respondents keyed-in all answers to the computer-presented questions. The research also employed APM.

To participate in the study, respondents had to be between the ages of 18 and 65 and own a 1989 or newer car. In addition, each respondent had to have been primarily responsible for choosing which car to buy. A total of 146 interviews was completed with respondents meeting these qualifications. Interviewing was conducted from March 4 through March 11, 1992.

Commodity Chemical. Data were collected using a self-administered paper-and-pencil questionnaire. Each survey was hand-delivered by a representative of the sponsoring organization and, on completion, returned by mail.

Qualified respondents were individuals involved in the purchasing or specification of a particular group of chemicals. A total of 220 questionnaires was completed. Data were collected in the Fall of 1991.

Questionnaire Content

For the Beer and Automobile studies, the following questions were asked during the structured APM interview:

- Brand Usage/Familiarity
- Attribute Importance Ratings (5-point, fully anchored scale)
- Brand Performance (5-point, Agree-Disagree scale)
- Pairwise Brand Preference

In each of these studies, the attributes included a reasonably balanced mix of feature and image attributes (See Tables 1 and 2).

Table 1

Automobile Attributes

- Is fuel efficient
- Has a smooth ride
- Is dependable
- Is available with the options I want
- Has a broad line of models from which to choose
- Has dash controls that are easy to use while driving
- Is a good value for the money
- Handles the road well
- Has good resale value
- Is economical to operate
- Is economical to repair
- Offers a good warranty program
- Offers rebates on a regular basis
- Is a high quality automobile
- Is fun to drive
- Is for people who take pride in their cars
- Makes a statement about my lifestyle
- Is a car for younger people
- Is a car for older people
- Is stylish
- Is good for a first time car buyer
- Is a status symbol
- Is sporty
- Is old-fashioned
- Is modern and up-to-date
- Is conservative
- Has advertising that I like

Table 2

Beer Attributes

- Has a robust, full-bodied taste
- Has a milder, lighter taste
- Is lower in calories
- Has a rich color
- Doesn't have an unpleasant aftertaste
- Is available everywhere beer is sold
- Is inexpensive
- Is thirst-quenching
- Is good with food
- Is good tasting
- Has the right amount of carbonation
- Is a less filling beer
- Is a good value for the money
- Is less expensive than most other beers
- Is for men
- Is for women
- Is for people who like to have fun
- Is an "old fashioned" brand
- Is a brand everyone likes
- Has advertising that I like
- Is for someone like me
- Is good to serve at parties
- Is for special occasions
- Is a "blue collar" beer
- Is for successful people
- Is a premium, high quality beer
- Is the beer to buy when you want to impress somone
- Is the beer I'd order in restaurants

Specific attributes are later referred to as being feature oriented, image oriented, or both. The author acknowledges the highly subjective nature of these classifications.

The Commodity Chemical study included the following types of questions:

- Overall Manufacturer Satisfaction
- Attribute Importance Ratings
- Supplier Performance Ratings

The attributes for this study were largely feature-oriented and included a mix of product and service characteristics.

Analytical Overview

The analytical plan was designed to examine the relationship between stated importance and derived importance. The following steps were completed for each data set.

Stated Importance

- Compute mean stated importance ratings and test for differences using repeated measures Analysis of Variance with SNKs (alpha=.05)

Derived Importance

- Compute bivariate correlations between brand/manufacturer performance ratings and measures of brand usage/manufacturer satisfaction

- Conduct two-group discriminant analyses using brand/manufacturer performance ratings. Group membership was based on brand used most often or manufacturer being rated.

All derived analyses were conducted with double-centered data using the following transformation:

((Raw Score - Respondent Mean) + (Evoked Set Mean - Grand Mean))

and with principal components scores based on the double-centered data.

RESULTS

Stated Importance

Mean importance ratings for the three sets of attributes are shown in Figures 1, 2, and 3. For clarity of presentation, the results of statistical testing are not shown in these charts. As a point of reference, however, a difference between mean ratings of about .3 scale points is significant ($p<.05$) for all three product categories.

Figure 1
Stated Attribute Importance
- Automobiles -

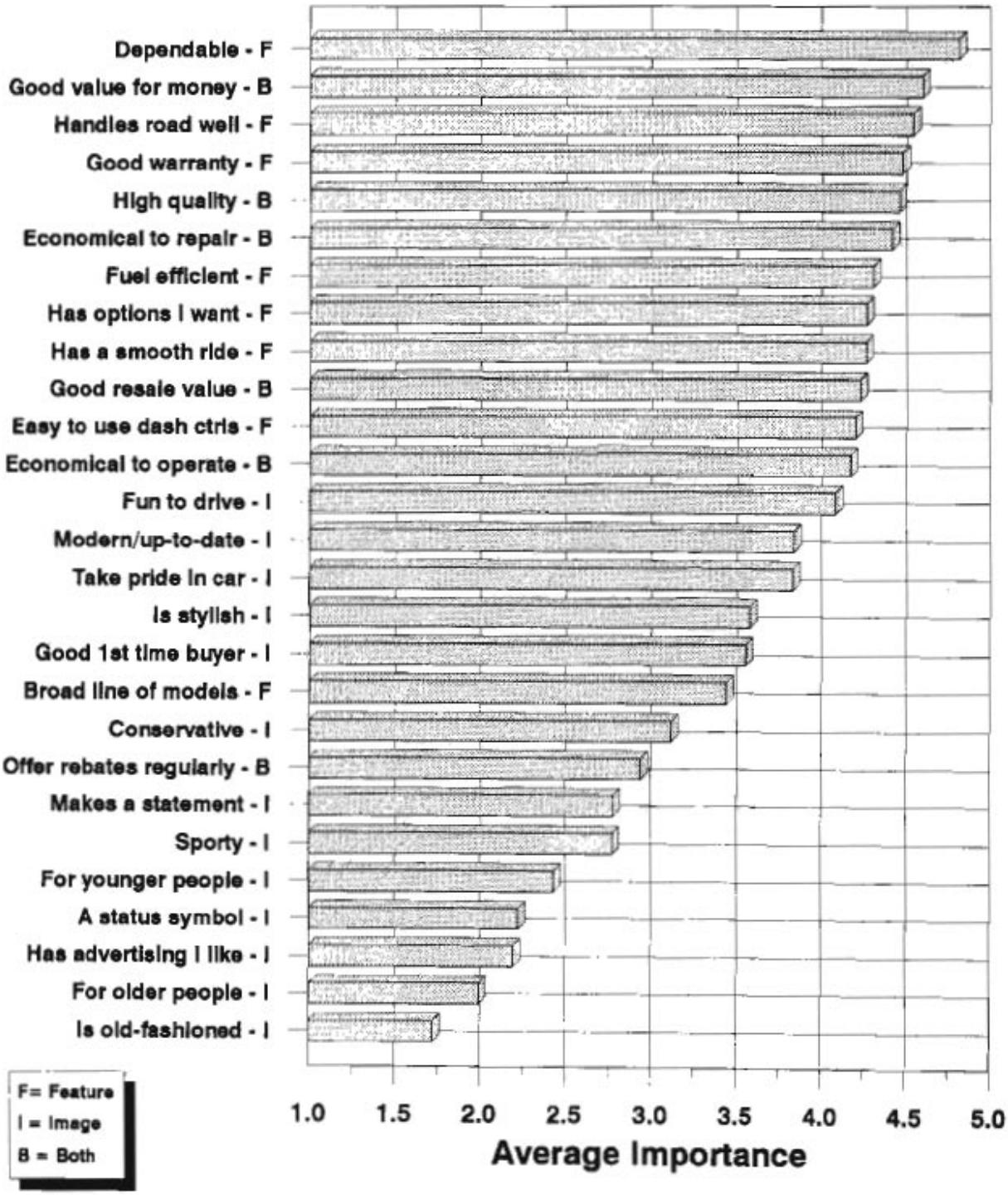


Figure 2
Stated Attribute Importance
- Beer -

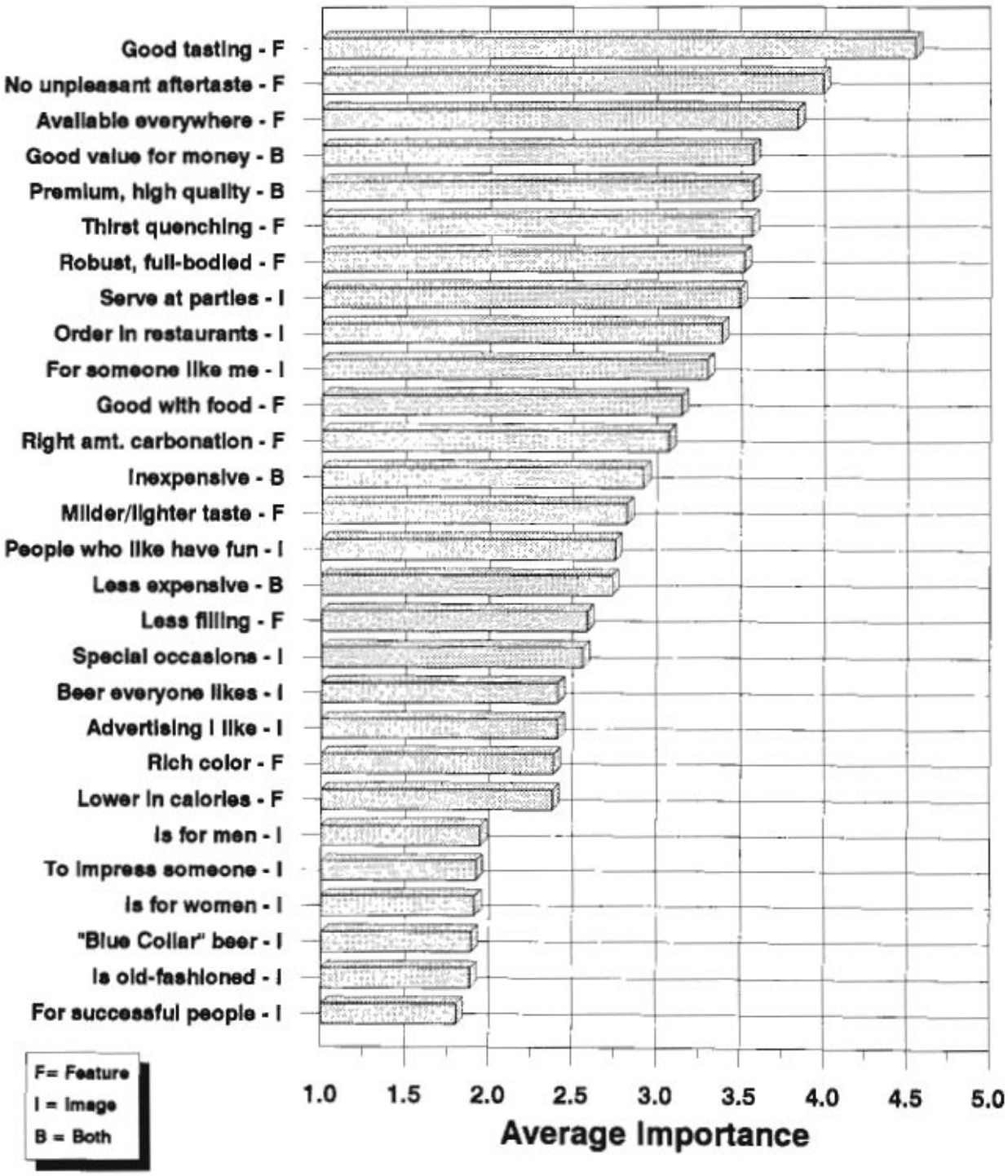
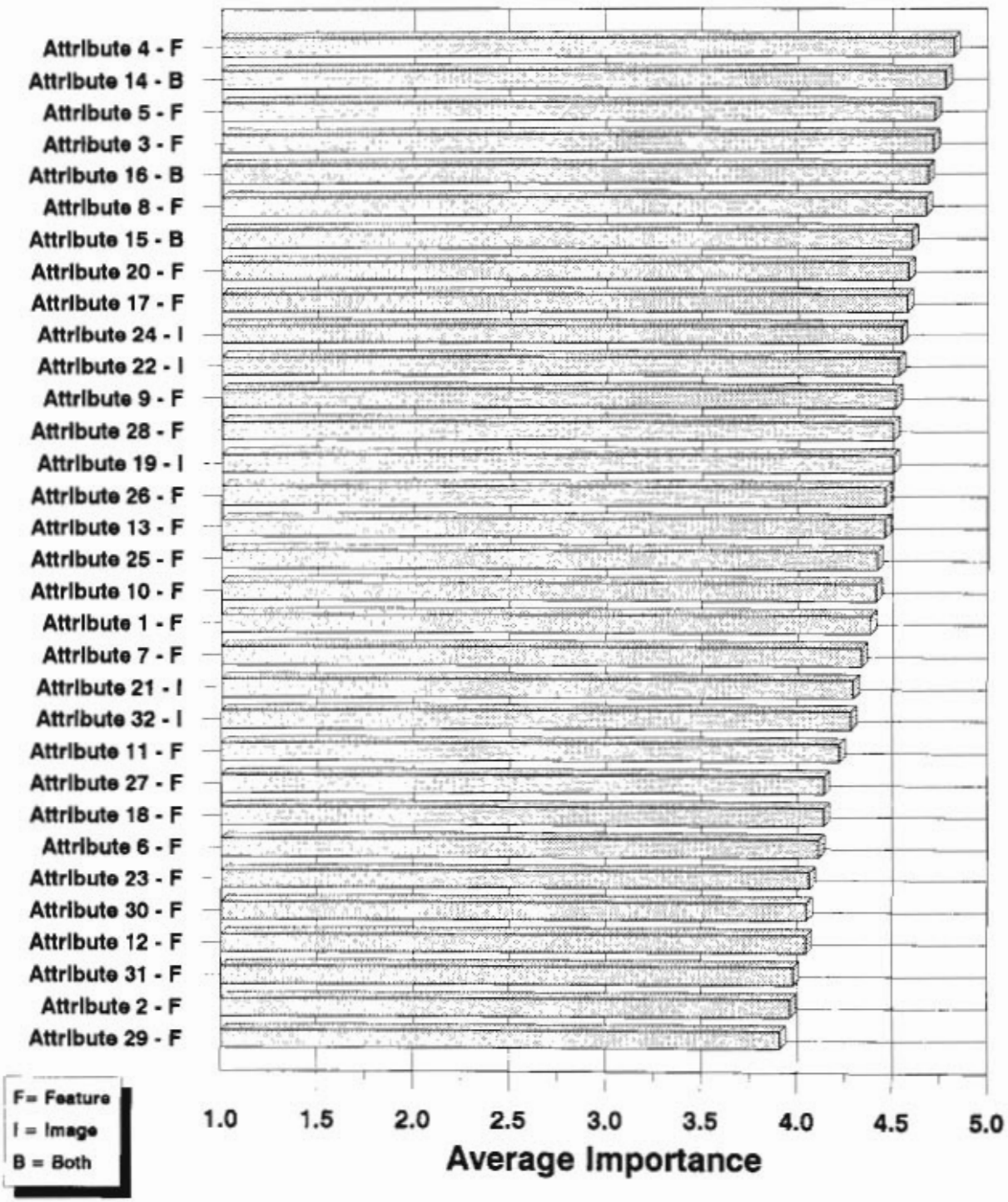


Figure 3
Stated Attribute Importance
- Commodity Chemical -



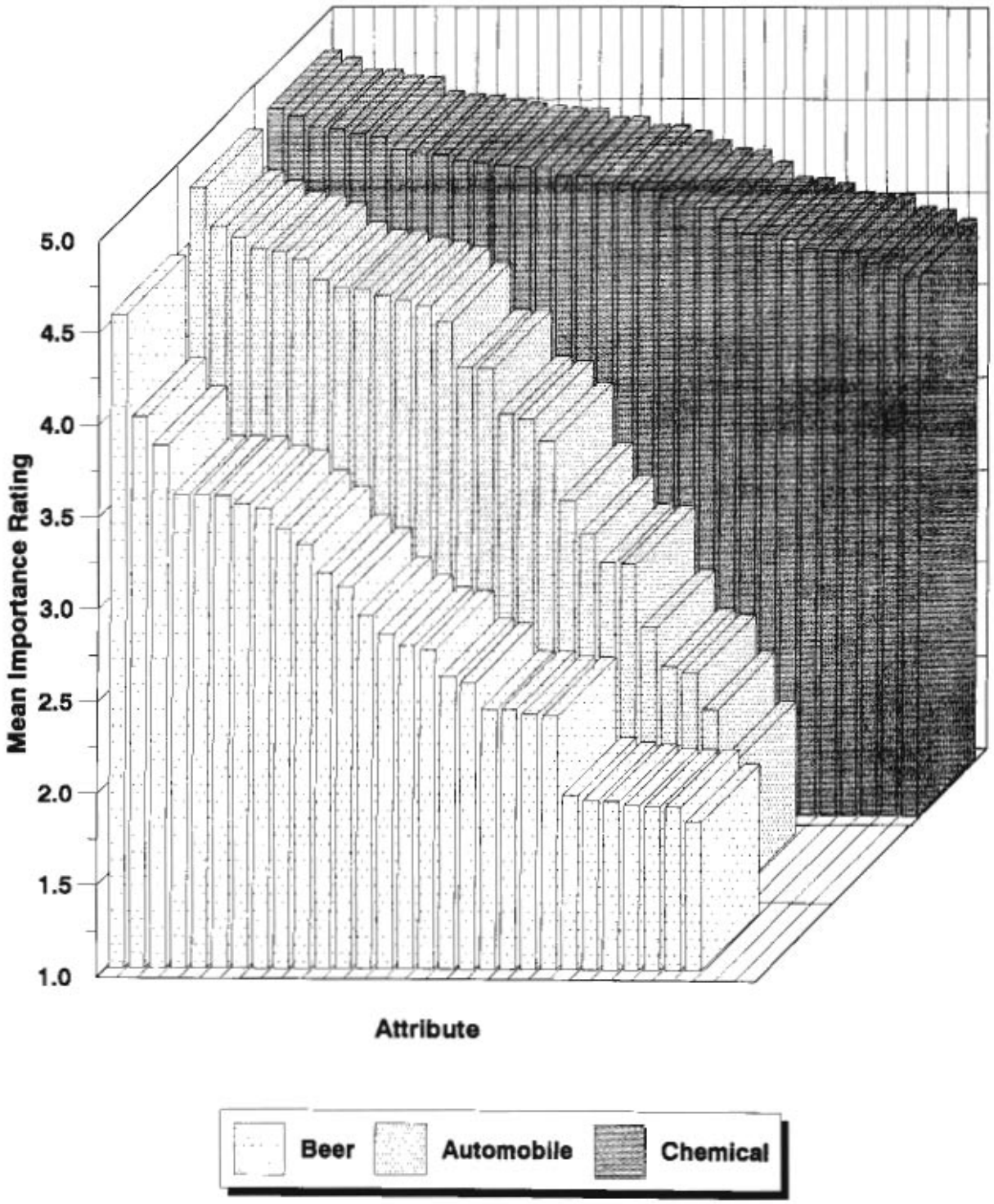
In each of the product categories, feature-based attributes garnered the highest mean importance ratings. For the Commodity Chemical study, this outcome should not be surprising given that the vast majority of attributes are, in fact, feature based.

More interesting, however, is the comparison of Beer and Automobile results. As can be seen, for Automobiles, the most important attributes on a stated basis tend to be descriptive of product features. While there are several attributes that could be characterized as both feature and image among the most important, these attributes tend to reflect economic/value issues. The most important image attribute, "Is fun to drive," is rated almost one full scale point lower in importance when compared to the most highly rated attribute: "Is Dependable."

For Beer, the attributes garnering the highest mean importance ratings are also feature-based. However, the rank-order positions of the image attributes are generally higher than in the Automobile data.

Finally, arraying the mean ratings for all three studies in the same chart (see Figure 4) leads to an interesting observation. The overall variability in mean stated importance ratings increases as the product category becomes increasingly image advertised. To wit, Beer, which is hypothesized as the most image driven of the three product categories, exhibits the greatest variability in mean importance ratings. The Chemical category, which is almost exclusively feature driven, presents very little variability in the mean ratings: every attribute is claimed to be extremely important.

Figure 4
Comparison of Stated Importance Ratings



Derived Importance

Correlation Analysis. As a first step, bivariate correlations were computed between the centered brand performance ratings and the relevant brand usage or satisfaction measures. Results of these analyses are presented in Figures 5, 6, and 7. In these charts, attributes are shown in order of stated importance.

Figure 5
Correlation of Performance Ratings with Automobile Ownership

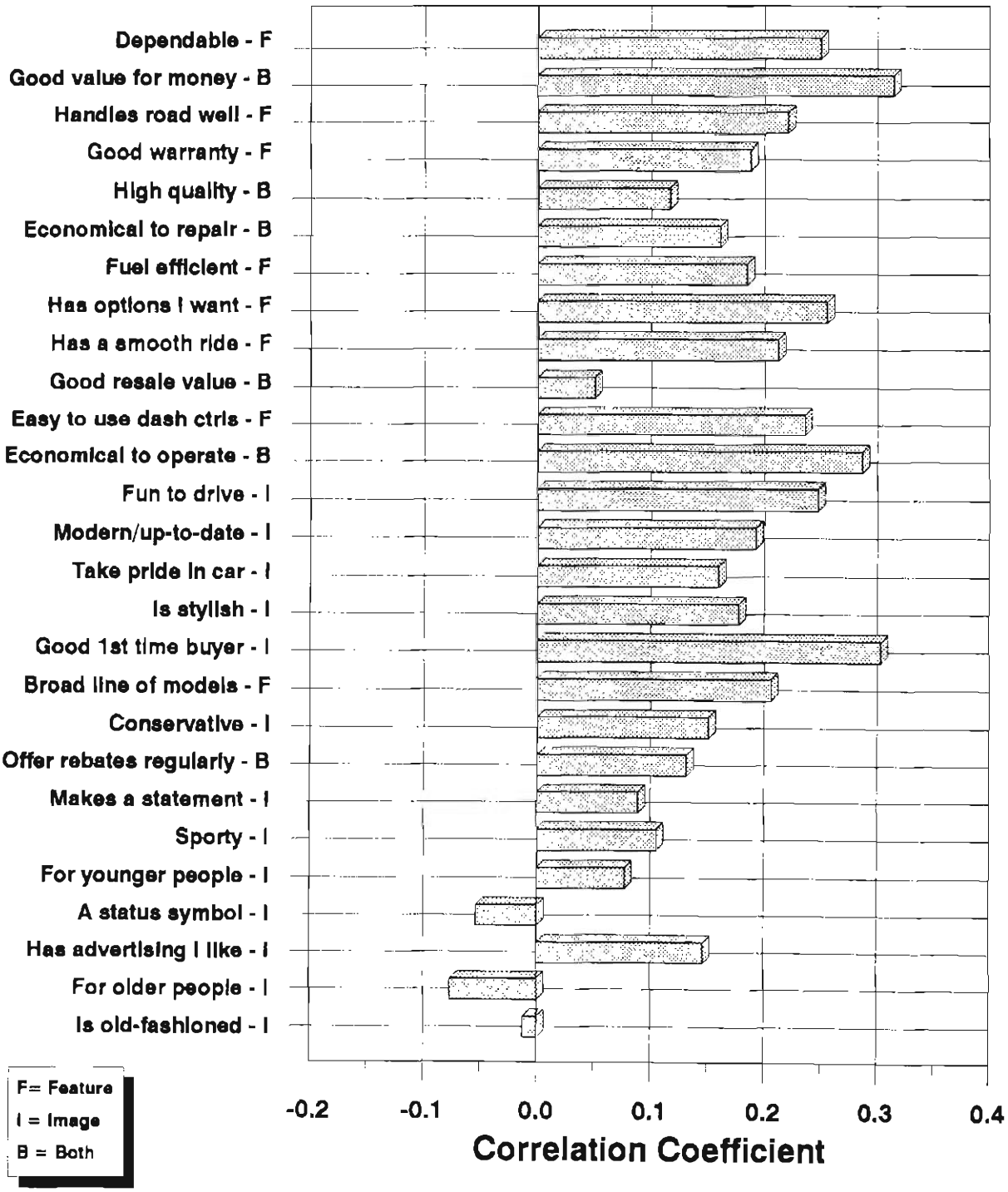


Figure 6
Correlation of Performance Ratings with Beer Brand Used Most Often

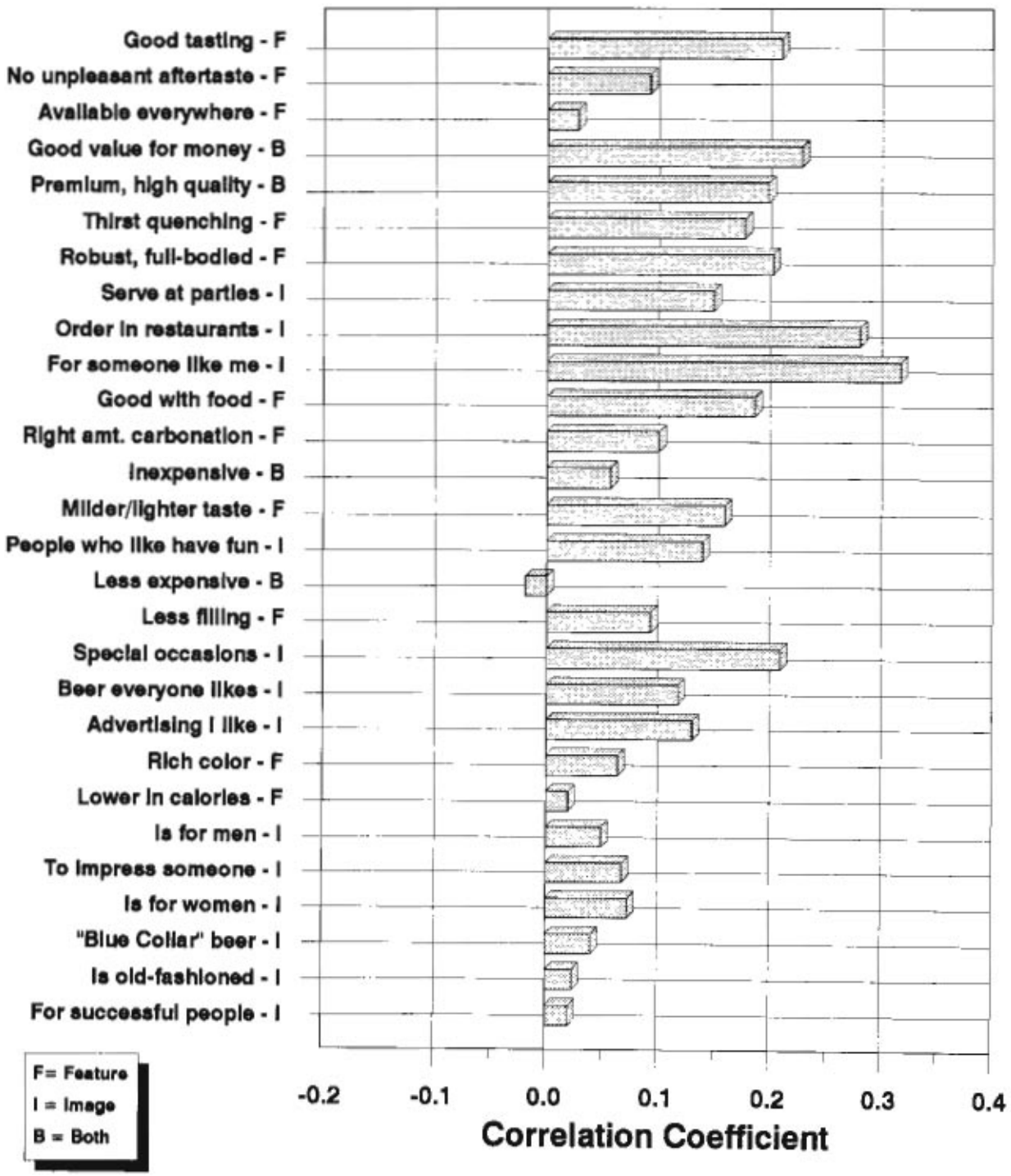
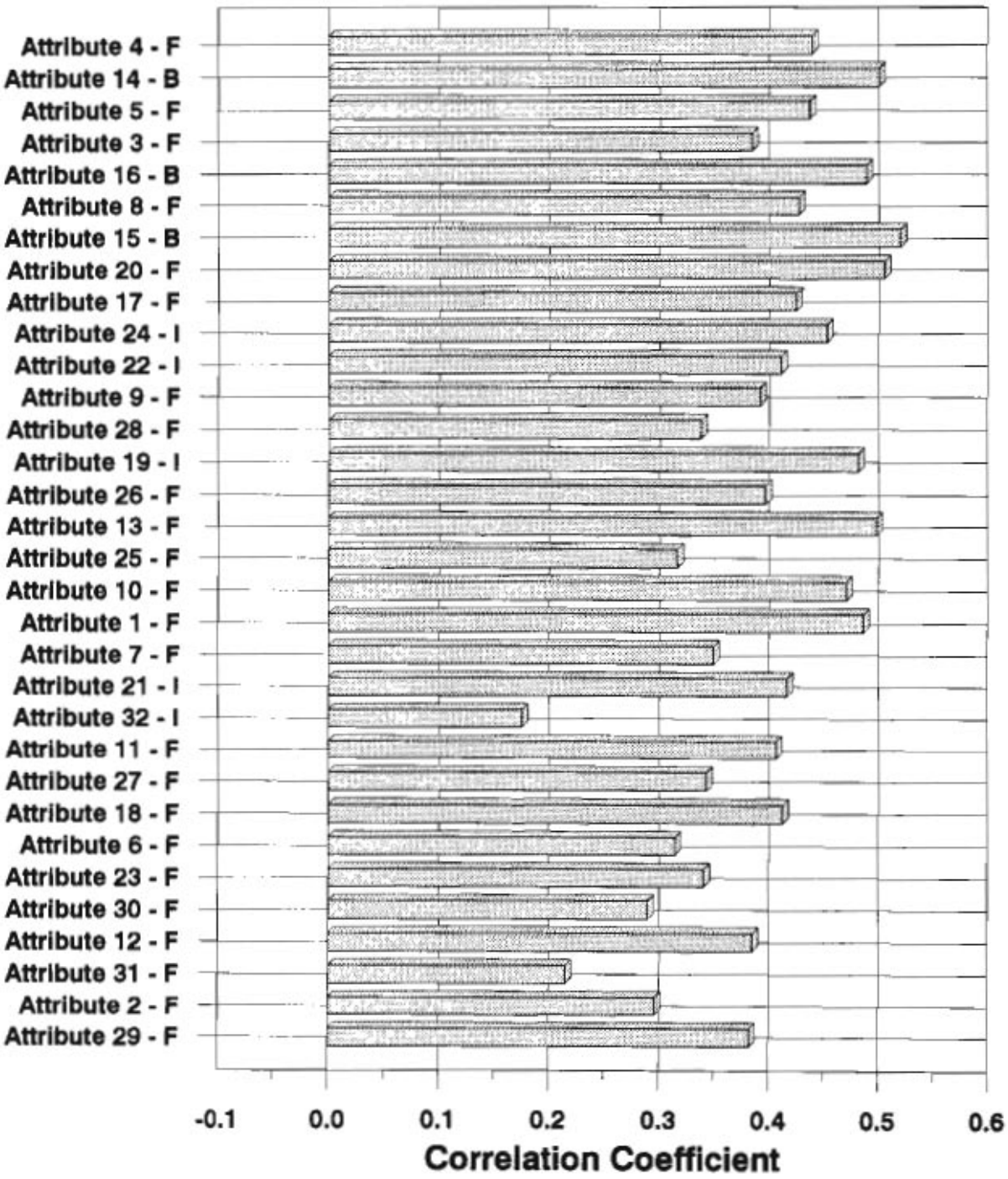


Figure 7
Correlation of Performance Ratings with Chemical Company Overall Satisfaction



For the Automobile data, the highest correlations between perceived brand performance and ownership occur on value/economic attributes (which have been classified as both feature and image). Other high correlations are observed on a mix of feature and image attributes:

- Is good for the 1st time car buyer
- Is dependable
- Has the options I want
- Has dash controls that are easy to use while driving
- Is fun to drive

Although the highest correlations on non-value related attributes are also noted for a mix of image and feature attributes in the Beer data, image attributes are more predominant than they are in the Automobile data:

- Is for someone like me
- Is the beer I'd order in restaurants
- Is good tasting
- Is for special occasions
- Has a robust, full-bodied taste

Because the attributes in the Chemical data are, for the most part, feature oriented, it is not surprising that they are also the attributes most highly correlated with overall supplier satisfaction.

Meaningful differences do occur between the data sets in both the magnitude and the variability of the bivariate correlations. For the Commodity Chemical data (Figure 7), virtually all correlations are .30, or higher, and average .41. For the Automobile data, the average correlation is .19; for Beer it is .15.

Regression-Based Analyses. Extensions of the simple bivariate derived analysis demonstrate the dangers inherent in doing derived importance analyses. Results typically do not do a very good job of explaining model variation.

First, the three data sets were reduced using principal components analyses and a varimax rotation. Solutions were based on a minimum eigenvalue of 1.0 criterion. The actual groupings of attributes by component are shown in Tables 3, 4, and 5. (Attributes are shown where loadings were .40 or greater.) Below is an overview of the number and nature of the solutions.

	Automobile	Beer	Chemical
Total Components	6	9	6
Mostly Feature	0	2	4
Both	4	3	2
Mostly Image	2	4	0
Variance Explained	54%	60%	65%

Table 3

Automobile Principal Components

Component 1 — The Car I Want

I Is stylish
I Is modern and up-to-date
F Has a smooth ride
I Is available with the options I want
I Is for people who take pride in their car
F Has dash controls that are easy to use while driving
B Is a high quality automobile
F Handles the road well
F Is dependable
I Is fun to drive
F Offers a good warranty program

Component 2 — Value/Economy

I Is good for a first time car buyer
B Is economical to operate
B Is economical to repair
B Is a good value for the money
F Is fuel efficient
I Is conservative

Component 3 — Long Term Quality

B Has good resale value
B Is a high quality automobile

Component 4 — For Older People

I Is a car for older people
I Is old-fashioned
I Is a status symbol

Component 5 — For Younger People

I Is a car for younger people
I Is sporty
I Makes a statement about my lifestyle

Component 6 — Advertising/Product Line

I Has advertising that I like
B Offers rebates on a regular basis
F Has a broad line of models from which to choose

F = FEATURE, I = IMAGE, B = BOTH

Table 4

Beer Principal Components

Component 1 — The Beer I Want
F Is good with food
F Is thirst-quenching
B Is a good value for the money
I Is for someone like me
F Is good tasting

Component 2 — Everyone's Beer
I Is a brand everyone likes
I Is good to serve at parties
F Is available everywhere beer is sold
B Is a premium, high quality beer

Component 3 — Appearance/Body
F Has a rich color
F Has a robust, full-bodied taste
F Has the right amount of carbonation
I Is the beer I'd order in restaurants

Component 4 — Inexpensive
B Is inexpensive
B Is less expensive than most other beers
I Is for special occasions (negative loading)

Component 5 — "Less Filling"
F Is a less filling beer
F Has a milder, lighter taste
F Is lower in calories

Component 6 — "Wouldn't It Be Great ..."
I Is for people who like to have fun
I Is a "blue collar" beer
I Has advertising that I like

Component 7 — YUPPIE Beer
I Is for successful people
I Is the beer to buy when you want to impress someone
F Doesn't have an unpleasant aftertaste

Component 8 — Gender
I Is for men
I Is for women

Component 9 — Old Fashioned Beer
I Is an "old fashioned" brand

F = FEATURE, I = IMAGE, B = BOTH

Table 5

Chemical Principal Components

Component 1 — Product Knowledge
F Attribute 23
I Attribute 24
F Attribute 26
F Attribute 28
F Attribute 25
I Attribute 19
F Attribute 27
I Attribute 22
F Attribute 29

Component 2 — Routine Services
F Attribute 9
F Attribute 10
F Attribute 11
F Attribute 13
F Attribute 8
F Attribute 12
F Attribute 17

Component 3 — Pricing Policies
B Attribute 16
B Attribute 14
B Attribute 15
F Attribute 20
I Attribute 21
F Attribute 7

Component 4 — Product Line/Quality
F Attribute 5
F Attribute 4
F Attribute 3
F Attribute 1

Component 5 — Leadership
F Attribute 31
I Attribute 32
F Attribute 30

Component 6 — Special Services
F Attribute 6
F Attribute 2
F Attribute 18

F = FEATURE, I = IMAGE, B = BOTH

Next, stepwise discriminant analyses were conducted in an effort to model brand ownership/ satisfaction as a function of the principal components scores. Results are discussed for each product category.

Automobiles

Of the six Automobile components, five were significant predictors of car ownership. The total variation in ownership explained by the model was a scant 16% (adjusted for degrees of freedom). Listed below are the correlations of the brand performance components with ownership and the proportion of variation explained by each.

Component	Correlation with Brand Ownership	Proportion of Explained Variance
The Car I Want (B)*	.24	34%
Value/Economy (B)	.28	47%
Long Term Quality (B)	-.02	--
For Older People (I)	-.07	3%
For Younger People (I)	.09	5%
Advertising/Product Line (B)	.14	11%

*F=Mostly Feature, I=Mostly Image, B=Both

Beer

Of the nine Beer components, seven were significant predictors of Beer brand used most often. The total variation in usage explained by the model was only 12% (adjusted for degrees of freedom). Listed below are the correlations of the brand performance components with brand usage and the proportion of variation explained by each.

Component	Correlation with Brand Usage	Proportion of Explained Variance
The Beer I Want (B)*	.24	41%
Everyone's Beer (B)	.11	9%
Appearance/Body (F)	.11	9%
Inexpensive (B)	-.05	--
"Less Filling" (F)	.12	10%
"Wouldn't It Be Great" (I)	.16	20%
YUPPIE Beer (I)	-.02	8%
Gender (I)	.10	--
Old Fashioned (I)	.07	3%

*F=Mostly Feature, I=Mostly Image, B=Both

Chemical

Of the six Commodity Chemical components, five were significant predictors of overall satisfaction with Chemical supplier. The total variation in usage explained by the model was a highly significant 46% (adjusted for degrees of freedom). Listed below are the correlations of the brand performance components with satisfaction and the proportion of variation explained by each.

Component	Correlation with Satisfaction	Proportion of Explained Variance
Product Knowledge (B)	.27	16%
Routine Services (F)	.32	22%
Pricing Policies (B)	.41	36%
Product Line/Quality (F)	.33	23%
Leadership (F)	.05	--
Special Services (F)	.09	3%

*F=Mostly Feature, I=Mostly Image, B=Both

Derived versus Stated Importance

A comparison of the stated attribute importance ratings and the derived bivariate correlations is presented in Figures 9, 10, and 11. For the three data sets, the Spearman rank-order correlations between ranks based on the stated importance ratings and ranks based on the bivariate correlations analyses are high and equivalent:

	Spearman r
Commodity Chemical	.66
Automobiles	.64
Beer	.61

Although this is a somewhat contrived measure of fit, the Spearman analysis does indicate that, generally speaking, rank-ordered stated importance measures correspond quite well to rank-ordered derived bivariate correlations in all three categories.

Given the high levels of statistical equivalence that exist in both the stated importance measures and the bivariate correlations, the strength of these relationships appears even more robust.

The results of the analyses presented here suggest the following:

- As products become increasingly feature advertised, stated importance ratings, on an absolute basis, are generally higher than they are for more image advertised products.
- There is less variability in stated importance ratings for feature advertised products when compared to image advertised products.

- Bivariate correlations between product performance ratings and satisfaction, as measures of derived importance, are higher for feature advertised products.

- Models of derived importance explain more variance in the criterion measure for feature advertised products (although all models leave much of the variance unexplained).

The derived importance analyses reported here represent only a few of the many ways that derived analyses are conducted. Until there is more compelling evidence that derived measures do a better job explaining variation in product usage than do stated measures, there is little to suggest that derived measures produce better insight into reasons for purchase.

As such, it is recommended that we be more inclined to believe respondents when they tell us what is important to them. Surely if we are willing to believe their perceptions of brand performance, we can also believe that they will reliably report their perceptions of attribute importance.

Figure 8
Comparison of Bivariate Correlations

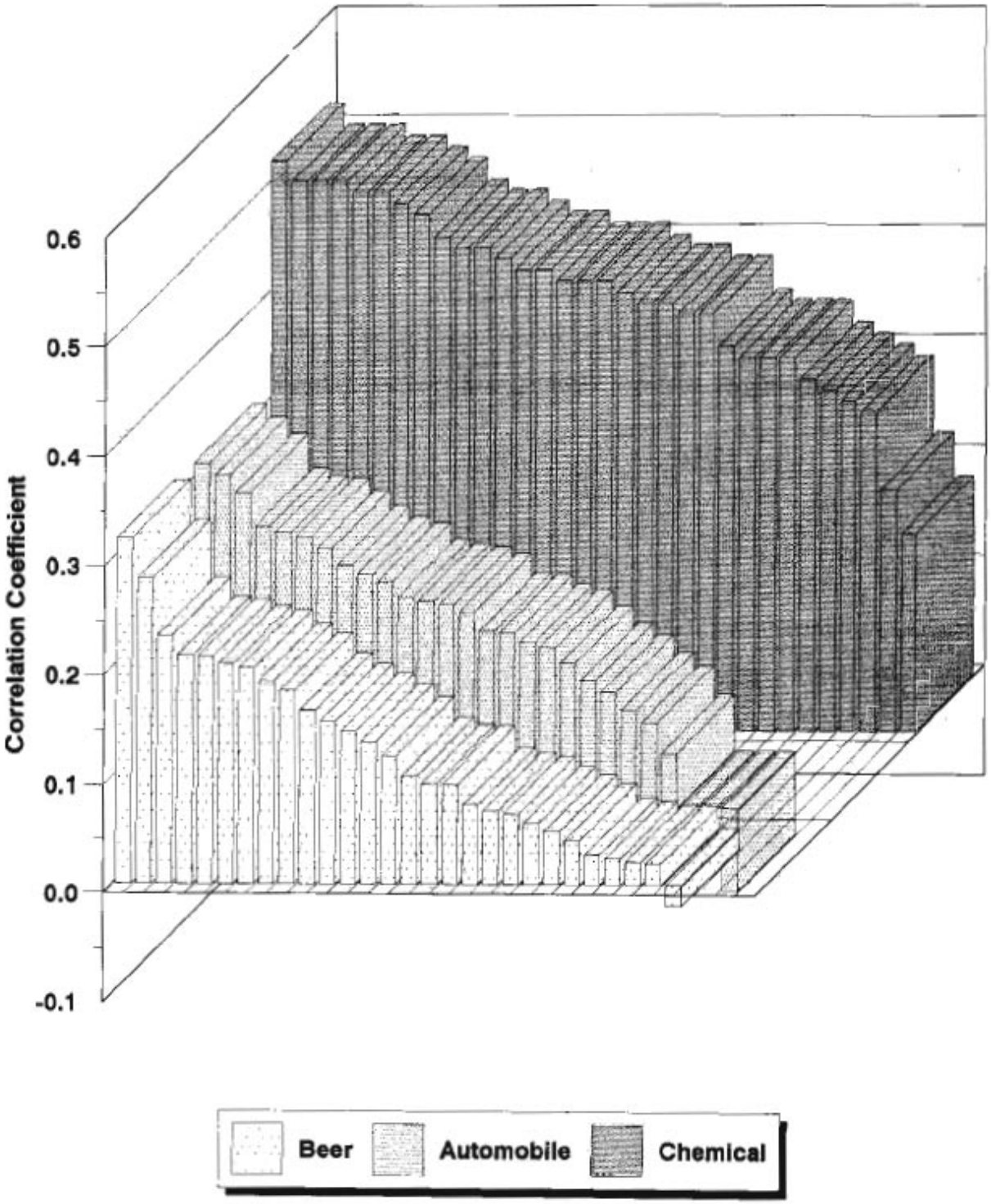


Figure 9
Stated versus Derived Importance
- Automobiles -

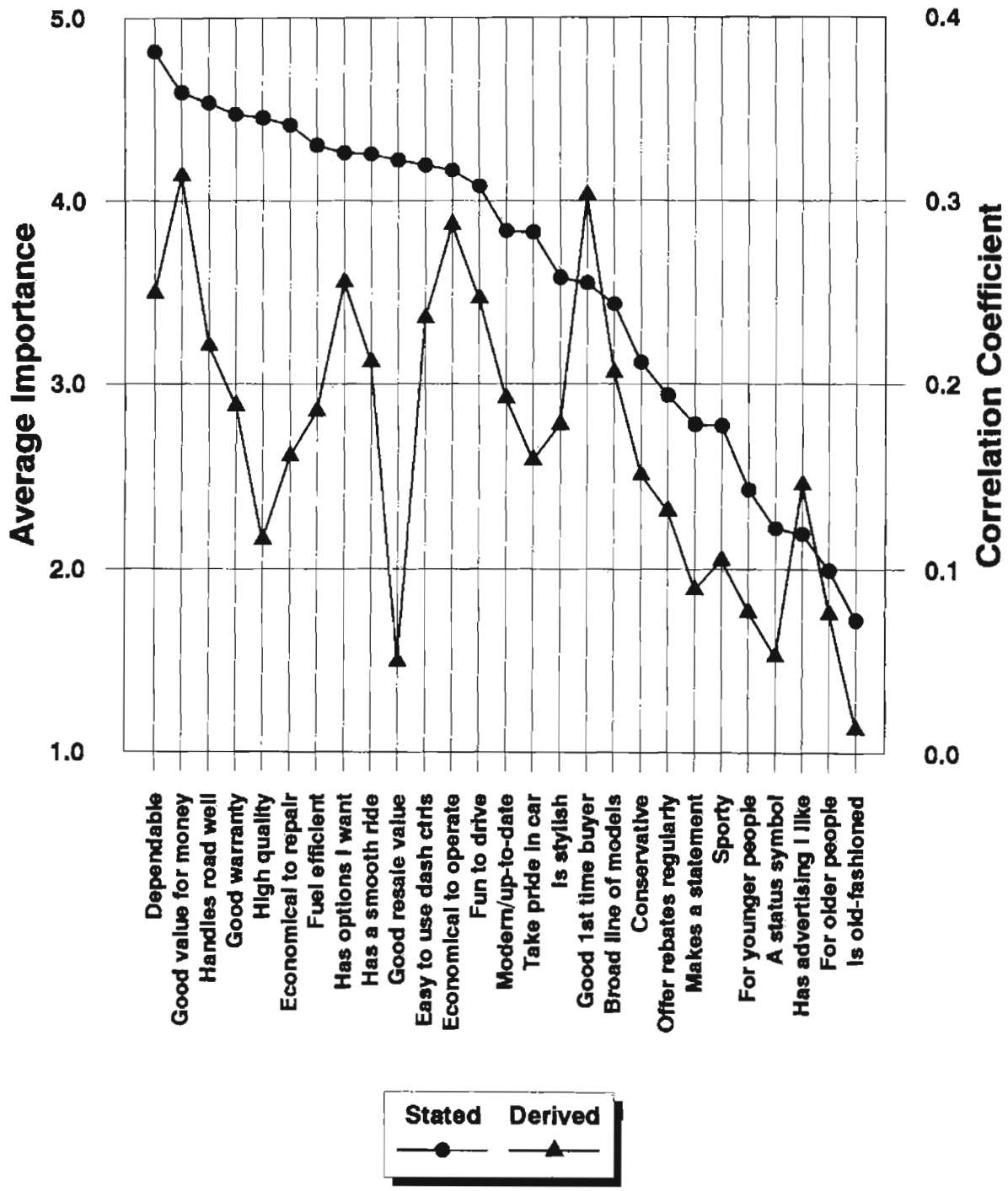


Figure 10
Stated versus Derived Importance
- Beer -

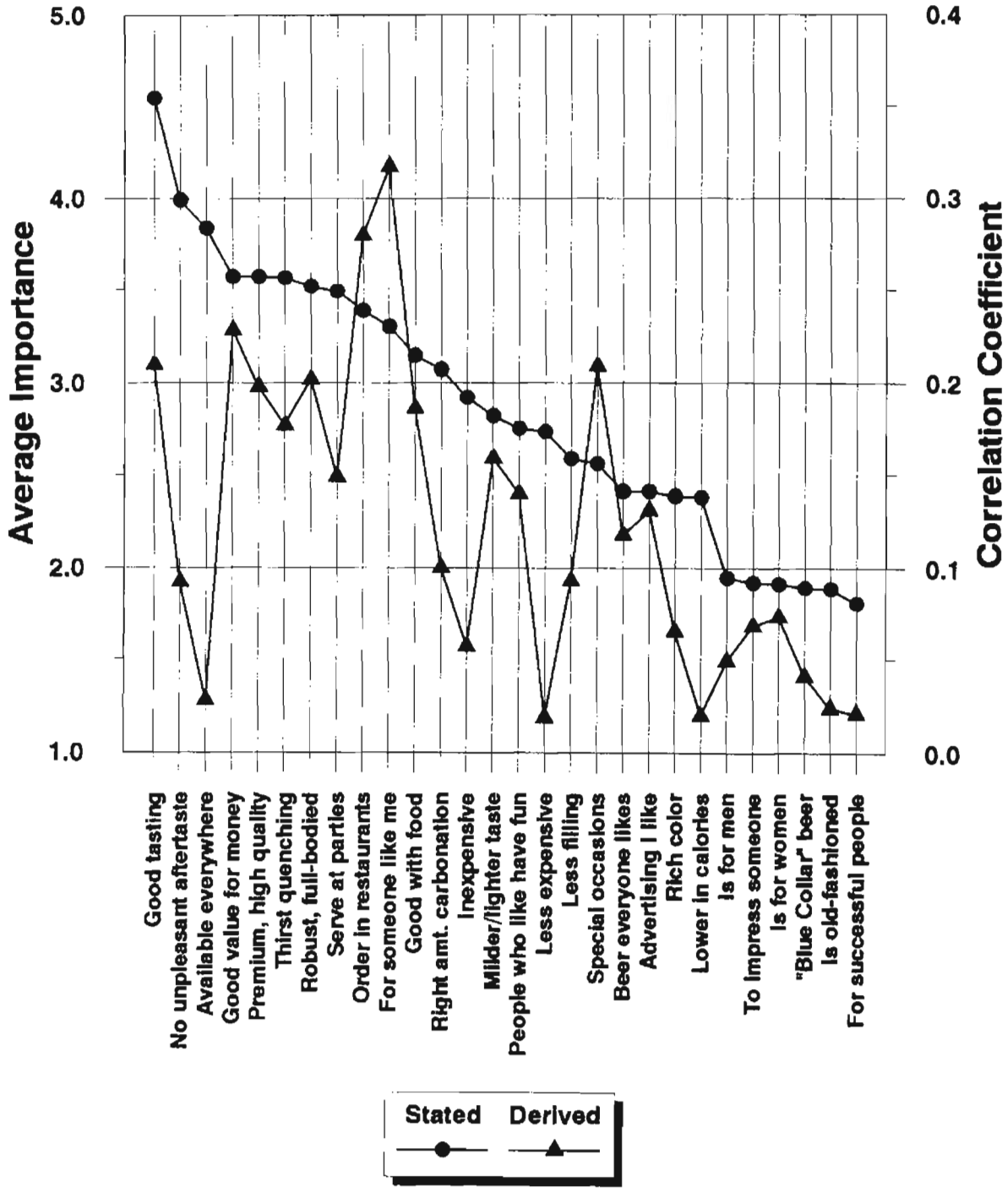
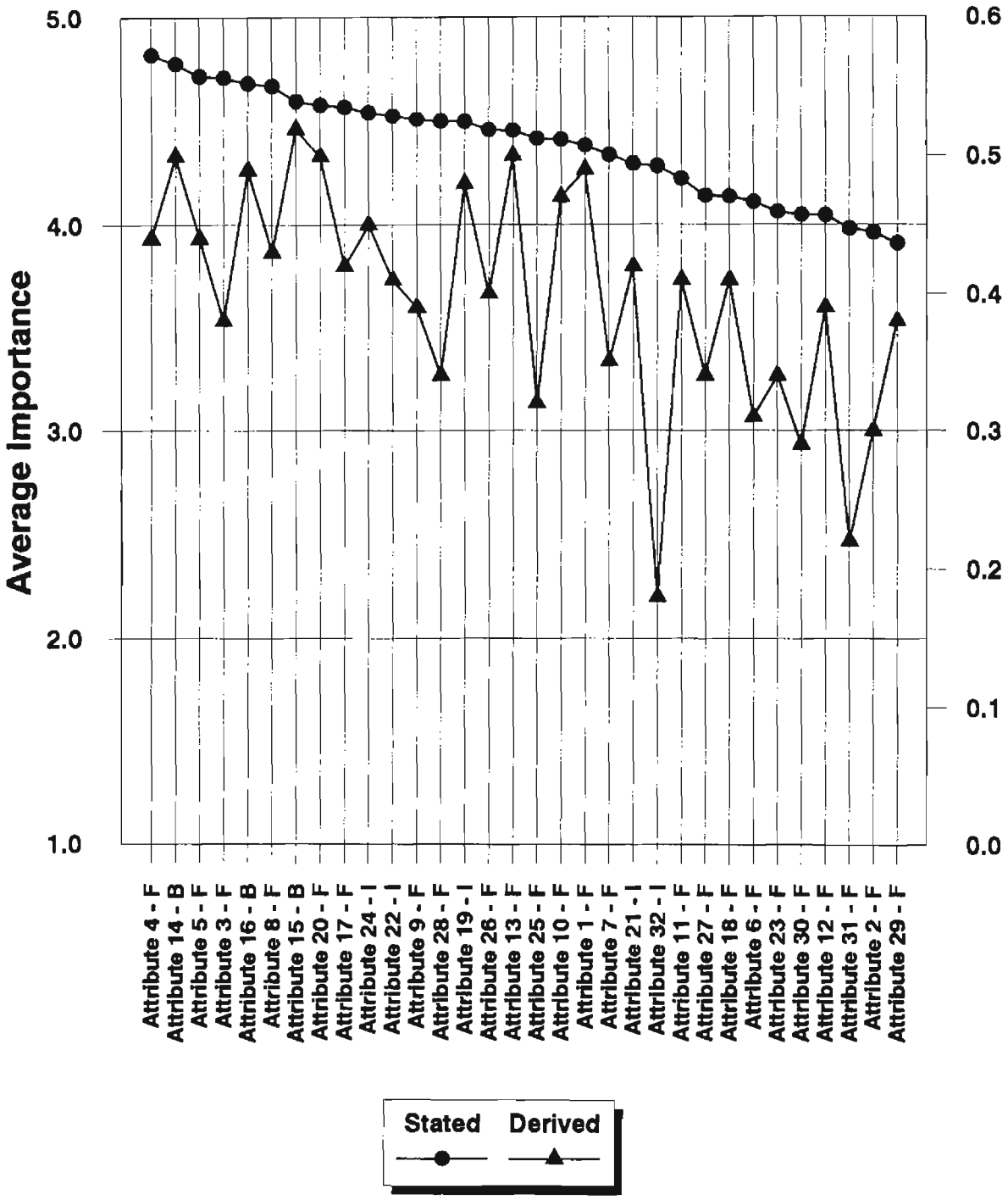


Figure 11
Stated versus Derived Importance
- Commodity Chemical -



Comment on McLauchlan

Joel Huber
Duke University

In his paper, "The Predictive Validity of Derived versus Stated Importance," Bill McLauchlan examines both two ways to measure attribute importance across three product classes. He uses impressive data displays to graphically illustrate the differences between direct and derived measures of attribute importances. His focus is on the difference in the two techniques when applied to feature versus image advertised products. The results given are clear and would very likely replicate in other contexts. McLauchlan closes by noting that given the rough equivalence of the measures, the more easily accessed direct measure should be used "until there is more compelling evidence that derived measures do a better job explaining variation in product usage."

In commenting on this paper, I will consider differences among derived and stated importance measures to propose that both have difficulties that limit their usefulness in practice. Figures 9 and 10 make a fine a graphical comparison of the two importance measures for automobiles and beer categories. The table below identifies attributes whose ranking by one criterion is quite different than by the other. Examining these incongruent attributes provides insight into how the two importance measures differ and some guidance on their relative strengths and weaknesses.

TABLE
ATTRIBUTES WHICH ARE
DIFFERENTIALLY IMPORTANT IF

	STATED		DERIVED	
		(Stated rank/derived rank)		
AUTOS:				
	Warranty	(4/11)	Options desired	(8/4)
	Resale value	(10/25)	First time car	(17/7)
	Dependable	(1/5)	Broad line	(18/9)
BEER:				
	Available	(3/24)	Someone like me	(10/1)
	Inexpensive	(15/25)	Special occasions	(19/5)
	Unpleasant aftertaste	(2/16)	Order in restaurants	(9/2)

In examining the table above, three generalizations are apparent.

1. **Stated importances emphasize normatively desirable features.**
Resale value, warranties and dependability loom large for stated importances, while actual purchases are better reflected in the car having the desired options, style (broad line), and the price of a first time car. As these data indicate, when consumers are directly asked what is important, they

respond more positively to attributes that they feel they ought to believe are important, despite the fact that these attributes may not be considered deeply when making a choice.

2. **Stated importances emphasize low probability problems, while derived importances will be very insignificant unless the attribute exhibits substantial variability.**

Most popular beers are available, so the importance of availability reflects one's reaction to finding a favorite beer out of stock. The same reaction is arguably true for aftertaste. Most beers have an acceptable aftertaste, but it is easy to remember a stale or bitter beer. These are low probability events; they do not happen often, but are important if they do. Thus the managerial message from an attribute with high stated and low derived importance is conditional: the attribute may not now determine choices; however, it could do so if distribution fails to make the product available, if operations makes a bad batch, or if marketing prices it too high.

In keeping with this idea, the derived importances of the attributes on the left column of the table are very low precisely because they have so little variation in the environment. Derived importances, reflecting the correlation between preference and the attribute ratings, depend critically on the variability of the attribute in the market. The problem is not that the relationship does not exist, but that we cannot detect it given the limited variability in the attribute. To see that, suppose initially one had a scatter plot indicating a strong negative relationship between brand preference and bitterness. Now consider what happens if companies with bitter beer go out of business or mend their ways by producing beer that is no longer bitter. Once that happens, the resulting scatter will show almost no relationship between preference and bitterness, not because bitterness is irrelevant, but because there is not enough range on the bitterness to exhibit a significant impact on preference.

Thus, correlational measures can be quite misleading in that they assume that the range of the attributes will remain constant. Stated importance measures, by contrast, implicitly ask respondents to indicate how important an attribute would be if it does change. This line of reasoning suggests that correlational measures are appropriate when the managerial action will preserve attribute expectations in the market, while stated importances are more appropriate when they will alter them.

3. **Causality is equivocal in derived importances**

The biggest problem with correlational measures of importance is that they only measure association, not causality. For example, given one likes a beer, that will tend to lead to its purchase in a restaurant or its use for special occasions. Thus, preference leads to a high rating on the attribute, not the reverse. Of course, it is also possible that one first uses a beer in those contexts and then learns to like it, indicating that usage creates preference. The point is not to resolve these two interpretations but to understand the problem with trying to tease them apart. Many hours have been spent with path models and causal modeling (LISREL) to try to parse causal paths, but the most common outcome from such careful analysis is an acknowledgement of how difficult it is to derive causal relationships from cross sectional data.

Between its sensitivity to range effects and its ambiguous causal implications, it is very hard to draw useful managerial conclusions from derived importances. Correlational measures may be good for generating ideas, but can produce very misleading managerial recommendations. Further, a better way to generate ideas may be through qualitative research in the form of focus groups or in-depth interviews.

In summary, both stated and derived measures of attribute importance have their drawbacks. Stated measures tend to focus attention differentially on low probability problems and be influenced by social norms, while derived importances result in equivocal managerial recommendations. Both methods tend to be dominated by a different measure of derived importance — that produced from a conjoint task.

In a conjoint exercise, respondents trade off differences in one attribute against differences in another. Since these differences are explicit, there is no indeterminacy with respect to range. Further, since conjoint's measure of importance is a function of how much of one attribute one is willing to give up on others, it tends to be less affected by normative values. Finally, since the levels of the profiles are experimentally manipulated, the direction of causality can only flow from the attribute levels to the choice.

All of this suggests that the critical question may not be deciding between derived and stated importance, but determining the contexts in which either should be used at all.

USING ANALYSIS OF RESIDUALS AND LOGARITHMIC TRANSFORMATIONS TO IMPROVE REGRESSION MODELING OF BUSINESS SERVICE USAGE

Michael G. Mulhern
Mulhern Consulting
Douglas L. MacLachlan
University of Washington

INTRODUCTION

Regression analysis is a statistical technique widely used to evaluate the impact of one or more independent predictor variables on a single criterion or dependent variable of interest. As Green and Tull (1978, p. 303) note, regression is the prototype of single criterion, multiple predictor analysis.

All statistical models, regression or otherwise, rely on simplifying assumptions which, with any given data set, may or may not be valid. The purpose of this paper is to illustrate via a case study the corrective measures an analyst may take when one or more of the regression model assumptions is violated.

REGRESSION AND RESIDUALS: A REFRESHER

Regression analysis is a statistical tool for assessing the relationship between a single continuous dependent variable (Y) and one or more continuous independent variable(s) (X_1, X_2, X_3, . . . , X_p). Most often, it is used when the independent variables are not controllable (for example, in survey research or observational study).

It is typically used to address these questions:

1. Can we characterize the relationship or find a linear composite between the dependent and independent variables that compactly expresses the relationship between them?

2. Can we develop a quantitative formula to predict the relationship between a set of variables?

3. If so, how strong is the relationship and how well can it predict values of the independent variables?

4. Which independent variables are most important in explaining the variation in the dependent variable?

5. Can we describe the relationship between dependent and independent variables while controlling for other variables?

In its population parameter form, the simple linear regression model is written as:

$$Y_i = \text{alpha} + \text{beta}\ X_i + \text{epsilon}_i$$

where

Y_i = value of the dependent variable in the i^{th} trial,

alpha and beta are parameters specifying the functional form of the relationship between dependent and independent variables,

X_i = value of the independent variable in the i^{th} trial,

epsilon_i = population error, a random variable, normally distributed with mean = 0, having constant variance over different values of X_i, and uncorrelated for different trials.

Normality of the errors is required to compute probabilistic estimation intervals for the population parameters and the predicted values of Y, as well as to statistically test hypotheses about the magnitude of the population parameters (that is, alpha and beta).

However, since the population parameters are usually unknown, the simple linear regression model for estimating the population parameters from sample data is written as:

$$\text{Y-hat}_i = a + b\ X_i$$

where

Y-hat_i = value of the independent variable predicted by X_i (that is, fitted or predicted Y)
a = intercept (the value of Y-hat when X = 0)
b = slope (the change in Y-hat per unit change in X)
X_i = the i^{th} value of the independent variable.

To obtain the values of a and b (estimates of alpha and beta, respectively), a method called least squares is commonly used. In essence, least squares is a curve fitting procedure that finds the line that minimizes the sum of squared deviations of the observed values Y from the estimated values of Y on the regression line (the Y-hats).

Assumptions of ordinary least squares estimation of simple linear regression equations are typically as follows:

1. The regression function is linear.
2. The error terms are independent.
3. The error terms have constant variance over all values of the independent variable (that is, homoscedasticity).
4. Outliers are few in number.

This version of the regression model can be expanded by adding independent variables to improve the explanatory or predictive power of the model. In most marketing situations, it's clear that a single

variable such as income will not likely explain a large portion of the variance in another variable such as frequency of use. Consequently, the multiple regression model incorporates more than one independent variable and can be symbolically expressed as:

$$Y_i = \text{alpha} + \text{beta}_1 X_{1i} + \text{beta}_2 X_{2i} + \ldots + \text{beta}_p X_{pi} + \text{epsilon}_i$$

Once additional independent variables are introduced into the model, the problem of multicollinearity may arise. Multicollinearity is the correlation between independent variables included in the regression model. This implies that the regression coefficient for any independent variable is a function of the other independent variables. The regression coefficient reflects the marginal or partial impact on the dependent variable given the other correlated independent variables in the model.

Since formal experiments are not typically conducted to control for variation in the independent variables, we say that the findings of most regression analyses are associative rather than causative. In other words, we describe the relationship statistically in terms of its existence and strength.

In order to quantify our measure of strength, we must first consider potential sources of error in our model. This is accomplished by analyzing different types of variance. The analysis of variance approach is based upon partitioning the sums of squared deviations and degrees of freedom associated with the independent variable Y.

Conceptually, variation of an observation from the mean of all observations can be decomposed into that portion explained by the regression model and that portion unexplained by the model.

$$(Y_i - \text{Y-bar}) = (\text{Y-hat}_i - \text{Y-bar}) + (Y_i - \text{Y-hat}_i)$$

where

Y_i - Y-bar = Total deviation (from the mean Y-bar)

Y-hat_i - Y-bar = Deviation explained by the regression

Y_i - Y-hat_i = Deviation unexplained by the regression

Total variation of all the observations on Y in a data set is measured by the sum of squared deviations of Y_i from their mean, Y-bar. Similarly, the sum of squared deviations of the fitted value of Y, Y-hat_i, from the mean, Y-bar, is the variation explained by the regression. The final component, variation unexplained by the regression is the sum of squared differences between the observed values of Y_i and the fitted values, Y-hat_i. These differences are also called the residuals.

Mathematically, we have

$$\text{sum}(Y_i - \text{Y-bar})^2 = \text{sum}(\text{Y-hat}_i - \text{Y-bar})^2 + \text{sum}(Y_i - \text{Y-hat}_i)^2$$

SSTO = SSR + SSE

SSTO measures the variation in the observations when X is not considered. Similarly, SSE measures the variation in the Y_i when a regression model using X is employed. Thus, a measure

of the impact of X in reducing the variation in Y is

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

r^2 is interpreted as the proportionate reduction in total variation associated with the use of the independent variable X. Thus, the larger the r^2 , the more of the total variation in Y is explained by introducing the independent variable X. r^2 measures the strength of the linear relationship in that it measures the reduction in the sum of squares of vertical deviations obtained by using the least squares regression line relative to using the naive model where X is ignored and the sample mean of Y is used to predict Y. When more than one independent variable is included in the equation, the measure extends directly to the coefficient of multiple determination, R^2. The latter must be adjusted by a factor to account for the fact that additional independent variables will always inflate R^2. Hence, adjusted R^2 is used for comparison of fit of equations containing different numbers of independent variables.

While R^2 is useful in evaluating the strength of the linear relationship between X and Y, it is of little help in testing for the aptness of the model (that is, whether the assumptions are violated). This is due to the fact that many different distributions of X and Y can generate similar deviations and, therefore, R^2. Furthermore, R^2 is strongly affected by outliers, especially in small samples.

DETERMINING MODEL QUALITY AND THE ROLE OF RESIDUAL ANALYSIS

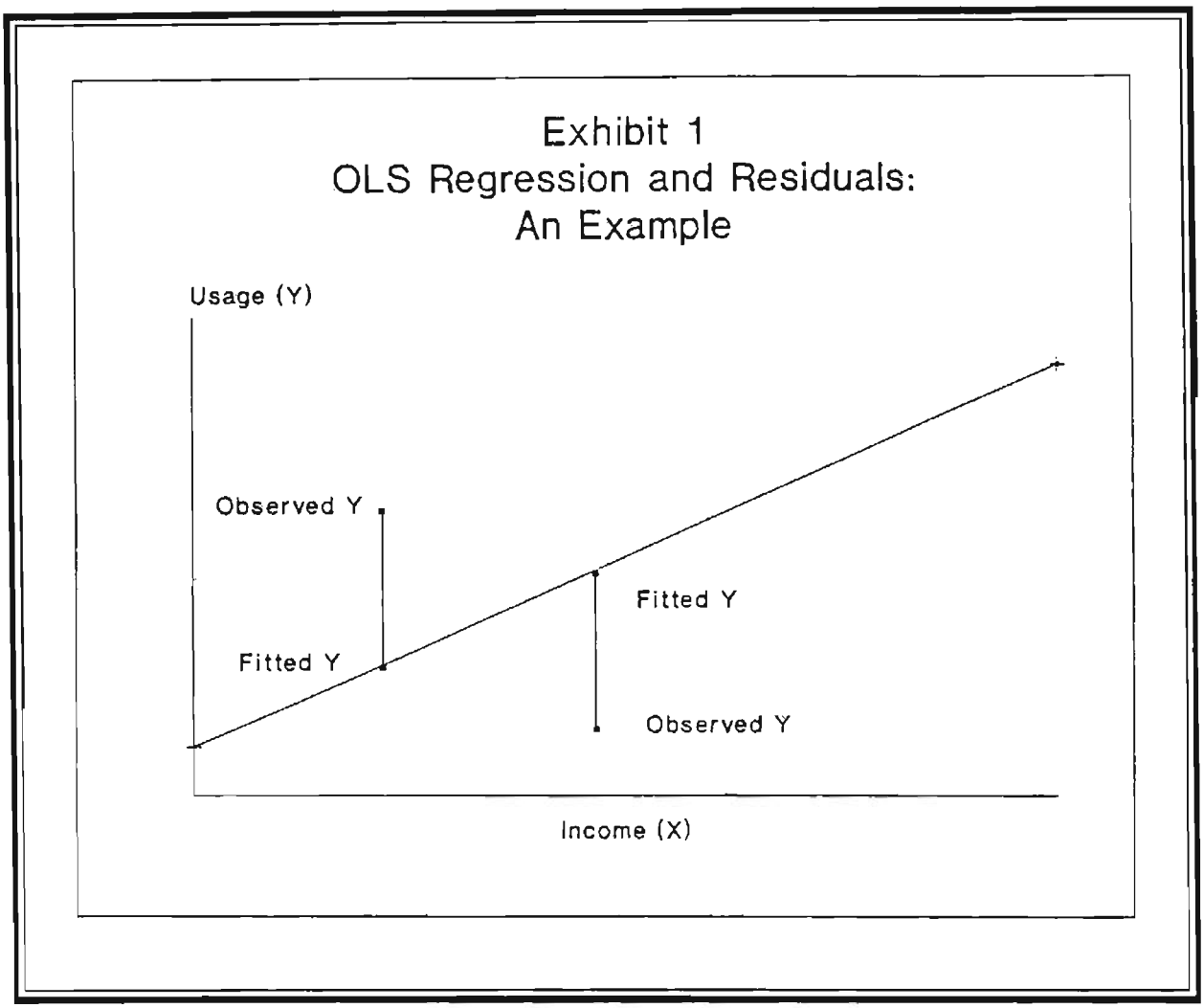
Model quality is often assessed via two yardsticks. The first, coefficient of determination or R^2, is a statistical measure of the proportion of variation in the dependent variable accounted for by the variation in the independent variables included in the model. As noted above, this statistic does not address completely how well the regression model fits the data set; that is, the aptness of the model.

APTNESS OF THE MODEL AND RESIDUAL ANALYSIS

The residual, e_i, is the difference between the observed value of Y (Y_i) and the fitted value of Y ($\hat{Y}_i$)

$$e_i = Y_i - \hat{Y}_i$$

and can be thought of as empirical realization of the population error (that is, the random variable epsilon $_i$). This is graphically portrayed in Exhibit 1.



As can be seen from the equation, residual analysis focuses on the dependent variable Y. Residual analysis consists of a set of diagnostic procedures for the dependent variable that are particularly useful in testing for the aptness of the least squares regression model. In essence, analysis of residuals gives insight into the assumed properties of the population error (for example, normal versus nonnormal distribution and constant variance over all values of the independent variables) and the functional form of the equation (for example, linear versus nonlinear).

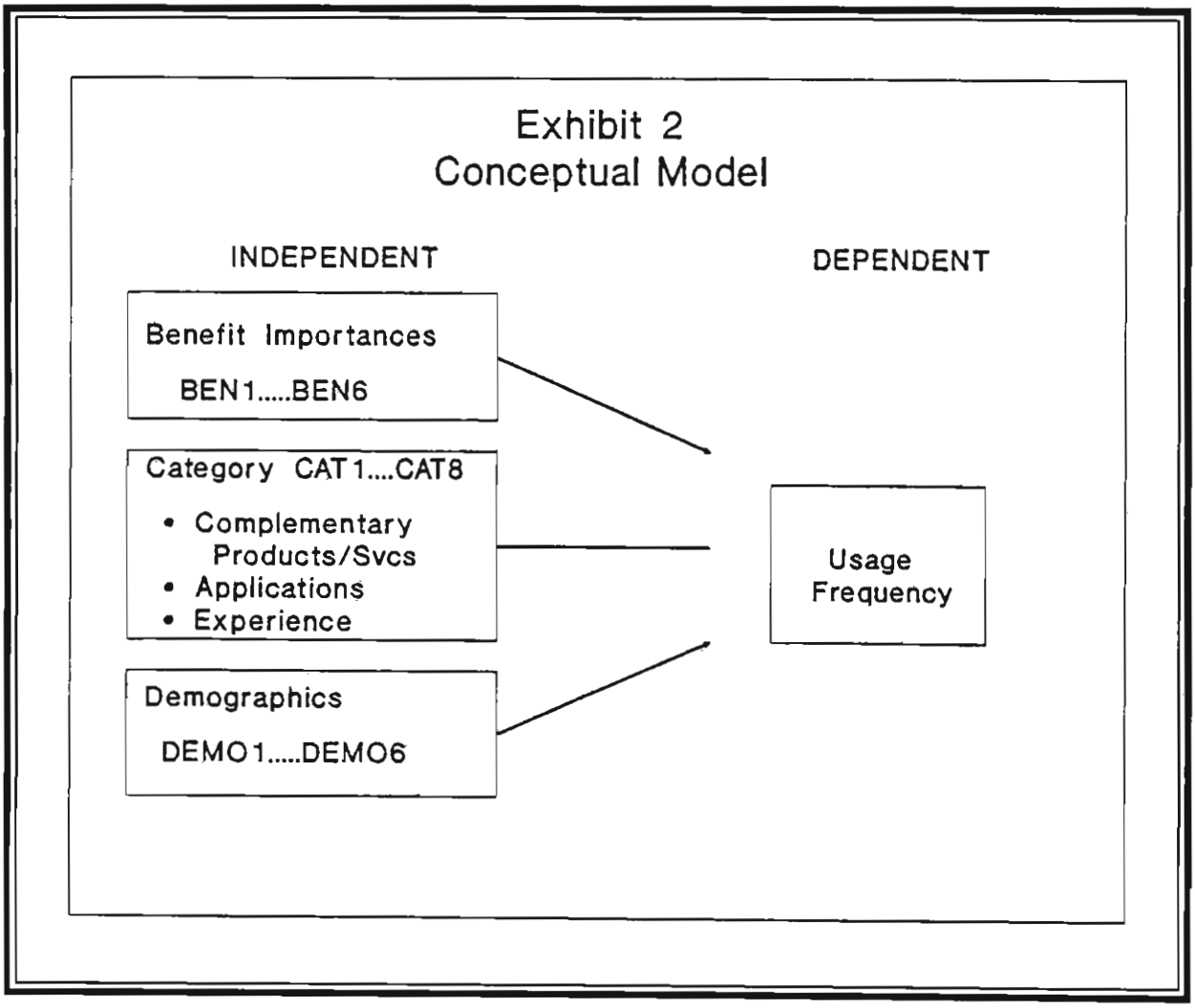
Either a graphical or statistical approach to residual analysis can be taken. In our case, the graphical approach was sufficient to diagnose and solve our problem. However, statistical tests are available and are discussed in Neter *et al.*, 1990, 130-140.

CASE STUDY: RESEARCH TASK, DATA, APPROACH

One of the authors was engaged in a consulting assignment where two primary issues confronted management. The first was to define and describe pricing segments and the second was to

determine the factors that had the greatest influence on usage. For the second issue, the project goal was to develop a model that determined the factors that were the greatest contributors to explaining the variation in usage of the client's service.

The client was a large business services firm headquartered in the Western United States. Survey and usage data from 160 respondents were analyzed. Since interval scaled data were available for both the dependent and independent variables, ordinary least squares regression was the analytical technique chosen. The dependent variable was frequency of usage and 20 independent variables were included in the model. The independent variables were represented by three general groups of marketing variables: 1) benefits, 2) category related variables which included application, length of usage/experience, and complementary products used, and 3) demographics. Exhibit 2 presents these variables. Several independent variables were nominally scaled and were treated as dummy variables in the regression model. An additive, linear model was assumed and ordinary least squares (OLS) regression estimation was employed to estimate its parameters. SPSS-PC+ was the analytical software used (Norusis, 1986).



MODEL BUILDING PROCESS

The process we followed to build an appropriate model incorporated these steps:

1. Assume that a straight line (or hyperplane in p dimensions) would fit the data.
2. Find the best fitting straight line (or hyperplane).
3. Determine the strength of the association among the variables.
4. Determine if the data set met the OLS regression model assumptions. If not, try various transformations to eliminate or minimize violations.

PROBLEMS IDENTIFIED IN THE INITIAL ANALYSIS

As noted above, a graphical analysis of the residuals identified several violations of the regression assumptions. A histogram of residuals and a normal probability plot of cumulative standardized residuals versus cumulative standard normal distribution both indicate the nonnormality of the error terms. (See Exhibits 3 and 4.) The latter shows the deviation of the cumulative residuals from that predicted by the normal curve as points falling off/diverging from the diagonal straight line. Characteristic curves emerge from skewed (nonsymmetric) distributions and distributions with tails too fat (platykurtic) or too thin (lėptokurtic), both forms of kurtosis. Exhibit 3 shows that the distribution of residuals is both skewed to higher values and too peaked to be normally distributed.

Exhibit 3
Histogram of Residuals: Before Transformation

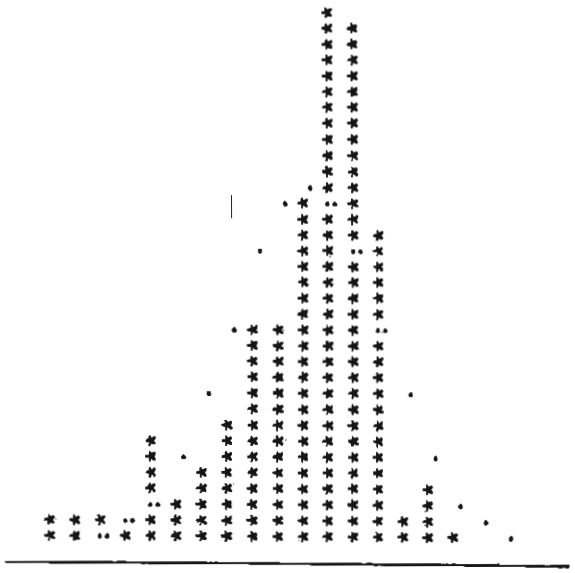
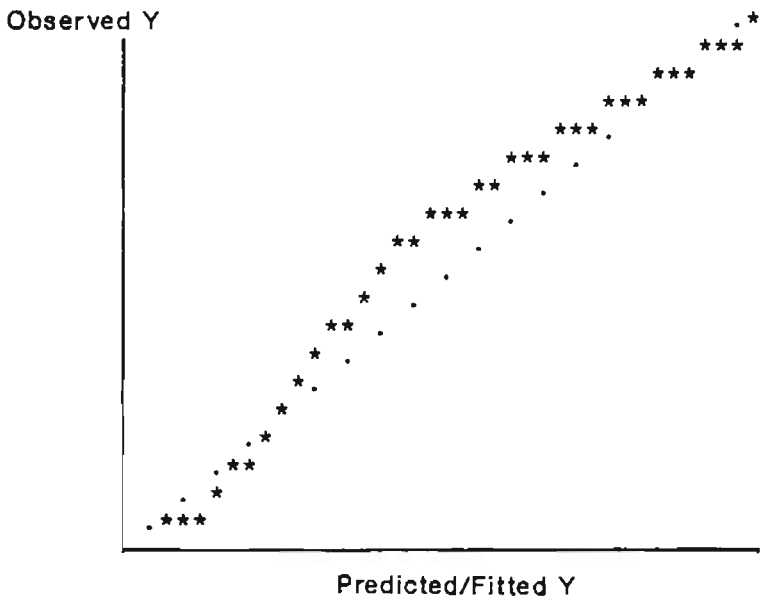
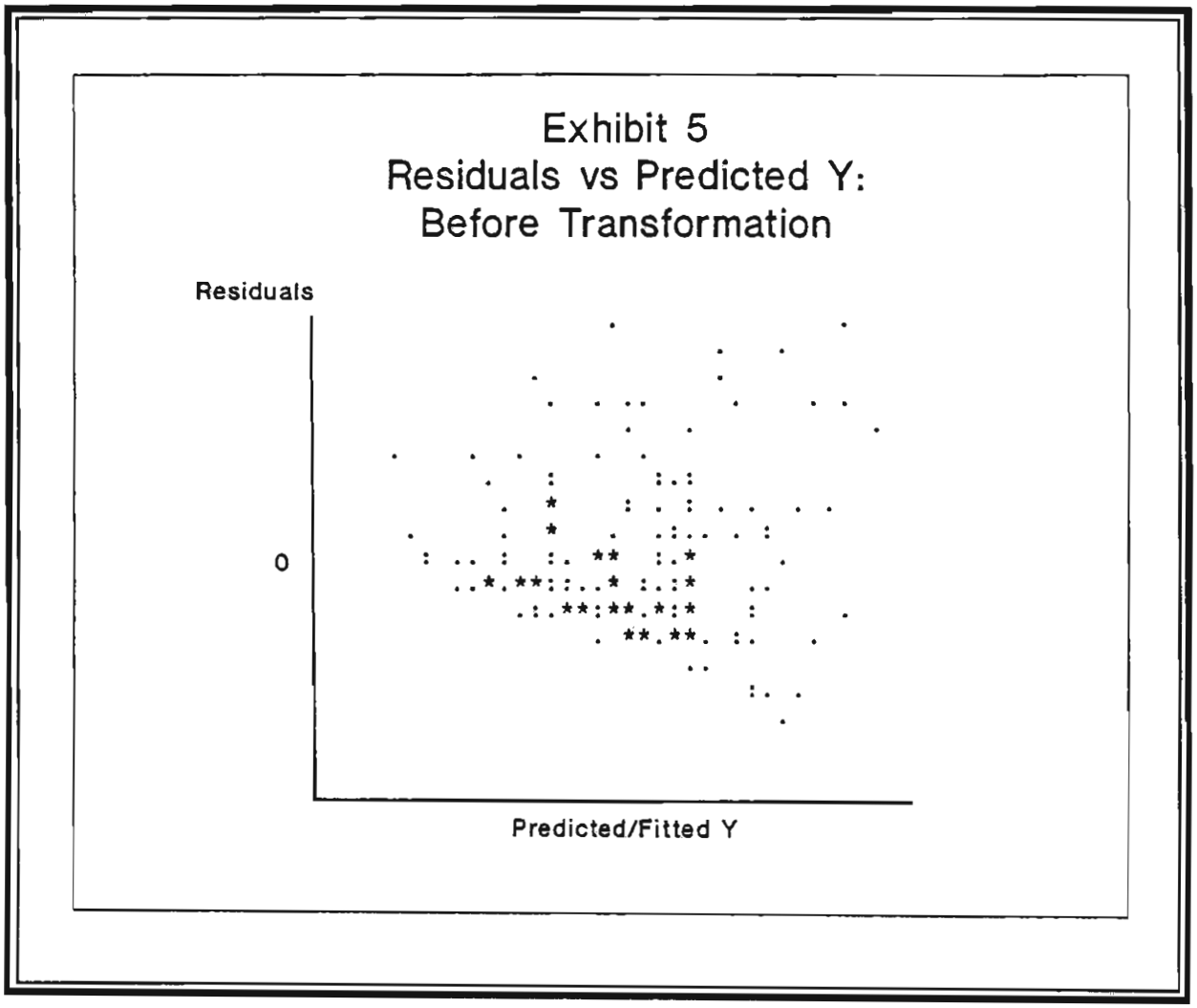


Exhibit 4
Normal Probability Plot: Before Transformation



Nonconstancy of variance or heteroscedasticity is demonstrated by the residuals versus predicted Y graph (Exhibit 5). If the residuals had constant variance with increasing values of predicted Y, the graph would show uniform scatter about the horizontal line at e=0. This is obviously not the case in Exhibit 5. Higher values of predicted Y have larger residuals. In some cases, nonlinearity can also be examined with such a scatterplot, for example, when there is an obvious curvilinearity to the plot. However, in the case where there are multiple independent variables, this is difficult to observe. However, sometimes it is evident when plotting residuals against particular independent variables.



SOLUTIONS

Several remedial measures were available. First, we could attempt various transformations to linearize the model, eliminate the nonconstant variance problem, and normalize the distribution of error terms. Alternatively, we could abandon the simple linear regression approach and seek other statistical models that fit our data set.

We chose to try various transformations prior to abandoning the regression approach. Our choices were threefold:

1) transform X,
2) transform Y, or
3) try some combination of X and Y transformations.

Since the residuals were nonrandomly distributed when plotted against predicted Y, we needed to change the shape and spread of the distributions of Y. Since the residuals generally increased with the predicted value of Y, a power transformation with power less than one or a logarithmic transformation was indicated. These transformations differentially impact the values of Y. That is, they have a larger impact on the higher rather than lower values of Y. As a result, they will tend to bring the residuals versus predicted Y plot into a more constant scatter around the horizontal line (that is constant variance for any X or predicted Y).

We chose to address this problem first with a logarithmic transformation to the dependent variable. (Note that weighted least squares can also be used to correct the variance nonconstancy violation. See Neter *et al.*, 1990, Chapter 14 for more information.)

Fortunately, this transformation not only solved the variance violation but also linearized the regression function and normalized the distribution of error terms. The logarithmic transformation had three major impacts:

1) It caused the distribution of residuals to more closely approximate a normal curve. (See Exhibits 6 and 7)

Exhibit 6
Histogram of Residuals:
After Transformation

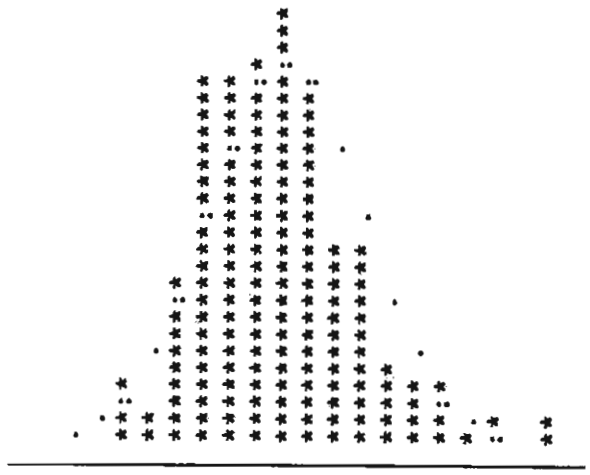
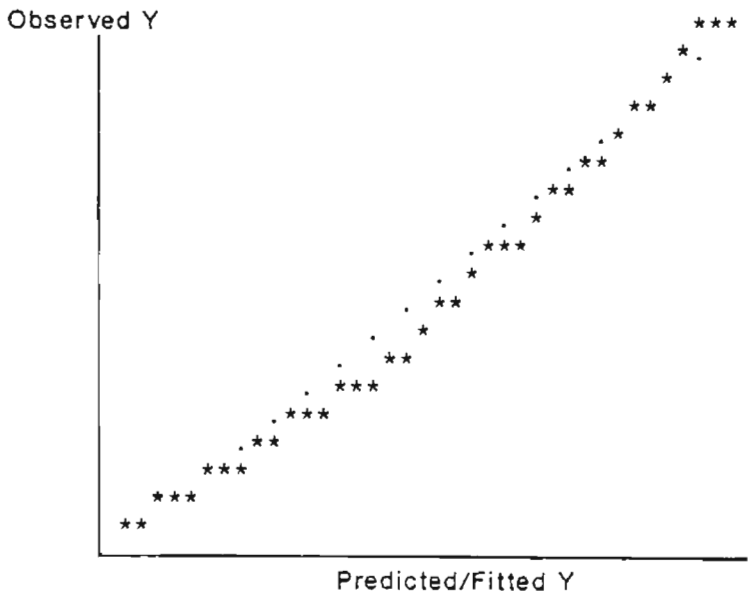
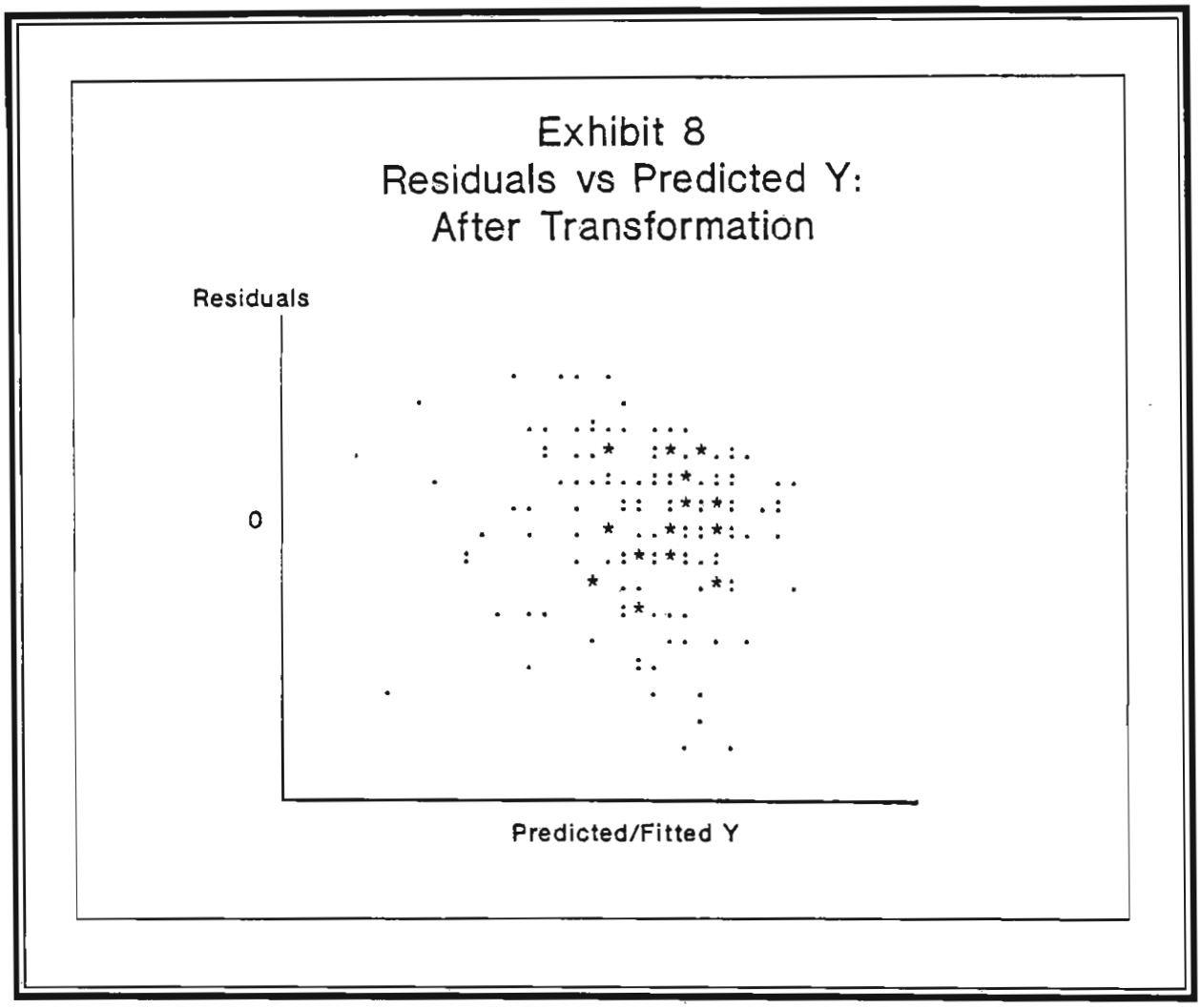


Exhibit 7
Normal Probability Plot: After Transformation



2) It eased the variance nonconstancy violation by bringing the residuals versus predicted Y plot into a more random scatter around the horizontal line (See Exhibit 8).



3) It changed the model's assumptions about linearity. By transforming the dependent variable, the estimation equation was linear, but the implied population model was nonlinear of the form Y = 10**(alpha + beta X + epsilon), where ** indicates that the term in parentheses is an exponent (that is, the power of 10). Hence, the logarithm to the base of 10 yields a linear form.

Transforming the dependent variable causes the R^2s to be noncomparable with those from regressions on the original variable due to the different units of Y and transformed Y. Consequently, an analyst should be careful about comparing R^2s derived from computer analyses of pre- and post-transformation regressions. However, it is possible to take the predicted values and transform them back individually into the original units for computation of R^2s that would be directly comparable. We did not do this in the present analysis.

The model developed for the total sample is:

$\log Y = 5.12 + 11.57 X_1 - .36 X_2 + 6.95 X_3 + .27 X_4$

where

logY = log transformed frequency of use
X_1 = Benefit variable 1 (BEN1)
X_2 = Category variable 8 (CAT8)
X_3 = Demographic variable 3 (DEMO3)
X_4 = Category variable 4 (CAT4)

SUBSEQUENT SEGMENTATION ANALYSIS

A portion of the client research addressed the question: Are there segments that have differing preferences for pricing options? Preference segments were derived by cluster analyzing conjoint utility data. Four preference segments emerged from the data.

Once a valid regression model was developed for the total sample, we wanted to know if the explanatory variable sets were the same or if a different regression model applied for each segment. Two segments were discarded for further modeling activity due to their insufficient sample size. The remaining segments were modeled using the log transformed regressions developed in the initial modeling effort.

As can be seen in Exhibit 9, this approach was particularly successful in identifying the explanatory factors driving Segment 1. It was also clear that the driving forces varied between segments. Note in the following equations that the primary factors impacting usage for Segment 1 are a benefit (BEN1) and a category (CAT8) variable. The drivers in Segment 2 are four category variables (CAT1, CAT2, CAT6, CAT8), a benefit (BEN1), and a demographic variable (DEMO1).

Exhibit 9
Modeling Summary
Key Explanatory Variables After Transformation

VARIABLES	TOTAL SAMPLE	SEGMENT 1	SEGMENT 2
BEN1	✓	✓	✓
CAT 1			✓
2			✓
4	✓		
6			✓
8	✓	✓	✓
DEMO 1			✓
3	✓		
R SQUARE	.28	.51	.28

Segments derived by clustering conjoint utilities

The model for Segment 1 is:

$$\log Y = 5.55 + .41 X_1 - .65 X_2$$

where

logY = log transformed frequency of use
X_1 = Benefit variable 1 (BEN1)
X_2 = Category variable 8 (CAT8)

The model for Segment 2 is:

$$\log Y = 4.77 + .21 X_1 - .65 X_2 + .70 X_3 + .32 X_4 + .79 X_5 + .36 X_6$$

where

logY = log transformed frequency of use
X_1 = Category variable 1 (CAT1)
X_2 = Category variable 6 (CAT6)
X_3 = Category variable 8 (CAT8)
X_4 = Demographic variable 1 (DEMO1)
X_5 = Category variable 2 (CAT2)
X_6 = Benefit variable 1 (BEN1)

WHAT WE'VE LEARNED

We learned or confirmed many of the basic tenets of model building:

1. We do not know, *a priori*, if the data will fit the mathematical model chosen.

2. If the data do not fit the model, there are remedial procedures available.

3. Although finding the proper transformation can be a trial and error procedure, graphical analysis can provide direction.

4. Several violations may occur simultaneously in a data set, yet a single corrective measure may eliminate or reduce all the violations.

5. Several measures of model quality are available that tell us different things:

   - R^2 estimates the strength of the relationship between dependent and independent variables

   - Residual analysis tests for the appropriateness of the regression model to the data at hand.

FURTHER RESEARCH

Additional research could move along at least three paths. First, the data could be factor analyzed to eliminate correlation among the independent variables and reduce the number of independent variables for potential inclusion in the regression. This would be particularly helpful for analyzing the smaller segments since the available degrees of freedom would be increased.

Secondly, alternative transformations could be applied to the data set to determine if they would improve the fit and/or the explanatory power of the equations.

Finally, a larger data set would enable us to evaluate the viability of this approach not only for the total sample, but also for each of the segments. In addition, more respondents would enable us to test the stability of the cluster solution.

REFERENCES

Green, Paul E. and D.S. Tull. *Research for Marketing Decisions.* 4th Edition. Englewood Cliffs NJ: Prentice Hall, 1978.

Kleinbaum, David G. and L.L. Kupper. *Applied Regression Analysis and Other Multivariable Methods.* North Scituate MA: Duxbury Press, 1978.

Neter, John, W. Wasserman and M.J. Kutner. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* Third Edition. Homewood IL: Irwin, 1990.

Norusis, Marija J. *SPSS/PC+ Manual.* Chicago IL: SPSS Inc., 1986.

Comments or questions are appreciated and should be directed to Mike Mulhern, Mulhern Consulting, 4732 NE 193rd, Seattle WA 98155, 206-365-6321, or Doug MacLachlan, Department of Marketing and International Business, University of Washington, DJ-10, Seattle WA 98195.

Comment on Mulhern and MacLachlan

Leland Wilkinson
SYSTAT, Inc. and Northwestern University

This paper illustrates the importance of transformations in market research. The problem is too frequently overlooked by some consultants who wish to provide a client an R-square statistic and ignore the aptness of the fitted model.

My only quibble with the paper is the chosen transformation. A square root transformation should bring these residuals closer to normality than does the logarithmic. While I haven't access to the data, I can illustrate this crudely. Figure 1 shows the histogram of the data on the dependent variable reconstructed from the raw residuals histogram in the paper and my assumptions about the model. I have tried several transformations on this histogram, shown to the right. The square root seems to fit the normal curve a little better and is at least less skewed.

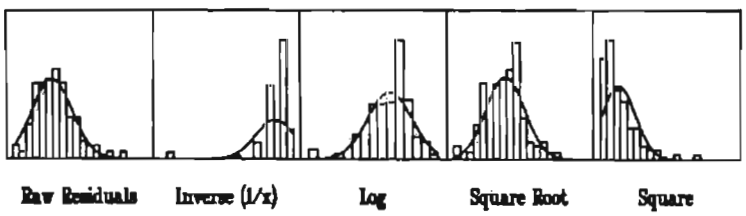


Figure 1. Transformations of Residuals

Other factors can affect the choice of transformation for these data — the model being fitted, the predictor variables, and the additive constant. I could only confirm my guess by transforming the raw data and refitting the model. Nevertheless, because the dependent variable consists of frequencies, I would guess my hunch is correct.

The authors' main point is worth repeating. Goodness of fit measures like R-square are not very useful for understanding a model. Examination of residuals is critical. I would go a step further. If your data are counts or frequencies, incomes, survival times, proportions, rates, ratios, and other types of non-additive derived statistics, a bell should ring in your head to remind you to examine histograms and scatterplots and consider a transformation to normality. The issues raised in this paper apply, with little qualification, to any linear modeling procedure, including factor analysis, principal components, conjoint analysis, latent variable causal modeling, and discriminant analysis. If transformations achieve an apt linear model, fine. Otherwise, proceed to nonlinear models or nonparametric analyses.

APPLICATION OF FACTORIAL SIMILARITY MEASURES IN THE DIFFERENTIATION OF TARGET MARKET CONSUMER GROUPS

Derek R. Allen, Ph.D.
University of Wisconsin — Milwaukee
Social Science Research Facility

1.1 RATIONALE

The intent of this analysis is to discuss a multivariate approach for exploring differences across consumer groups that are not revealed by either univariate or bivariate techniques. In short, this strategy entails a comparison of the factor structures of two consumer groups as derived from either factor or principal components analysis.

The first portion of this analysis will present a brief comparison of several authors' approaches to measures of factorial similarity. As will be shown, there is some disagreement as to the proper manner in which to determine the extent to which two or more factor solutions differ. Specifically, I will review several authors' prescriptions for comparing two or more factor solutions as derived through principal components or factor analysis. My pursuit of this topic was precipitated by an actual application involving the differentiation of a target market consumer group from the overall population. Based upon this foundation, I will present an actual application of an attempt to differentiate two consumer groups based upon their lack of factorial similarity. By doing so, I hope to demonstrate the efficacy of several of the approaches discussed below.

1.2 FACTORIAL CONGRUENCE

Factorial congruence involves the extent to which two separate factor structures, derived either through factor or principal components analysis, are similar. Measures of factorial similarity should be differentiated from efforts aimed at "matching" factors. The latter, according to Kaiser, Hunka, and Bianchini (1971, p. 410) involves rotating "two sets of arbitrary reference factors, such as principal axes, to find two new sets of factors which are as close together as possible, thus matching pairs of factors." In contrast, relating factors involves only the assessment of the degree to which the factors appear similar without manipulating the vector spaces defined in the original studies. According to Harman (1976, p. 347), factorial matching should be further differentiated from another related technique involving the transformation of one factor solution to fit a specific target matrix. The quantification of factorial similarity is the main thrust of this analysis. Of interest are a number of coefficients and measurements designed to assess the degree to which two factor solutions are in agreement.

Congruent or prescribed factor solutions usually are discussed in one of two contexts. The first involves the broader topic of confirmatory factor analysis. As Dillon and Goldstein (1984, p. 99) noted, confirmatory factor analysis simply involves the *a priori* specification of structure prior to model estimation. This means, according to the authors, "that some of the factor loadings and

some (but not necessarily all) of the common factor correlations are set equal to zero." The reliance of confirmatory factor analysis upon the Maximum Likelihood Method is traceable to Lawley (1940, 1942).

Kim and Mueller (1978, p. 55) categorize confirmatory factor-analytic approaches as either involving (a) one group or (b) two or more groups. In the former case, one posits either that a specific number of factors will emerge or offers hypotheses concerning the magnitude of each variable's loadings across the factors. One may also specify the orthogonality (or lack thereof) among the factors.

Confirmatory factor analysis involving a comparison of factorial similarity across two or more groups is most appropriately addressed by the "multiple-group method," according to Nunnally (1978, p. 394). According to the author, the method "can be used to test for the presence of a general factor, much as is done with the general-factor solution. At the other extreme, it can test for the presence of any number of hypothesized factors." Bernstein (1988, p. 209) also identifies Oblique Multiple Groups (OMG) method for assessing factorial similarity. This will be discussed again below.

Factorial similarity is also addressed in the broader topic of building a compelling body of knowledge in support of a theory and demonstrating factorial invariance. As Rummel (1970, p. 449) suggests,

> "...to build a science requires that findings be sufficiently explicit to make possible evaluation, replication, and comparison with other studies. Each study in its own right may contribute a bit of knowledge — a datum — to building a science. But these data output of different studies must be integrated into general propositions and given meaning in terms of a theoretical framework. This requires that comparison between findings be possible so that the replicable substantive patterns can be identified, and the unique, research-design-specific results can be discarded."

For factor analysts, relating numerous studies and factor solutions has not proven to be an easy task. Pinneau and Newhouse (1964, p. 271) noted that relating factor solutions from different studies has been a problematic methodological issue. Ever since Thurstone (1947) suggested that factor invariance be a goal in a programmatic factor analytic approach, attempts have been made to resolve the problem.

Harman (p. 341) identified this as a situation in which factorial similarity is of interest. It involves *fixed variables* and different samples as opposed to the more troublesome scenario characterized by *fixed samples* and different variables. Rummel (p. 450) suggests that when comparing the results of factor analyses involving either of the situations discussed by Harman, a host of characteristics may be compared.

In his discussion of objects of comparison in the assessment of factorial similarity, Rummel (p. 451) identifies a number of factor solution attributes which may be contrasted. The correlation inverses from two analyses, for example, can be compared. Since one basic goal of factor-analytic approaches is to minimize the off-diagonal elements of the inverse, the extent to which two or more solutions achieved this can be compared. Rummel also identifies the factor correlation, higher-order matrix, factor regression matrix, and factor score matrix as possible objects of comparison. The

rotated or unrotated factor loading matrices of one or more factor analytic studies may also be the object of comparison when gauging factorial similarity.

Rummel (1970:452-453) differentiates the substance of comparison from the object of comparison. The former involves *what* is being compared. Rummel suggests that two factor solutions can be compared on the basis of configuration, complexity, variance, number of factors, and communality. Configuration, for example, refers to the "pattern and magnitude of the loadings." Closely related is the notion of complexity which involves a comparison of the distribution of a variable's loadings across factors. That a variable meets the simple structure criterion in one study but not in another may be revealing. Rummel also identifies variance as a substance of comparison. Should one factor consistently account for a lot of variance across different studies, then this, too, should be of interest. Contrasting the number of factors that emerge from different factor-analytic studies will permit conclusions concerning parallel dimensionality. Finally, comparing communalities across studies may help differentiate variables that are consistently unique from those that are interrelated.

Clearly there are a variety of ways in which two factor solutions may be compared. *This analysis will focus on techniques that compare factor solutions derived from two groups of subjects involving the same variables.* It is this area in particular that appears to be a source of disagreement among factor analysts. The central question concerns how to make valid and defensible comparisons between two factor solutions that rely upon the same variables but employ different samples.

The source of disagreement among authors who have discussed the measurement of factorial similarity is the role of the factor loadings in the pattern matrix (or structure matrix in the oblique case). In particular, it has been argued that measurements that rely upon the covariation of loadings across two analyses are not valid. The reason, according to Nunnally (p. 432) is that the loadings are not the factors. Instead, they are representative of the extent to which each variable is associated with the factors. As Nunnally warns, "...it is important to remember that factors are linear combinations of the variables — actual linear combinations in component analysis and hypothetical linear combinations which are estimated in common-factor analysis." His conclusion that it is invalid to compare pattern matrix configurations — particularly to correlate loadings — is echoed by other authors. Bernstein (p. 202), for example, proscribes correlating loadings without considering the underlying correlational structure of the variables. As he suggests, "...two linear combinations can appear to be quite different from one another yet correlate very highly if the underlying variables that give rise to them are highly intercorrelated, a situation which is at least modestly probable in many applications."

Many authors, however, have suggested that correlating the loadings from two different analyses' pattern matrix is perfectly acceptable. Gorsuch (1974, p. 224), for example, offered this approach as valid in determining the extent to which two factor solutions are parallel. Harman and Rummel both restrict their discussions of the measurement of factorial similarity to coefficients that are derived from comparisons of the pattern matrices without regard for the underlying correlational relationships among the variables.

The school of thought represented by Nunnally and Bernstein *does* perceive some utility in measurements that rely upon comparisons between loadings. In particular they acknowledge that when only the pattern matrices are available the correlational approach must be used. And, such is often the case when comparing several authors' studies.

The next two sections of this paper will review a variety of methods for assessing factorial similarity. First, correlational approaches that rely exclusively on the pattern matrices will be examined. Next, various other strategies — some of which account for the underlying correlational structure among the variables — will be reviewed.

1.3 PATTERN MATRIX COMPARISONS

The measurement approaches reviewed in this section disregard the correlational relationships among the variables used in the analysis. They generally are restricted to a comparison of two studies' pattern matrices. Such an approach may be compelled if access to the correlation matrices is not possible. Still, as noted earlier, exclusive reliance upon this approach to compare factor structures may be misguided. That is, some authors have insisted that, if available, any attempt to relate pattern matrices take into account the original correlation matrix.

1.3.1 ROOT-MEAN-SQUARE DEVIATION

The most rudimentary technique for relating two factors from different studies relying upon the same set of variables is referred to by Harman (p. 343). Harman suggested using the *root-mean-square* deviation to assess the extent of agreement between two factors. To compare factor *p* of study 1 with factor *q* of study 2, the following was offered:

$$rms = \sqrt{\sum_{j=1}^{n} ({}_1a_{jp} - {}_2a_{jq})^2 / n}$$

Harman concedes that the root-mean-square deviation is a simplistic approach to determining factorial similarity. One of its most serious drawbacks is that while perfect agreement between two factors would yield a rms value of zero, the upper-end value is determined by the number of variables used in the study.

1.3.2 COEFFICIENT OF CONGRUENCE

Harman (p. 343), Bernstein (p. 208), and Rummel (p. 462) all refer to the *coefficient of congruence* as an additional means of assessing factorial similarity. They disagree with respect to *when* it should be applied, however. That is, as described above, Bernstein and others (Nunnally in particular) proscribe its use when the original correlation matrices or raw factor scores are available. This point will be revisited later.

The coefficient of congruence is similar to a correlation between two sets of pattern matrix loadings. It is not a correlation, however, since mean deviations are not considered and the summations are across the variables, not the individuals. Its strength is that, unlike the root-mean-square deviation, it is a coefficient and, as such, varies between zero and one. It has a long history and is traceable to Burt (1948, p. 185) who employed a proportionality criterion referred to as the "unadjusted correlation" when comparing sets of factor coefficients in an effort to relate two factor solutions. Later, Tucker (1951, p. 43) referred specifically to the coefficient of congruence. Wrigley and

Neuhaus (1955) in an explicit treatment of the problems associated with relating two factors, refer to the "degree of factorial similarity." Their measure was identical to that proposed by Tucker. Harman (p. 344) uses the following notation to describe the coefficient:

$$\varphi_{pq} = \frac{\sum_{j-1}^{n} ({}_1a_{jp} * {}_2a_{jq})}{\sqrt{(\sum_{j-1}^{n} {}_1a_{jp}^2)\ (\sum_{j-1}^{n} {}_2a_{jq}^2)}}$$

Pinneau and Newhouse (p. 275) have questioned the utility of this coefficient based upon a number of potential problems. First, it should be noted that there are no tests of significance that can be applied to the coefficient. Tucker (1951, p. 19), however, proposed that values greater than about .94 were evidence of congruent factors. In contrast, values less than .46 were regarded as "definitely so low that this factor will not be considered as a congruent factor."

One of the most serious drawbacks associated with the coefficient of congruence involves certain ambiguities with respect to interpretation. As Pinneau and Newhouse (p. 275) noted, the coefficient is strongly affected by sign. Two factors that have loadings with identical signs will have high coefficient values. Pinneau and Newhouse (p. 275) correctly identify this as a particularly serious problem for centroid factor analysis in which "the first factors from two different matrices (for fixed variables and different subjects) will almost always have high coefficients of congruence because of the high proportion of large positive loadings" in the first factors.

1.3.3 INTRACLASS CORRELATION COEFFICIENT

As a whole, correlation coefficients provide an inadequate gauge of factorial similarity. As Pinneau and Newhouse (p. 275) warned, the conversion of loadings to mean deviation data can lead to erroneous conclusions. In particular, the authors offered a hypothetical situation in which the factor loadings in one study ranged from 0.00 to +0.85 and a second study yielded loadings that varied from -0.85 to +0.85. The conversion of these pattern matrix data to standard scores would result in assigning equivalent values to very dissimilar loadings. That is, the 0.00 extreme in the first study would be equivalent to the -0.85 extreme in the second. Pinneau and Newhouse (p. 276) conclude that "one thus equates a variable which contains none of the common variance of the factor with one which shares a great deal of common variance of the factor on which it loads."

Rummel (p. 462) suggests that the intraclass correlation coefficient may have utility in assessing factorial similarity. In the present context, derivation of this coefficient will provide a rough ratio of the difference of within- and between-factor loading variation to the total variation of loadings on the factors. Rummel (p. 299) attributes this formula to Kendall and Stuart (1961). The total within-class variance, denoted W is written as:

$$W = \frac{\sum_{i-1}^{n} [(x_{ij} - \overline{X_i})^2 + (x_{ik} - \overline{X_i})^2]}{n}$$

In the case that two sets of factor loadings are exactly the same, *W* would equal zero. Thus, *W* is a good measure of the magnitude of similarity between two sets of factor loadings. In contrast, the second portion of the equation, *B*, measures the *between-class* variance. This is determined by calculating the grand mean of all the variables on the two factors and summing the squared deviations of this grand mean from the two factor-specific means.

$$B = \frac{\sum_{i=1}^{n} (\overline{X_i} - \overline{X_{jk}})^2}{n-1}$$

In the calculation of both *B* and *W*,

$$\overline{X_i} = (x_{ij} - x_{ik}) / 2$$

$$\overline{X_{jk}} = (\overline{X_j} + \overline{X_k}) / 2$$

Given the preceding definitions, the intraclass correlation coefficient can be computed as the ratio of the difference between *B* and *W* to the sum of *B* and *W*:

$$r_i = \frac{B-W}{B+W}$$

The intraclass correlation coefficient varies from -1.00 to +1.00. Interestingly, Rummel (p. 302) notes that this coefficient can be computed for any number of variables and, when computed for more than two variables, the maximum possible *negative* value is -1/(m-1). In the case of two variables (or factors), as discussed above, the maximum negative correlation is -1/(2-1)=-1.00.

Clearly, the three pattern matrix comparisons reviewed above differ with respect to the types of factorial differences they reveal. Each of these vector comparison methods involves pattern-magnitude similarity. The most rudimentary is the root mean square coefficient which measures any deviation between two vectors. It was considered to be quite stringent in this regard by Rummel (p. 461). And, it has been referred to as a more discriminating measure than the coefficient of congruence by at least one author (Joreskog, 1963, p. 128).

In contrast to the root mean square coefficient, the coefficient of congruence is the cosine of the angle between the factors in the *m* dimension vector space. As discussed in the preceding, when it is applied to factor loadings, it represents a measure of the ratio of the difference of between-factor loading variation to within-factor variation to the total loading variation across the factors.

Finally, the intraclass correlation coefficient represents an analysis of variance coefficient in that it divides the variance of two (or more) variables, *k*, into *k* parts. For each case measured on each variable, variance is attributable to either *within-class variation* or *between-class variance*.

It should be noted that these measures are all quite straightforward and can be computed readily by hand. Their emergence preceded the advent of high-speed digital computers. Thus, they represent the foundations upon which more recent attempts to meaningfully compare factor solutions have been built.

2.0 OTHER METHODS FOR ASSESSING FACTORIAL SIMILARITY

As noted earlier, several authors have proscribed measures like those introduced in the preceding section. Bernstein (p. 202), for example, emphasized that inferences concerning factorial similarity must be based upon covariation across *factor scores* not factor loadings. In particular, Bernstein warned that "*Sets of factors are equivalent to the extent that they or their corresponding factor scores correlate highly.*" Nunnally (p. 394) strongly concurred with this admonition.

2.1 USAGE OF SCALING AND STRUCTURE MATRICES

Bernstein (p. 206), in a discussion of the comparison of factor solutions with the same variables using different samples suggests the following double cross validation approach:

$$\mathbf{R}_{ab.a} = \mathbf{D'}_a\mathbf{W'}_a\mathbf{R}_a\mathbf{W}_b\mathbf{D}_b$$

$$\mathbf{R}_{ab.b} = \mathbf{D'}_a\mathbf{W'}_a\mathbf{R}_b\mathbf{W}_b\mathbf{D}_b$$

Two matrices are derived because the correlations among the original variables may differ between the two studies. In each case, the correlation matrix **R** is pre- and post-multiplied by both a factor score weight matrix **W** and *scaling matrix* **D**. The latter is derived by taking the reciprocals of the square roots of the diagonal of $\mathbf{W'}_a\mathbf{RW}_a$ in the case of $\mathbf{R}_{ab.a}$ and $\mathbf{W'}_b\mathbf{RW}_b$ for $\mathbf{R}_{ab.b}$. Bernstein (p. 206) suggests that while it is unlikely, the two matrices of inter-factor correlations *could* differ due to substantive correlational differences among the variables in the two original datasets.

2.2 CONFIRMATORY FACTOR ANALYTIC AND OTHER APPROACHES

Nunnally (p. 433) warned that if different subjects are used in two studies with the same variables, there is no reasonable manner to assess factorial similarity. Again, this is because comparison of factor scores is precluded. One way around this, according to Nunnally (p. 433), involves deriving and comparing weighted factor scores. In short, each analysis is associated with two sets of factor scores. The first set is derived as usual from the factor or component analysis. The second set of scores, however, is obtained by applying to the factor scores of the first analysis a series of weights derived from the factor scores of the second analysis. Thus, this explains the existence of correlation between the two sets of actual and derived factor scores.

A number of authors including Nunnally (p. 394) and Bernstein (p. 209) suggest that confirmatory analytic approaches may be most appropriate for assessing factorial similarity. The Oblique Multiple Groups (OMG) method and other confirmatory approaches (for example, Procrustes methods) permit hypothesis testing with respect to the presence or absence of specific factors. Regrettably, the scope of this analysis cannot accommodate a lengthy discussion of confirmatory analytic methods. Generally speaking, these approaches are more sophisticated than any of the coefficients discussed thus far and are to be preferred in the assessment of factorial similarity.

3.0 APPLICATION OF FACTORIAL SIMILARITY MEASURES

Differentiation of target market consumer groups from the overall population has always been a concern of marketing research practitioners. Demonstrating that one's target market behaves or thinks in a way that is dissimilar from the general consumer population has obvious advantages. These differences, for example, can be leveraged in broad advertising communications and utilized in more direct, targeted messages.

To have utility, a target market definition should result in a homogeneous group of consumers who share certain characteristics *apart* from the definitional criteria that grouped them. That is, a substantive target market definition will result in group differences across variables that did not constitute membership criteria. Thus, a target market definition such as *"...females between the ages of 35 and 44 with household incomes of more than $50,000..."* should yield a group that differs from the general population in more ways than age, gender, and socio-economic class. Indeed, their lifestyle and psychographic profiles will differ from those of other groups defined on the basis of similar demographics. More detailed target market definitions will increase the probability of encountering appreciable differences between groups.

This paper was precipitated by an attempt to differentiate a target market consumer group from the overall U.S. population. The target market group (n=300) was defined based upon demographic characteristics. It was hypothesized that their product preferences (that is, mean appeal ratings) would differ, significantly, from those of a group (n=300) that did not meet the target market criteria. The target market group was considerably more affluent than the non-target group.

Twenty-six different product attributes were rated by members of both target and non-target groups. A 10-point Likert scale was used in each case with magnitude relating directly to favorability. The proprietary nature of this study precludes release of detailed information concerning consumer preferences. Of course, this should be of little consequence because it is the assessment of factorial congruence that is of interest here.

A comparison of mean favorability ratings revealed a startling lack of significant differences between the target market and non-target market groups. That members of the target market group were virtually indistinguishable from the overall population in terms of their perceptions of the importance of the 26 characteristics was somewhat disconcerting. A separate components analysis of each group was then undertaken to reveal whether members of the two groups might differ with respect to the relationships they perceived between the 26 variables.

The components analyses suggested that the target market group may have been more sophisticated or discriminating in evaluating the 26 characteristics than the non-target group. The target market analysis revealed seven components while the non-target market analysis revealed only five. This result suggests that target market consumers may be able to differentiate among the characteristics more effectively.

In an effort to explore the extent to which any two factors were parallel, the measures of factorial similarity introduced earlier were calculated. A comparison of the various coefficients will be the focus of the next section of this analysis. The rotated factor patterns introduced in this section will be used to assess agreement between the various measures of factorial similarity.

3.1 COMPARISON OF THE MEASURES

When assessing the similarity of two pattern matrices, it is important to compare every factor in the first matrix with every factor in the second (Harman, p. 345). In each of the cases presented below, all of the factors are paired with one another. As noted earlier, the root-mean-square deviation is a simple distance metric; the lower its value the closer the vectors can be construed to be. In the case of Table 1, the relationship between component 3 of the non-target market analysis and component 2 of the target analysis emerged as strongest.

Table 1
Root-mean-square Deviations

(Target Market Factors)

Non Target	1_a	2_a	3_a	4_a	5_a	6_a	7_a
1_b	.30	.42	.38	.29	.38	.29	.38
2_b	.21	.42	.35	.29	.30	.29	.34
3_b	.38	**.13**	.30	.38	.30	.36	.36
4_b	.34	.31	.18	.27	.25	.31	.29
5_b	.32	.34	.32	.30	.30	.29	.16

Strongest relationship is highlighted.

Table 2
Coefficients of Congruence

(Target Market Factors)

Non Target	1_a	2_a	3_a	4_a	5_a	6_a	7_a
1_b	.75	.47	.55	.77	.53	.76	.53
2_b	.84	.37	.54	.68	.66	.68	.53
3_b	.42	**.93**	.61	.36	.61	.40	.40
4_b	.38	.48	.79	.53	.57	.33	.42
5_b	.45	.39	.35	.42	.38	.42	.81

Strongest relationship is highlighted.

Table 2 presents the coefficients of congruence for every factor pairing. As was the case with respect to the root-mean-square deviation, the relationship between target market component 2 and non-target component 3 emerged as the strongest. A brief visual inspection reveals considerable agreement between the coefficient of congruence and the root-mean-square deviation.

The intraclass correlation coefficients between the two groups' components are presented in Table 3. As was the case in the preceding tables, two components emerged as strongly related.

Table 3
Intra-class Correlation Coefficient

(Target Market Factors)

Non Target	1_a	2_a	3_a	4_a	5_a	6_a	7_a
1_b	-.16	-.66	-.60	-.08	-.63	-.09	-.58
2_b	.21	-.78	-.57	-.21	-.32	-.19	-.52
3_b	-.62	**.73**	-.24	-.66	-.24	-.56	-.59
4_b	-.59	-.33	.29	-.22	-.19	-.52	-.41
5_b	-.41	-.41	-.50	-.36	-.45	-.33	.45

Strongest (positive) relationship is highlighted.

Tables 4 and 5 present the scaled factor score matrices as proposed by Bernstein (p. 206). Again, there appears to be considerable agreement between the two scaled matrices with respect to the strongest inter-component relationships. The strongest link between the two groups of consumers seems to involve target market component 2 and non-target component 3. This was the case with every other measure investigated.

Table 4

Scaled Factor Score Matrix $R_{ab.a}$

(Target Market Factors)

Non Target	1_a	2_a	3_a	4_a	5_a	6_a	7_a
1_b	.26	.01	-.01	.50	-.66	.50	-.03
2_b	.64	-.11	.17	.19	.21	.21	.17
3_b	-.03	**.88**	.26	-.13	.07	-.01	.01
4_b	-.10	.00	.70	.25	.12	-.12	.08
5_b	.11	.07	-.03	.03	-.06	.06	.85

Strongest (positive) relationship is highlighted.

Table 5

Scaled Factor Score Matrix $R_{ab.b}$

(Target Market Factors)

Non Target	1_a	2_a	3_a	4_a	5_a	6_a	7_a
1_b	.54	.14	.20	.72	-.86	.78	.08
2_b	.66	-.12	.13	.33	.13	.39	.08
3_b	-.03	**.88**	.22	-.06	.05	.12	.02
4_b	-.13	-.01	.60	.26	.03	-.07	.12
5_b	-.01	.05	.00	.11	-.02	.09	.86

Strongest (positive) relationship is highlighted.

Table 6 demonstrates the extent to which the various measurements of factorial similarity were in (rank order) agreement. Each of the coefficients and measurements found the relationship between target market component 2 and non-target component 3 to be the strongest. In the case of the coefficients whose range was from -1.00 to +1.00 this was the strongest *positive* relationship. That is, there were inverse relationships with absolute values greater than the positive value which was encountered between these two particular components.

The data suggest that the most comparable measurements were the root-mean-square deviation and the intra-class correlation coefficient. The root-mean-square deviation appeared to be the most independent of the measures. That is, it was associated with two of the weakest (.66) rank order relations in the table. Overall, however, there was a surprising level of agreement among the measurements.

Table 6
Measurement Congruence: Spearman Rank Order Correlations

	RMS	CC	ICC	$R_{ab.a}$	$R_{ab.b}$
RMS	1.00				
CC	.66	1.00			
ICC	.94	.76	1.00		
$R_{ab.a}$	.78	.86	.87	1.00	
$R_{ab.b}$	.66	.82	.74	.88	1.00

Note: all coefficients are significant at =.01

4.0 SUMMARY AND CONCLUSIONS

The intent of this paper has been to demonstrate an application of several factorial similarity measurements. The application itself has been assigned a secondary role. To a large extent this approach was taken because of the proprietary nature of these data. It should be clear, however, that the type of information revealed here could be of considerable utility in a marketing environment. In particular, it was demonstrated that while the two groups of consumers were virtually identical with respect to their mean product attribute ratings, they tended to be quite different in a multivariate sense. Further, the only substantive factorial link between the two groups illuminated a strong source of similarity between them.

Theoretically, two groups of subjects *randomly* selected from the same population should yield parallel factor structures. In the present case, the two groups were differentiated by a number of demographic characteristics. And, while they were virtually identical in a univariate sense, they tended to depart tremendously when examined in a multivariate context.

As I noted earlier, a discussion of confirmatory factor-analytic techniques was beyond the scope of this paper. Confirmatory factor analysis probably represents a more robust approach to testing hypotheses concerning the presence of specific factors in more than one group.

With respect to the debate concerning the applicability and validity of each of these measures, there seems to be substantive evidence here to suggest that they may actually be quite similar. In fact, there was a surprising amount of agreement among the various measures that were investigated here.

REFERENCES

Bernstein, Ira H. (1988). *Applied Multivariate Analysis.* New York: Springer-Verlag.

Burt, C. (1948). "The Factorial Study of Temperamental Traits." *British Journal of Mathematical and Statistical Psychology.*

Dillon, William R. and Matthew Goldstein (1984). *Multivariate Analysis: Methods and Applications.* New York: John Wiley & Sons.

Gorsuch, R.L. (1974). *Factor Analysis.* Philadelphia: Saunders.

Harman, Harry H. (1976). *Modern Factor Analysis.* Chicago and London: The University of Chicago Press:

Horst, Paul (1961). "Relations Among *m* Sets of Measures." *Psychometrika.* 26:129-149

Joreskog, K.G. (1963). *Statistical Estimation in Factor Analysis.* Stockholm: Almqvist and Wiksell.

Kaiser, Henry F., Steve Hunka and John C. Bianchini (1971). "Relating Factors between Studies Based upon Different Individuals." *Multivariate Behavioral Research.* 10:409-421.

Kendall, M. and A. Stuart (1961). *The Advanced Theory of Statistics.* New York: Hafner. Vol. II. *Inference and Relationship.*

Kim, Jae-On and Charles W. Mueller (1978). *Factor Analysis: Statistical Methods and Practical Issues.* Sage University Press series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills and London: Sage Pubns.

Lawley, D.N. (1940). "The Estimation of Factor Loadings by the Method of Maximum Likelihood." *Proceedings of the Royal Society of Edinburgh.* 60:64-82.

Lawley, D.N. (1942). "Further Investigations in Factor Estimation." *Proceedings of the Royal Society of Edinburgh.* 60:64-82.

Nunnally, Jum C. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Pinneau, Samuel R. and Albert Newhouse (1964). "Measures of Invariance and Comparability in Factor Analysis for Fixed Variables." *Psychometrika.* 29:271-281.

Rummel, R.J. (1970). *Applied Factor Analysis.* Evanston, Illinois. Northwestern University Press.

Thurstone, L. L. (1947). *Multiple-factor Analysis.* Chicago: University of Chicago Press.

Tucker, L. R. (1951). "A Method for Synthesis of Factor Analysis Studies." Personnel Research Section report No. 984. Washington, D.C.: Department of the Army.

Wrigley, C., and J. O. Neuhaus (1955). "The Matching of Two Sets of Factors." Contract Memorandum Report A-32, Task A. Urbana: University of Illinois.

Comment on Allen

Michael G. Mulhern
Mulhern Consulting

I would like to cover three basic areas in my remarks on Allen's paper. First, I'll comment on some strengths of the paper. Next, a weakness will be discussed and, finally, a direction for future research will be proposed.

A major strength of this paper is its exploration of an ignored area within marketing research. Market researchers typically focus on choosing the best technique for generating perceptual maps and pay little, if any, attention to the congruence or similarity of the factor matrices and the measurement or managerial implications of this similarity (Pilon, 1989 and Pilon, 1992).

Another strength of this paper is its solid grounding in psychometric theory. Often, applications oriented papers are merely case studies with no basis in theory. Allen's work, however, is a welcome exception.

A third strength is the proper selection of perceptual mapping techniques. As discussed elsewhere (Shocker, 1987), the selection of mapping techniques will depend on the application. In general, discriminant based maps excel at differentiating among objects (for example, products, brands, companies) while factor based maps are preferred for discovering the underlying structure of the data. The choice of factor analysis for this application was appropriate.

When reading Allen's paper, I was unsure whether it addressed a practitioner or academic audience. In future versions, this minor weakness should be addressed by targeting a single audience and writing in the style most relevant to that audience.

With respect to future research, a question arises regarding the impact of factor extraction methods on the congruence measures. It appears that Allen used a single extraction method to generate the factor matrices. If alternative factor extraction methods generate different factor structures, then the findings could be quite different.

As an informal empirical test of the hypothesis that different extraction methods generate different factor structures, I factor analyzed a database of 600 respondents with SPSS-PC+ using seven different extraction methods. (For a comparison of the various extraction methods, see Kim and Mueller, 1978, 12-29). The factor structures were not rotated. The sample was drawn from a population of western Washington business service users.

The table below summarizes the results:

EXTRACTION METHOD	NUMBER OF FACTORS	TOTAL VARIANCE EXPLAINED	VARIANCE EXPLAINED BY FIRST FACTOR
PC	5	74%	35%
PAF	5	68	34
ML	5	68	26
ALPHA	5	67	34
IMAGE	5	64	33
ULS	5	68	34
GLS	5	68	26

Total Variance Explained = Total variance explained by all factors with eigenvalues over 1.0.

Variance Explained by First Factor = Variance explained by the factor with the largest eigenvalue.

PC = Principal components
PAF = Principal axis factoring
ML = Maximum likelihood
ALPHA = Alpha factoring
IMAGE = Image factoring
ULS = Unweighted least squares
GLS = Generalized least squares

To summarize the findings from the table, the number of factors is the same across methods, the variance explained by the factors with eigenvalues greater than 1.0 is within a 10 percentage point range (64%-74%), and the variance explained by the first factor covers a nine percentage point range (26%-35%). The factor extraction methods that produced results that were most different were maximum likelihood estimation and generalized least squares. These preliminary results indicate that there may be some value in more closely evaluating the impact of the factor extraction method not only on the factor matrices but also on the subsequent convergence score comparisons.

In summary, I believe Allen's paper is well worth reading; particularly for those of us who would like to explore an alternative view of mapping or desire an introduction to psychometric theory and its relationship to factor analysis.

REFERENCES

Kim, J. and C.W. Mueller (1978). *Factor Analysis: Statistical Methods and Practical Issues.* Sage University Press Series on Quantitative Applications in the Social Sciences, 07-014. Beverly Hills CA: Sage Publications.

Norusis, Marija J. (1986). *SPSS/PC+ Advanced Statistics Manual.* Chicago IL: SPSS Inc.

Pilon, Thomas L. (1989). "Discriminant versus Factor Based Perceptual Maps: Practical Considerations." *Sawtooth Software Conference Proceedings,* 166-182.

_____ (1992). "Comparison of Results Obtained from Alternative Perceptual Mapping Techniques." *Sawtooth Software Conference Proceedings,* (this volume).

Shocker, Allan (1987). "Perceptual Mapping: Its Origins, Methods, and Prospects." *Sawtooth Software Conference Proceedings,* 121-142.

THE NUMBER OF LEVELS EFFECT IN CONJOINT: WHERE DOES IT COME FROM, AND CAN IT BE ELIMINATED?

Dick R. Wittink
Cornell University
Joel Huber
Duke University
Peter Zandan
IntelliQuest, Inc.
Richard M. Johnson
Sawtooth Software

INTRODUCTION

In 1981 Currim *et al.* found, in a conjoint study with six attributes, three attributes defined on three levels and three defined on two levels, that the three-level attributes achieved much higher average importances than the two-level attributes. When they designed their study they had not anticipated this result. Subsequently, in various experimental studies, such level effects have been demonstrated to exist for all popular data collection and parameter estimation methods.

In this paper we define the attribute-level problem, we review results obtained in various experimental studies, and we discuss its relevance to commercial applications of conjoint analysis. Given a multitude of possible sources of this systematic effect, we motivate an experiment that is designed to provide further insight into the plausibility of alternative explanations for its existence. We interpret the experimental results, and we end with a discussion of how the level effect can be eliminated.

THE ATTRIBUTE-LEVEL PROBLEM

Currim *et al.* studied subscription series to performing arts events, using a variation of the tradeoff matrix data collection method (Johnson, 1974). Each tradeoff matrix involved one three-level and one two-level attribute. Pairs of objects were chosen from the matrices. Each pair involved a tradeoff. For example, one object would have the lower price, and the other object would have a superior benefit. Respondents were asked to indicate which of two objects they preferred for each of several pairs of objects. The strict preference data were used together with an assumed preference order for attribute levels (for example, the lowest price is most preferred) to derive complete rank orders for all the objects in a tradeoff matrix.

For managerial purposes, they summarized the results by computing average attribute importances. Importance was defined as the difference in partworths between the best and worst levels for an attribute. For each matrix involving one three-level and one two-level attribute, the importances for the two attributes summed to 100. In this study, the best and worst levels (for example, lowest and highest prices) also necessarily had the highest and lowest partworths. They found that the

three-level attributes had average importances no less than 0.55 and no more than 0.66. The two-level attributes had average importances between 0.36 and 0.45. Computed separately for six segments, the three-level attributes' importances were no less than 0.49. The two-level attributes' importance never exceeded 0.50. Thus, attribute importances appeared to depend on the number of levels.

In subsequent studies, Wittink *et al.* (1982) and Wittink *et al.* (1989) showed that experimental manipulations of the number of attribute levels resulted in systematic differences in estimated attribute importances. Systematic level effects were found for full-profile rank orders, full-profile paired comparisons, full profile ratings, and tradeoff matrix ranks, using both metric (least-squares regression) and nonmetric (for example, LINMAP, in Srinivasan and Shocker, 1973) estimation methods. In one study involving five product categories, the relative importance of price was, on average, seven absolute percentage points higher when price had two additional intermediate levels (Wittink *et al.* 1989).

To investigate the research hypothesis that ACA (ACA System for Adaptive Conjoint Analysis by Sawtooth Software) might not suffer from a level effect, Wittink *et al.* (1991) compared results for ACA and full-profile ratings. The ACA results did include a level effect, although its magnitude was found to be only about half what occurred in the full-profile data. Their study, however, did not include experimental manipulations that might distinguish between different sources for the effect. For example, Green and Srinivasan (1990) suggest that the addition of intermediate levels for an attribute in the conjoint design may make a respondent pay more attention to that attribute. Currim *et al.* proposed a mathematical or algorithmic explanation.

Managerial Relevance. The existence of a level effect can have profound implications for the managerial conclusions obtained from a conjoint study. Consider for example, the results contained in Currim *et al.* One of their three-level attributes is a discount percentage. It was the least important of the three three-level attributes, but considering all six attributes it was the third most important. Currim *et al.* adjusted their results by considering the minimum and maximum possible importances for the attributes based on the number of levels in the conjoint design. After this adjustment, discount percentage was the least important of all six attributes. We expect managers to take different actions on the discount attribute when it is the least important as opposed to the third most important, in a list of six attributes.

For managerial purposes it is, nevertheless, much more pertinent to use market simulations than to rely on attribute importances. It is easy to demonstrate, however, that if attribute importances are affected by attribute levels, preference share predictions can be as well. Suppose, for example, that two attributes, A and B, can both have four levels. And imagine that in respondent X's mind the differences between the best and worst levels of both attributes are truly equally important. But if attribute A gets 4 and B gets 2 levels, A will obtain more relative importance. Conversely, if A gets 2 and B 4 levels, B will obtain higher importance.

To be precise, suppose that A will obtain a relative importance of 54 if it has 4 levels (and B would have a relative importance of 46 with 2 levels). But A will obtain a relative importance of 46 if it has 2 levels (so that B has an importance of 54 when it has 4 levels).

Now imagine that in a market simulation there are two products. Product I has the best level of A and the worst of B, while product II has the worst level of A but the best level of B. Then, based on

the first-choice rule, respondent X would be predicted to choose product I if A has 4 (and B has 2) levels (because with 4 levels A is more important, and product I is superior on attribute A). But respondent X would be predicted to choose product II if A has 2 (and B has 4 levels).

If the choice rule involves a predicted probability of choosing an object, it is easier yet to demonstrate a level effect on market simulations. Since the importances are based on differences in partworths, it should be clear that subsequent computations that involve differences in or ratios of predicted utilities for objects will be systematically affected by the level effect.

THE LEVEL EFFECT IN ACA

Wittink *et al.* (1991) manipulated the number of levels for four attributes in a study of refrigerators. The attributes, Capacity, Energy Cost, Compressor Noise, and Price, had either two or four levels. For example, for Energy Cost the best and worst levels were $70 and $100, respectively. Half the respondents saw only those two levels, while the other half also saw two intermediate levels, $80 and $90. For full profile ratings, the relative importance of Energy Cost, based on the difference between the highest and lowest partworths (out of five attributes) was found to be 8.0 with two levels but 20.5 with four levels. For ACA the relative importances were 13.7 for two and 19.9 for four levels of Energy Cost.

Across the four manipulated attributes, the difference in relative importance between the two- and four-level conditions was between 9 and 12 absolute percentage points for full profile, and between 5 and 6 absolute percentage points for ACA. Thus, the level effect was a much more serious phenomenon in the full-profile task than in ACA. But, because the ACA task proceeds in stages, Wittink *et al.* (1991) were able to trace the source of the level effect for ACA data.

They found no evidence of a level effect in the self-explicated data that provide the initial partworth solution in ACA. Thus, the level effect had to occur in the section in which preference intensity judgments are collected for one out of each pair of objects. Ideally, ACA chooses pairs of objects such that the predicted overall utility difference (based on the initial or updated partworth solution) between the objects is close to zero. An inspection of possible situations showed that if a respondent's self-explicated importances are incongruent (negatively correlated) with the numbers of levels assigned to the attributes for that respondent, it is impossible to achieve equality in the predicted utilities for objects defined on two attributes.

Wittink *et al.* reasoned that:

a) a bias in providing preference intensities toward the center of the scale can produce the levels effect;
b) such a bias is less likely to occur if the predicted utility differences are close to the center of the scale;
c) ACA is better able to produce pairs of objects with equal predicted utilities when respondents have higher importances for attributes with more levels.

This reasoning was used to construct the hypothesis that there should be a smaller levels effect for "congruent" respondents, that is, those for whom the more important attributes have more levels. Indeed, for respondents with congruent self-explicated importances and attribute levels, no level

effect in ACA was found. On the other hand, for respondents with incongruent self-explicated importances and attribute levels, the magnitude of the level effect for ACA varied from 9 to 11 percentage points (or about twice the average level effect in ACA). This suggests that the magnitude of the level effect in ACA depends on the difference in predicted utility between two objects. Thus, the results in Wittink *et al.* (1991) suggest an algorithmic explanation for the level effect. Yet, alternative explanations cannot be ruled out.

EXPERIMENTAL DESIGN

To obtain further insight into the plausibility of alternative explanations we conducted another study with four manipulations: (1) the number of attribute levels (for four attributes); (2) the balancing of paired objects in terms of predicted utilities; (3) a tutorial on the meaning of the attributes; and (4) the inclusion of prompts in the preference intensity section of ACA.

Based on the consistent results obtained in the earlier studies referred to, we expect an attribute's importance to be higher with more levels. The second manipulation (balancing) is expected to influence the magnitude of the level effect. The closer the predicted utilities of the objects, the smaller the potential for a level effect.

The third and fourth manipulations represent attempts to study potential behavioral explanations of the effect. The third manipulation consisted of making tutorial information available to half the respondents. Respondents who do not fully understand the attributes may react to cues unintentionally provided by the researcher. For example, respondents may infer that attributes with more levels should have more importance. If so, we would expect additional information about the attributes to increase respondents' understanding, such that the results better reflect their "true" importances. In this manner, the levels effect may be reduced under the tutorial treatment.

The fourth manipulation, "prompting," was intended to jar respondents into a higher state of awareness. A bored respondent may provide answers with a substantially random component and may have a larger midscale bias. Thus, we hypothesized that an occasional challenge to an answer provided by the respondent might produce better responses. At randomly chosen points in the interview the screen turned red with the message: "Are you sure about your answer of x? Please think some more about the strength of your preference. Press any key now to answer the question again."

The authors' expectations differed for this experiment. Huber, Johnson, and Zandan favored a "behavioral" explanation for the level effect, while Wittink favored an "algorithmic" one. A behavioral source had been illustrated in a study reported by Johnson (1992). Respondents were asked the dollar values of improvements in TV sets. Values of improvements from the worst to the best level of each attribute were found to be greater if attributes had intermediate levels. Since importance was stated directly by the respondent, rather than estimated by an algorithm, that study suggested a behavioral origin of the effect.

The algorithmic argument, on the other hand, concerned the ability of ACA to produce pairs of objects with nearly equal estimated utilities for each respondent. ACA ordinarily tries to produce pairs that are "balanced" in the sense of both objects having nearly equal utility. Although we couldn't produce a version of ACA that did a better than usual job of this, we were able to modify

ACA to do a worse job, by disabling the part of the program that does the balancing. The algorithmic hypothesis would suggest that respondents who received balanced pairs would display a smaller level effect than those for whom the pairs were not balanced.

The experimental design to test hypotheses about behavioral and algorithmic explanations of the level effect was a 2^4 full factorial design. The product category chosen was the notebook computer. Six attributes were used: (1) brand name; (2) notebook size; (3) weight; (4) battery life; (5) performance; and (6) purchase price. Half the respondents saw two levels of notebook size and battery life but four levels of weight and purchase price. The allocation of these levels to the attributes was reversed for the other half of the respondents.

RESULTS

Data were obtained from 403 respondents (40 percent) out of 1,008 surveys mailed. The sample was drawn from the office intensive file of Dun and Bradstreet and screened over the telephone by the market research firm of IntelliQuest to locate an individual with responsibility for purchasing or using notebook computers.

For each respondent, attribute importances were calculated based on the difference between the largest and smallest partworths for each attribute. Relative importances (by making the importances sum to 100) were analyzed as a function of all main and interaction effects for the experimental manipulations. The relative importance equations of the two attributes for which no level manipulation occurred had no significant explanatory power. For the other four equations, the level effect was significant ($p < .01$) in each case. And the level effect was significantly higher without balance than with balance for two of the attributes. No other consistent effects were obtained.

The interaction effect between level and balance was negative for all four attributes. Thus, in all four cases did the absence of balance in ACA increase the level effect. This effect cannot be attributed to a behavioral phenomenon involving the level manipulation. And, because neither of the two manipulations designed to produce a behavioral effect showed a significant interaction effect, we conclude that the evidence from this study is entirely consistent with an algorithmic explanation.

The Balance Manipulation. ACA includes a section that eliminates dominated pairs of objects from consideration. For example, if Price and Battery Life are the two attributes, the commercial version of ACA would not select one object to be better than the other on both Price and Battery Life. The selection of only nondominated pairs is what we call the existence of balance.

By eliminating this section from ACA for half the respondents, we allowed the pairs of objects to be both nondominated and dominated. The result is that the expected difference in predicted utilities for paired objects is now further from zero. If respondents have a tendency to provide preference intensity ratings toward the midpoint of the scale, the opportunity for distortion is greater when dominated pairs of objects are allowed.

It turns out, however, that the nature of the possible interaction between the level and balance manipulation is very complex. We will address this issue in detail in a future paper. We restrict ourselves here to the empirical results.

We show the average relative importances for all respondents, and separately for the combinations of levels and balance, in Table 1. Although we show separate averages for all four attributes, only the difference (in average difference between 2 and 4 levels) between the balance alternatives for Size and Price are significant ($p < .01$). Considering all four attributes, the level effect is between 1.34 and 4.00 when balance exists, but it is between 4.60 and 6.76 in the absence of balance. Thus, the commercial version of ACA with nondominated pairs is favored over a version that also allows for dominated pairs of objects.

Table 1

Average Relative Importances, Overall and for Level and Balance Combinations

		Experimental Manipulations				
		Balance	Yes		No	
Attribute	Overall Average	Levels	2	4	2	4
Size	9.52		7.25	11.25	6.43	13.19
			4.00		6.76	
Weight	12.05		10.22	13.78	9.80	14.40
			3.56		4.60	
Life	15.97		14.48	17.04	13.74	18.50
			2.56		4.76	
Price	22.06		22.42	23.76	18.14	23.92
			1.34		5.78	
Average Difference Between 2 and 4 Levels			2.84		5.48	

Noise-Adjusted Importance Measure. Even though the number of levels was manipulated for attributes that can be expected to have a monotone preference function, it is possible for the partworths to violate monotonicity. And, of course, violations of monotonicity are more likely to happen for attributes with a larger number of levels. Thus, it is possible that the magnitude of a level effect is reduced when we use an importance measure that is not influenced by statistical noise.

An alternative importance measure is to define importance based on the difference in partworths for the extreme levels. All four attributes with level manipulations in our study are assumed to be monotonically related to preference. That is, the lower the price, other things being equal, the more preferred an object. Similarly, the smaller the size, the lower the weight, and the longer the battery life, the more attractive a notebook is expected to be. By defining the importance using the difference in partworths for the extreme levels, we also ensure consistency across the level manipulations. For example, this measure now captures the importance of the difference between $3,600 and $2,300 (the extreme prices) for Price, in all cases (as is also true for the self-explicated importance in ACA). However, we note that in the first paper on level effects (Currim *et al.*) statistical noise played no role. In that paper the highest and lowest partworths necessarily occurred for the best and worst levels defined *a priori*.

We estimated the effects of the experimental manipulations on this noise-adjusted measure of importance in exactly the same manner as for the original importance measure. Significant level effects occurred for Size ($p < .01$), Life ($p < .01$), and Price ($p < .05$), but not for Weight. The balance manipulation produced a significant interaction effect with the attribute levels only for Price ($p < .05$). However, with the exception of Weight, the interaction effect has the expected sign. Neither the prompting nor the tutorial manipulations showed significant interactions with the level manipulation. For comparison purposes we show the average relative (noise-adjusted) importances for all respondents, and separately for the combinations of levels and balance, for all four attributes in Table 2.

The overall average importances do not appear to be very different when we compare the first column in Table 2 with the first one in Table 1. The importances of Life and Price are now somewhat higher, implying that one or both of the other attributes (Brand Name and Performance) have lower importances. The level effect in the balance condition is now between -0.10 and 2.24, while in the other condition it is between 0.68 and 4.24. On average, the level effect is approximately half in Table 2 of what it is in Table 1.

Table 2

Average Relative Noise-Adjusted Importances, Overall and for Level and Balance Combinations

Attribute	Overall Average	Experimental Manipulations: Balance	Yes: Levels 2	Yes: Levels 4	No: Levels 2	No: Levels 4
Size	9.33		8.24	9.92	7.68	11.48
			1.68		3.80	
Weight	12.70		12.33	13.73	12.03	12.71
			1.40		0.68	
Life	17.99		16.73	18.97	16.49	19.77
			2.24		3.28	
Price	25.80		27.06	26.96	22.42	26.76
			−0.10		4.24	
Average Difference Between 2 and 4 Levels			1.31		3.00	

DISCUSSION

The results of our study indicate that if ACA did not exclude dominated pairs of objects from consideration, the level effect would be larger (at least in this application). On average, the difference in importance when an attribute has four versus two levels is 2.84 absolute percentage points for the current version of ACA but 5.48 absolute percentage points when dominated objects are allowed to occur in the preference intensity section. The noise-adjusted importance measure shows that the level effect is reduced, but not eliminated.

Our study did not provide statistical support for the manipulations designed to capture behavioral effects. This does not mean that respondents do not react to the number of levels included for a given attribute. All we can claim is that the prompt and tutorial manipulations in this application were insufficient. It is possible that other manipulations can provide a substantial reduction in the level effect. The results of this study suggest that the source of the effect is algorithmic.

Given that the balance manipulation reduced the level effect to about half its magnitude otherwise, it is appropriate to ask whether ACA can be improved further. Wittink *et al.* (1991) have shown that the level effect, in a study of refrigerators, was zero for respondents with congruent attribute levels and self-explicated importances. This means that if a respondent indicates that attribute A is very important and B is not important, A should have, say, 4 and B have 2 levels. In that case, the predicted utilities for objects defined on attributes A and B, selected by ACA, are approximately equal. Thus, if ACA is modified to increase the likelihood of selecting pairs of objects with equal predicted utilities, the level effect will become smaller.

The more important an attribute is, the more reason we have to include intermediate levels, both to avoid level effects and to maximize useful learning. That is, there is no need to learn the partworths for intermediate levels if an attribute is unimportant. Thus, the more important an attribute is for a given respondent, the more reason we have to include intermediate levels in order to understand the shape of the function and to minimize the level effects. Our results suggest that the next ACA version would benefit if modified to reflect this idea. In the meantime, conjoint analysts can improve the results by using a larger number of levels for the attributes with higher (expected) importances. This information can be obtained in a pre-test of respondents (or in a procedure such as the self-explicated portion in ACA).

Finally, we want to emphasize that despite the absence of evidence in favor of psychological explanations of the level effect, it seems to us it is always a good idea to make respondents as motivated and as smart as possible.

We thank Karlan Witt of IntelliQuest for her assistance with various aspects of the study.

REFERENCES

Currim, Imran S., Charles B. Weinberg, and Dick R. Wittink (1981). "The 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 Research: New Developments and Directions." *Journal of Marketing*, 54 (October), 3-19.

Johnson, Richard M. (1974). "Trade-Off Analysis of Consumer Values." *Journal of Marketing Research*, 11 (May), 121-7.

Johnson, Richard M. (1992). "Comment on Attribute Level Effects." *Second Annual Advanced Research Techniques Forum Proceedings*, American Marketing Association, 62-4.

Srinivasan, V. and Allan D. Shocker (1973). "Linear Programming Techniques for Multidimensional Analysis of Preferences." *Psychometrika*, 38, 337-69.

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 Effects of Differences in the Number of Attribute Levels on Conjoint Results." *Marketing Letters*, (2), 113-23.

Wittink, Dick R., Joel C. Huber, John A. Fiedler, and Richard L. Miller (1991). "The Magnitude of and an Explanation/Solution for the Number of Levels Effect in Conjoint Analysis." working paper, November.

WITHIN- AND ACROSS-ATTRIBUTE CONSTRAINTS IN ACA AND FULL PROFILE CONJOINT ANALYSIS

Ivo A. van der Lans
University of Leiden (The Netherlands)
Dick R. Wittink
Johnson Graduate School of Management, Cornell University
Joel Huber
Fuqua Graduate School of Business, Duke University
Marco Vriens
University of Groningen (The Netherlands)

INTRODUCTION

Studies by Srinivasan *et al.* (1983) and van der Lans and Heiser (1990) indicate that the predictive validity of partworth estimates in conjoint analysis can be improved by imposing constraints on the estimates. Srinivasan *et al.* imposed inequality constraints among partworth estimates within attributes. These constraints were based on *a priori* desirability orders for the levels of attributes and ensured that partworth estimates are consistent with these *a priori* orders. If *a priori* orders are valid then the imposition of constraints will bring the partworth estimates closer to their true value and this will improve the predictive validity of these estimates. For instance, Srinivasan *et al.* assumed that consumer preferences for checking accounts offered by banks can only increase with the hours of operation. Compared to unrestricted partworth estimates, they obtained better predictions with restricted partworth estimates for the full-profile method but not for the tradeoff matrix approach. (Predicative validity was defined at the level of an individual respondent in two ways: *i*) the ability to predict the rank order of preferences for a set of holdout stimuli, and *ii*) the ability to predict the most preferred stimulus in that set.)

As an alternative to *a priori* desirability orders of attribute levels that are specified by the researchers, Srinivasan *et al.* suggested that self-explicated desirability orders could also be used to impose within-attribute constraints. This suggestion was followed and extended by van der Lans and Heiser. They derived both within- and across-attribute constraints using the self-explicated utility model (Huber, 1974). The self-explicated utility model implies that, for instance, the partworth $h_{(40,\ \text{HOURS OF OPERATION})}$ for 40 hours of operation per week is the product of the desirability value $e_{(40,\ \text{HOURS OF OPERATION})}$ for that level and the importance value $v_{(\text{HOURS OF OPERATION})}$ for that attribute, or:

$$h_{(40,\ \text{HOURS OF OPERATION})} = v_{(\text{HOURS OF OPERATION})} e_{(40,\ \text{HOURS OF OPERATION})}. \quad (1)$$

More generally, the self-explicated utility model implies that for any level i of attribute j:

$$h_{ij} = v_j e_{ij} \quad (2)$$

Within- and across-attribute constraints (constraints consisted of inequality and/or equality constraints) upon partworth estimates were obtained by: *i*) within- and across-attribute constraints among estimates for the e_{ij}'s, and *ii*) across-attribute constraints among estimates for the v_j's.

Constraints among estimates for the e_{ij}'s were based upon self-explicated attribute level desirability ratings, and constraints among estimates for the v_i's were based upon self-explicated attribute importance ratings. Again, if the self-explicated utility model and the *a priori* orders from self-explicated desirability and self-explicated importance ratings are valid, then the imposition of constraints will bring the partworth estimates closer to their true values and this will improve the predictive validity. Van der Lans and Heiser compared the predictive validities of *i*) partworth estimates under within-attribute constraints, based upon the self-explicated desirability ratings only, *ii*) estimates under within- and across-attribute constraints, and *iii*) unconstrained estimates, for the full-profile method. Like Srinivasan *et al.* (1983), they obtained better predictions for constrained than for unconstrained estimates. Also, the addition of across-attribute constraints resulted in better or equal predictive validities compared with using only within-attribute constraints. (Predictive validities were measured both at the individual level, in terms of cross-validated Pearson correlations and correct first-choice predictions for a set of holdout stimuli, and at the aggregate level, based on mean absolute errors in first-choice share predictions.)

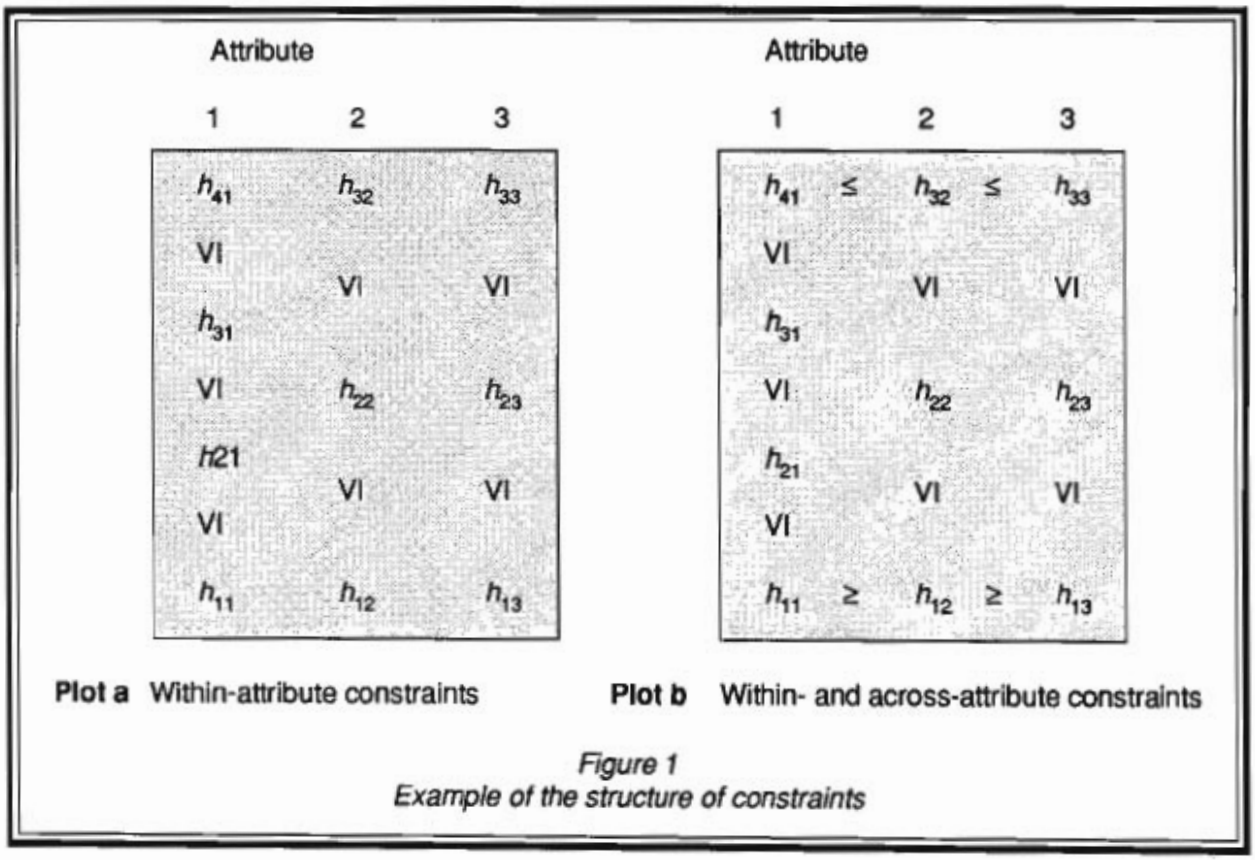
In this paper we derive respondent-specific within- and across-attribute constraints from the self-explicated data in ACA (ACA System by Sawtooth Software). Using both ACA and full-profile data collected by Huber *et al.* (1991), three sets of partworth estimates — unconstrained, within-attribute constrained, and within- and across-attribute constrained — are compared with respect to their predictive validity. Although there has recently been a lot of interest in the comparative predictive validity of ACA and full profile (see Agarwal and Green, 1991; Finkbeiner and Platz, 1986; Green *et al.*, 1991; Green and Srinivasan, 1990; Huber *et al.*, 1991, 1993; Johnson, 1991), we are not aware of studies comparing the two under constrained estimation. In addition to comparing predictive validities, we also determine the effects of the constraints on the existence/magnitude of a number of levels effect in derived attribute importances (Wittink *et al.*, 1991). We present conclusions and end with suggestions for ways in which conjoint analysis can be improved and with suggestions for further research.

STRUCTURE OF THE WITHIN- AND ACROSS-ATTRIBUTE CONSTRAINTS

In a typical ACA interview, respondents are first asked to rank order the levels of each attribute separately according to desirability. If the researcher assumes an *a priori* desirability order among the levels for all respondents, this rank order is not obtained from the respondents. Subsequently, individuals are asked to give importance ratings of the *differences* between the most desirable and least desirable levels per attribute on a four-point rating scale. From the self-explicated desirability orders, we derive inequality constraints among partworth estimates for levels within each attribute, as suggested by Srinivasan *et al.* These are the within-attribute constraints.

From the self-explicated importances, we derive inequality constraints among *differences* between partworth estimates for the most desirable and the least desirable levels per attribute. These are the across-attribute constraints. An example may serve to clarify the structure of constraints. Suppose that a respondent gave the following desirability orders (from most desirable to least desirable): ATTRIBUTE 1 (level 4, level 3, level 2, level 1), ATTRIBUTE 2 (level 3, level 2, level 1), and ATTRIBUTE 3 (level 3, level 2, level 1). The within-attribute constraints on the partworth estimates are then given in Figure 1, Plot a. In Figure 1 $h_{41} \geq h_{31} \geq h_{21} \geq h_{11}$, shown in the first column, means that the partworths for attribute 1 are constrained to be consistent with this indicated pattern of inequalities. Thus, if attribute

1 is the price of a product, the constraints prevent a higher price from having a higher partworth than what occurs for a lower price.



Plot a Within-attribute constraints **Plot b** Within- and across-attribute constraints

Figure 1
Example of the structure of constraints

If in addition to these desirability orders, the respondent gave importance ratings of "1" to ATTRIBUTE 1, of "2" to ATTRIBUTE 2, and of "3" to ATTRIBUTE 3, we define the within- and across-attribute constraints depicted in Figure 1, Plot b. In plot b the row constraints prevent the partworths of the extreme (best and worst) levels from being inconsistent with the self-explicated (or *a priori* assumed) importances. If the self-explicated importances are equal for two attributes, no constraints are imposed on the conjoint-based importances.

The across-attribute constraints upon differences between partworths of the extreme levels correspond to stretching desirability values per attribute to ensure equal ranges before multiplying them in the self-explicated utility model with importance values as recommended by Srinivasan (1988), but advised against by Green *et al.* (1991). It is important to note that we are forced to impose constraints upon differences between products of desirability values and importance values, due to the incommensurability of the desirability responses across attributes. It is also important to note that the structure of within- and across-attribute constraints we propose is different from the one imposed by van der Lans and Heiser, who do not stretch desirability values per attribute before multiplying them with importance values.

Clearly, whether or not constrained estimation (under within-attribute constraints or within- and across-attribute constraints) will improve predictions depends upon the validity of the desirability orders and the order of the importance ratings.

Interestingly, the proposed across-attribute constraints are also likely to reduce the number of levels effect in conjoint analysis (Currim *et al.*, 1981; Green and Srinivasan; Wittink *et al.*, 1982, 1989, 1991). This effect implies that when intermediate levels are added to the levels of an attribute in a conjoint design, this attribute will get a higher relative importance than when these intermediate levels would not have been added, other things being equal. Wittink *et al.* (1991) found a significant number of levels effect with both the full-profile method and ACA, the effect size for the full-profile method being about twice as large as for ACA. However, they did not find a significant number of levels effect for self-explicated importances. Thus, the across-attribute constraints would not be dependent upon the number of levels and any effects in the conjoint-based importances will be reduced by these constraints. ACA already incorporates the self-explicated data in the estimation of partworths. This is one reason why the number of levels effect is larger for the full-profile method than for ACA, and we may therefore expect a larger reduction in the level effect for the full-profile method, by imposing across-attribute constraints, than for ACA. For a more detailed discussion of the levels effect, see Wittink *et al.* (1991).

METHOD

To compare the predictive validity of unconstrained, within-attribute constrained, and within- and across-attribute constrained partworth estimates, for both ACA and the full-profile method the data from Huber *et al.* (1991) are used. These data consist of responses of 400 respondents to refrigerators, refrigerator attributes, and refrigerator attribute levels. The attributes and attribute levels used are given in Table 1.

Table 1 Refrigerator attributes and levels

A. Brand Name	General Electric; Sears/Kenmore; Whirlpool
B. Capacity	19; 20*; 21*; 22 (cubic feet)
C. Energy Cost	$70; $80*; $90*; $100 (annual)
D. Compressor	Extremely quiet; somewhat quiet*; somewhat noisy*; extremely noisy
E. Price	$700; $850*; $1,000*; $1,150
F. Design	Freezer on left (side by side); Freezer on top
G. Warranty	1 year; 3 years
H. Refrigerant	Soft CFC (environmentally safe); Chlorofluoro-hydrocarbon (hurts environment)
I. Dispenser	Dispenses ice and water through the door; No door dispenser for ice or water

* In full profile, half the respondents saw all four levels for attributes B (Capacity) and E (Price), and only the extreme levels for attributes C (Energy Cost) and D (Compressor). The other half saw four levels for C and D, but only the extremes for B and E. In ACA, attribute D and E were each given at four levels for half the respondents, and attributes B and C had four levels for the other half.

Respondents were interviewed at super-regional malls in 11 cities. Prior to actual interviewing they were screened for being over 18 and having refrigerators in their homes, and promised $5 each for completing the interview. Each respondent provided both full-profile ratings (on a 9-point likelihood to purchase scale) and ACA judgments (self-explicated data as before, and graded paired comparisons on a 9-point scale). Half the respondents completed the full-profile task first whereas the other half completed the ACA task first. Both tasks were administered by computer in one sitting. In the ACA task, within-attribute desirability orders for levels of capacity, energy cost, compressor, and price, were assumed to be known *a priori* and to be common across respondents. Thus, no desirability orders were asked for these attributes.

Other experimental manipulations in Huber *et al.* (1991) are: 1) half the respondents saw nine attributes and half the respondents saw five attributes (A,B,C,D,E), 2) the number of levels was varied for attributes B, C, D, and E, as indicated in the footnote of Table 1, and 3) the order in which attributes were listed in the full-profile method was varied. Respondents who saw nine attributes gave 16 full-profile ratings and 16 graded paired comparison ratings (ten pairs differing on two attributes and six pairs differing on three attributes). Respondents who saw five attributes gave 16 full-profile ratings and 12 graded paired comparison ratings (ten pairs differing on two attributes and two pairs differing on three attributes). The order in which attributes appeared in the self-explicated part of the ACA task was not manipulated, but was equal to one of the orders of the attributes in the full-profile task. In ACA's graded paired comparisons task, the order in which attributes were listed varied between pairs of profiles.

In addition to the full-profile task and the ACA task, respondents were asked to complete a choice task in which they indicated their most preferred alternative in two pairs and two triples of refrigerator profiles. Respondents were also asked to indicate their least preferred alternative in the two triples. The profiles were defined in terms of the five attributes and the 13 levels that were common to all respondents in the full-profile and ACA tasks. Profiles within each pair/triple were chosen such that it could be expected that no profiles were dominated on all attributes by other profiles. The choice task was administered twice, once after the first preference elicitation task and again after the second preference elicitation task.

For each of 385 respondents (7 respondents provided incomplete data and another 8 gave equal ratings to all full profiles and/or pairs) we computed nine sets of least squares partworth estimates. Three sets were based on the full-profile ratings, three on both the ACA priors and the graded paired comparisons, and three on the graded paired comparisons only. The latter three sets were included to compare the way in which ACA uses the self-explicated data with a theoretically more justifiable way of using only the order information in the self-explicated data. We used the MORALS-algorithm (Young, De Leeuw and Takane, 1976) to compute unconstrained partworth estimates, and within-attribute constrained partworth estimates for the full-profile method. All other partworth estimates were computed by an alternating least squares procedure given by van der Lans (1992, Chapter 3; see also van der Lans, 1991). To reduce the likelihood of solutions that correspond to a local minimum, we computed solutions from multiple starts and retained the best ones. The implementation of constrained estimation requires additional considerations. Two or more attributes could obtain the same importance rating. These ties were treated by the so-called primary approach to ties (Kruskal, 1964). That is, differences between partworths of extreme levels were not constrained to be equal in the final conjoint solution for attributes that obtained equal self-explicated importances. The primary approach to ties was chosen because the coarse-grained importance

rating scale could be expected to result in a large number of ties, and van der Lans and Heiser found better predictive validity for the primary approach to ties even with a finer-grained importance scale.

The different sets of partworth estimates were used: 1) to predict individual choices in the (replicated) choice task, 2) to predict aggregate choice shares via the first choice rule and the multinomial logit rule, and 3) to investigate the number of attribute levels effect.

RESULTS

Individual-Level Choice Prediction

In Table 2 we show proportions of correctly predicted individual choices. For computing the entries in Table 2, each triple was converted to three pairs (except for the last subtable). Thus, with two pairs and two triples each evaluated twice we have sixteen pairs per respondent. The successive columns show the proportions of correctly predicted individual choices (hit rates) for unconstrained, within-attribute constrained, and within- and across-constrained partworth estimates. The last column gives the highest proportion across the three sets of estimates. The highest proportion (based on three digits precision) within each row is underlined.

Part A of Table 2 gives the overall proportions. In the rightmost column we see that the best set of estimates for each method yields approximately equal proportions correctly predicted choices. However, without constraints ACA provides the highest predictive validity. Its validity does not improve when constraints are imposed. On the other hand, Full Profile and Paired Comparisons Only do gain from constrained estimation under both within-attribute constraints and within- and across-attribute constraints. The low proportion for unconstrained partworth estimates from Paired Comparisons Only is due to the nonorthogonality of the paired comparisons design which results in a large variance of the partworth estimates. This variance is reduced by the imposition of constraints.

In Part B of Table 2 hit rates are decomposed according to the number of attributes that the respondents saw. Looking at the highest hit rate within each row, we see that, somewhat surprisingly, ACA and Paired Comparisons Only do better than Full Profile with five attributes, but not with nine. This seems to contradict (at least up to about ten attributes) the suggestion by Huber *et al.* (1991) that ACA should become increasingly attractive as the number of attributes to be included in the study increases. However, they did not consider the benefit of imposing constraints. Without constraints, ACA is superior for both five and nine attributes. A tentative explanation for the result would be that ACA's graded paired comparisons design matrices were more efficient for estimating the tradeoff between the attributes involved in the choice sets with five attributes than with nine attributes. This might also explain why with five attributes the predictive validities of ACA and Paired Comparisons Only become slightly worse under within- and across-attribute constraints compared to no constraints and within-attribute constraints. Unlike the case with nine attributes, for five attributes the tradeoff between attributes may already be estimated precisely, and the constraints (which contain some error) only worsen the predictive validity. On the other hand, for Full Profile with nine attributes, compared to with five attributes, the effect of *i*) this increased task complexity, and *ii*) the lower ratio of observations to parameters, on the predictive validity, seems to be counteracted by the imposition of constraints. Furthermore, with nine attributes within- and across-attribute constraints do better than within-attribute constraints.

Table 2 Individual-Level Choice Predictions (Proportion Predicted Correctly)				
	No Constraints	**Within-Attribute Constraints**	**Within- and Across-Attribute Constraints**	**Highest**
A: Overall				
Full Profile	.68	.71	.72	**.72**
ACA	.73	.73	.72	**.73**
Paired Comparisons Only	.66	.73	.72	**.73**
B: Number of Attributes				
		Nine Attributes		
Full Profile	.65	.69	.72	**.72**
ACA	.71	.71	.70	**.71**
Paired Comparisons Only	.61	.70	.71	**.71**
		Five Attributes		
Full Profile	.71	.72	.72	**.72**
ACA	.76	.76	.74	**.76**
Paired Comparisons Only	.70	.75	.73	**.75**
C: Consistency				
		≤4		
Full Profile	.57	.56	.56	**.57**
ACA	.59	.60	.59	**.60**
Paired Comparisons Only	.57	.59	.60	**.60**
		5 or 6		
Full Profile	.59	.64	.66	**.66**
ACA	.69	.69	.68	**.69**
Paired Comparisons Only	.62	.69	.68	**.69**
		7		
Full Profile	.69	.71	.72	**.72**
ACA	.74	.73	.72	**.74**
Paired Comparisons Only	.67	.72	.72	**.72**
		8		
Full Profile	.78	.81	.83	**.83**
ACA	.82	.82	.80	**.82**
Paired Comparisons Only	.71	.82	.79	**.82**

Part C of Table 2 decomposes the hit rates according to the respondents' consistency in the replicated choice task. Each respondent chose from two pairs and two triples twice. By decomposing the choices with each triple into three pairs, this yielded eight replicated choices. Both ACA and Paired Comparisons Only do better than Full Profile when the number of consistent choices is less than 4, or 5 or 6. With eight consistent choices the difference between Full Profile, ACA, and Paired Comparisons Only completely disappears. Apparently, respondents who are not perfectly consistent in their choices gain most from the simplicity of the ACA task. Constrained estimates do not improve over unconstrained estimates for Full Profile when respondents are very inconsistent (≤4). With more than 4 consistent choices constrained estimates do improve for Full Profile. In that case, within- and across-attribute constraints also improve over within-attribute constraints.

Hit rates were also decomposed on the combination of task order (ACA first or Full Profile first) and number of attributes (see Table 2, Part B). Apart from the main effect of the number of attributes, two interesting results were found. First, within-constrained partworth estimates always improve upon unconstrained partworth estimates (difference of about five percent correctly predicted choices) for Full Profile except when ACA comes first and respondents see only five attributes. Apparently, ACA serves as a warm-up by which respondents become able to judge full profiles based on five attributes in a manner that is consistent with the desirability orders on the attribute levels. Secondly, within- and across-attribute constraints improve upon within-attribute constraints for Full Profile only when Full Profile comes first and the number of attributes is nine (difference of four percent correctly predicted choices). Apparently, only when no warm-up task precedes the full-profile judgments and the number of attributes is relatively high do the respondents seem to be less accurate in making the tradeoff between attributes in their full-profile judgments. This latter result differs from the result in van der Lans and Heiser, where within- and across-constraints improve upon within-constraints for Full Profile, given six attributes and given that full-profile judgments were preceded by self-explicated desirability and importance ratings.

In Part A of Table 3 we distinguish between the proportions of correctly predicted choices for pairs and triples. As suggested by Huber *et al.* (1991), two reasons for examining the triples separately are *i*) the fact that real world choices are usually not restricted to pairs, and *ii*) ACA's paired comparisons intensity ratings may be more similar to the pairs in the choice tasks than the triples. We see that the large difference in hit rates between Full Profile and ACA for the pairs from triples found by Huber *et al.* (1991) (with unconstrained partworth estimates), becomes very small when partworth estimates for Full Profile are constrained. And, except for when there are no constraints, hit rates for Paired Comparisons Only are virtually equal to the hit rates for ACA.

Table 3 Individual-Level Choice Predictions Split by Type of Choice (Proportion Predicted Correctly)

	No Constraints	Within-Attribute Constraints	Within- and Across-Attribute Constraints	Highest
A: Choice Task				
		Pairs from Triples		
Full Profile	0.67	0.71	0.72	**0.72**
ACA	0.74	0.74	0.72	**0.74**
Paired Comparisons Only	0.65	0.73	0.72	**0.73**
		Pairs from Pairs		
Full Profile	0.70	0.70	0.71	**0.71**
ACA	0.72	0.72	0.71	**0.72**
Paired Comparisons Only	0.66	0.72	0.71	**0.72**
B: Triples				
		Most Likely		
Full Profile	0.52	0.59	0.61	**0.61**
ACA	0.64	0.65	0.62	**0.65**
Paired Comparisons Only	0.53	0.63	0.61	**0.63**
		Least Likely		
Full Profile	0.57	0.60	0.61	**0.61**
ACA	0.63	0.63	0.61	**0.63**
Paired Comparisons Only	0.52	0.62	0.61	**0.62**

Part B of Table 3 shows the results for the choices from triples separately for choosing the "most likely" and the "least likely" to-be-purchased profile from the triple. Note that the reduced percentages for most and least likely choices (relative to the pairs) simply result from the greater difficulty in predicting choices from triples than from pairs. Differences between the three methods are larger for most likely choices than for least likely choices. Of course, the most likely choices are the ones for which predictions are most critical. Constrained partworth estimates for Full Profile give the largest improvement compared to unconstrained partworth estimates for most likely choices.

Aggregate-Level Choice Share Prediction

In Table 4 mean absolute errors in choice share predictions are given for each method under each type of constraint. Predicted and actual choice shares were computed for each choice (from two pairs and two triples, that is, for a total of ten profiles) across respondents and across the replication of the choice task. Predicted choices were determined via the first choice rule and the multinomial logit rule. For the multinomial logit rule, slope parameters were computed first by maximizing the log likelihood of the multinomial logit model across respondents, choice sets, and the replication. In the

multinomial logit model, the chance $p_{(k)is}$ that respondent k will choose profile i from choice set s is given by:

$$p_{(k)is} = \frac{e^{bV_{(k)i}}}{\Sigma_{j \varepsilon s} e^{bV_{(k)j}}}, \qquad (3)$$

in which $V_{(k)i}$ gives the predicted overall utility that profile i has for respondent k. One benefit of the multinomial logit rule is that it adds random error to the predicted overall utilities which may improve choice share predictions when the noise in the choices is greater than the noise in the predicted conjoint-based overall utilities for the profiles (Elrod and Kumar, 1989).

Looking at Table 4 we see that Full Profile within-attribute constrained partworth estimates with the first choice rule give by far the lowest mean absolute error in predicted choice share. Interestingly, the absolute error in predicting choice shares increases when random error is added for Full Profile, whereas for ACA it decreases. This result can be explained by using the following argument: It seems that overall utilities predicted from Full Profile contain more noise than the choices. Adding random error to the predicted utilities can only increase the mean absolute error in predicted choice shares. Furthermore, it seems that for Full Profile, within attribute-constraints reduce the noise in the predicted utilities to the level of the noise in the choices, and within- and across-attribute constraints reduce the noise in the predicted overall utilities even further (below the level of noise in the choices). On the other hand, the overall utilities predicted from ACA seem to contain less noise than the choices. The fact that adding random error does not lower the mean error further than 5.6 may be due to systematic differences in ACA judgments and holdout choices. Clearly, systematic differences cannot be eliminated by adding random error. Adding constraints in ACA reduces the noise even further below the noise present in the choice data. Note that this explanation, and the somewhat counterintuitive finding that the method that does better at individual-level choice prediction (ACA) does worse at aggregate choice share predictions, is consistent with Hagerty's (1986) results. His results imply that for individual-level predictions noise in partworth estimates is quite important, whereas for aggregate-level predictions the noise tends to cancel itself out across respondents. Results for Paired Comparisons Only seem somewhat erratic. However, they can be explained if unconstrained partworth estimates for Paired Comparisons contain less noise than unconstrained partworth estimates for Full Profile.

Table 4 Mean Absolute Error in Aggregate-Level Choice Share Prediction (%)				
	No Constraints	**Within-Attribute Constraints**	**Within- and Across-Attribute Constraints**	**Lowest**
		First Choice Rule		
Full Profile	5.3	1.7	3.8	**1.7**
ACA	10.4	11.5	12.2	**10.4**
Paired Comparisons Only	4.1	10.0	11.3	**4.1**
		Multinomial Logit Model		
Full Profile	10.0	6.7	6.4	**6.4**
ACA	5.6	5.9	6.2	**5.6**
Paired Comparisons Only	12.5	6.9	14.4	**6.9**

Number of Attribute Levels Effect

The plots in Figure 2 show the number of attribute levels effect for the attributes whose numbers of levels were manipulated. As in Wittink *et al.* (1991), importances were first normalized per respondent such that the importances of the five common attributes sum to 100. We see that the different sets of partworth estimates yield different effect sizes. Unconstrained partworth estimates for Full Profile and Paired Comparisons Only display the largest number of level effects throughout. ACA shows much smaller effects.

Multivariate analyses of variance were run to test the significance of the two-way interactions between number of levels and type of constraints, and between number of levels and method, as well as the three-way interaction, upon average importance. The number of levels x type of constraints interaction has a significant effect for all attributes ($p<.01$). The number of levels x method interaction has a significant effect for energy ($p<.05$) and price ($p<.05$). The three-way interaction is significant for energy ($p<.05$), capacity ($p<.01$), and price ($p<.01$), but not for compressor noise.

The large effects for unconstrained partworth estimates in Full Profile and Paired Comparisons Only are reduced considerably by imposing within-attribute constraints, except for compressor noise and price with Full Profile. Thus it seems that the orders of compressor noise levels and price levels are less often violated than the orders of energy cost levels and capacity levels. Apparently, price is not confounded with quality as Srinivasan *et al.* (1983) warned. Adding across-attribute constraints reduces the number of attribute levels effect about equally for all three methods.

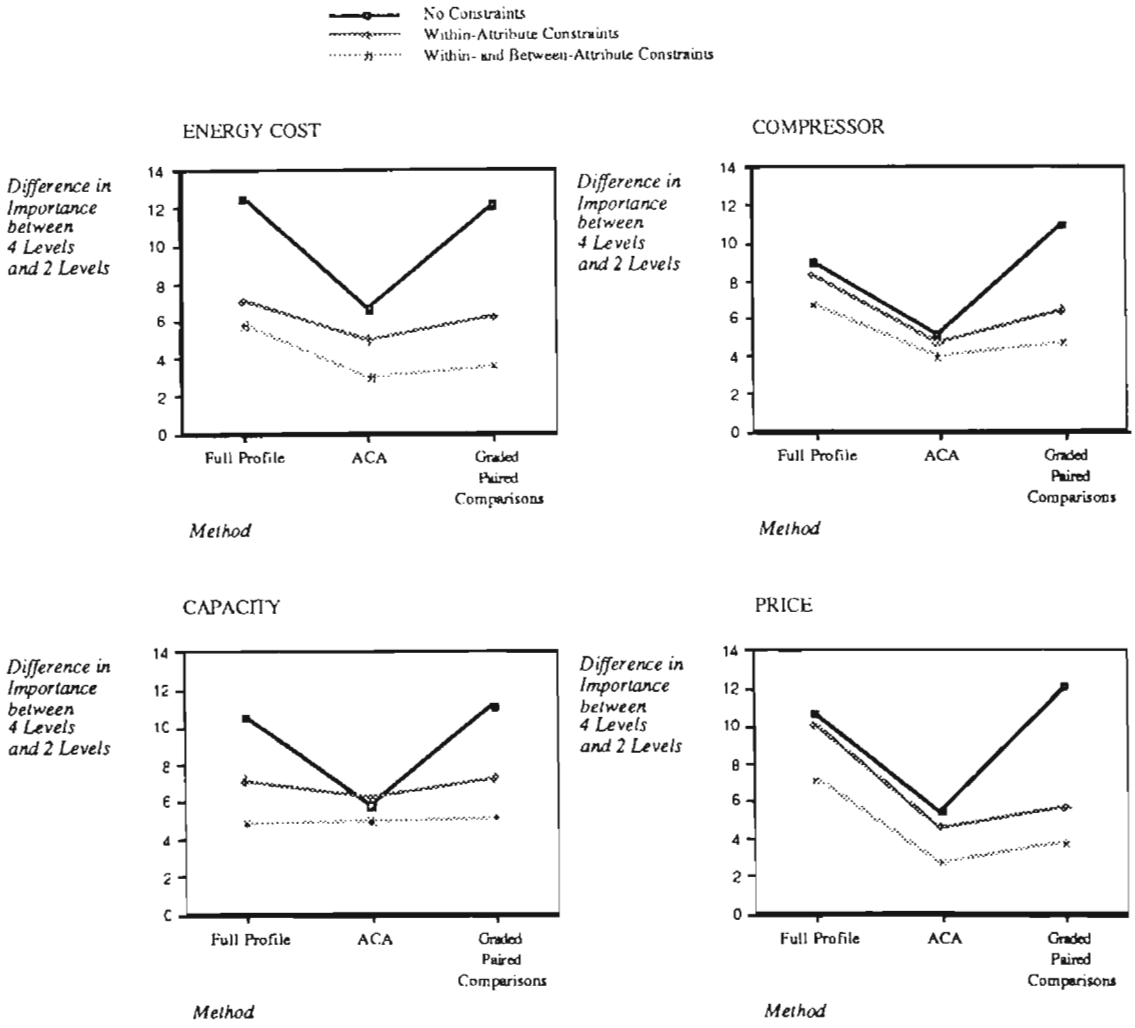


Figure 2
Number of attribute level effects for different attributes, methods, and constraints

CONCLUSIONS

The differences in predictive validity for unconstrained ACA and Full Profile conjoint analysis are substantially reduced when constraints are imposed on the partworths. For individual-level choice predictions, Full Profile, ACA, and Paired Comparisons Only are about equally valid when within- and across-attribute constrained partworth estimates are used for Full Profile; unconstrained or within-constrained partworth estimates are used for ACA; and within-constrained partworth estimates are used for Paired Comparisons Only. ACA and Paired Comparisons Only outperform Full Profile

when *i*) the Full Profile task comes first and the number of attributes is five, and *ii*) when respondents have low consistency in their choices. The almost equal predictive validities of ACA and Paired Comparisons Only seem to indicate that the way in which ACA combines the self-explicated priors with the paired comparisons data accomplishes the same result as the theoretically more justifiable idea of imposing constraints (at least for individual-level predictions).

For aggregate choice share predictions we have used both the first choice rule and the multinomial logit model. Although the differences between the two choice rules, the different conjoint analysis methods, and the different constraints can be explained by results from Elrod and Kumar and Hagerty, we think that our results should be interpreted carefully until more evidence favouring either of the two choice rules is found. Nevertheless, the much higher mean absolute error in choice share prediction for Full Profile compared to ACA when using unconstrained partworth estimates (see, Huber *et al.*, 1993) is considerably reduced when imposing constraints upon the partworth estimates for Full Profile. The difference between the effect of constraints in Full Profile and Paired Comparisons Only might be due to the fact that the design matrix for the paired comparisons is nonorthogonal. As such, imposing constraints upon some partworth estimates may alter other partworth estimates for levels that are not involved in the constraints. It is conceivable that this causes a deterioration in the predictive validity.

As expected, the imposition of constraints reduces the number of levels effect considerably for Full Profile and Paired Comparisons Only, and only slightly for ACA.

Further research could consider the use of finer-grained importance scales and commensurable desirability scales across attributes as suggested by Green, Krieger, and Agarwal. Finer-grained importance scales seem especially desirable because of the finding of van der Lans and Heiser that the primary approach to ties does better than the secondary approach, even with very fine-grained importance scales.

REFERENCES

Agarwal, Manoj K. and Paul E. Green (1991). "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results." *International Journal of Research in Marketing*, 8, 141-146.

Currim, Imran S., Charles B. Weinberg and Dick R. Wittink (1981). "The Design of Subscription Programs for a Performing Arts Series." *Journal of Consumer Research*, 8 (June), 67-75.

De Leeuw, Jan, Forest W. Young and Yoshio Takane (1976). "Additive Structure in Qualitative Data: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika*, 41 (December), 505-29.

Elrod, Terry and S. Krishna Kumar (1989). "Bias in the First Choice Rule for Predicting Share." *Sawtooth Software Conference Proceedings*, 259-271.

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

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

Green, Paul E. and Catherine M. Schaffer (1991). "Importance Weight Effects on Self-Explicated Preference Models: Some Empirical Findings." *Advances in Consumer Research*, 18, Association for Consumer Research.

Green, Paul E. and V. Srinivasan (1990). "Conjoint Analysis in Marketing Research: New Developments and Directions." *Journal of Marketing*, 54 (October), 3-19.

Hagerty, Michael R. (1986). "The Cost of Simplifying Preference Models." *Marketing Science*, 5, 298-319.

Huber, George P. (1974). "Multi-Attribute Utility Models: A Review of Field and Field-like Studies." *Management Science*, 20, 1393-1402.

Huber, Joel C., Dick R. Wittink, John A. Fiedler, and Richard L. Miller (1991). "An Empirical Comparison of ACA and Full Profile Judgments." *Sawtooth Software Conference Proceedings*, 189-202.

Huber, Joel C., Dick R. Wittink, John A. Fiedler, and Richard L. Miller (1993). "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice." *Journal of Marketing Research*, forthcoming.

Johnson, Richard M. (1991). "Comment on "Adaptive Conjoint Analysis: Some Caveats and Suggestions." *Journal of Marketing Research*, 28 (May), 223-5.

Kruskal, Joseph B. (1964). "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika*, 29 (March), 1-27.

Srinivasan, V. (1988). "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences." *Decision Sciences*, 19, 295-305.

Srinivasan, V., Arun K. Jain and Naresh K. Malhotra (1983). "Improving Predictive Power of Conjoint Analysis by Constrained Parameter Estimation." *Journal of Marketing Research*, 20 (November), 433-8.

van der Lans, Ivo A. (1991). "Optimal Scaling under Order Restrictions upon Parameters and (Differences between) Products of Parameters." paper presented at the Third Conference of the International Federation of Classification Societies, Edinburgh, Scotland, August.

van der Lans, Ivo A. (1992, in preparation). "Nonlinear Multivariate Analysis for Multiattribute Preference Data." Dissertation, University of Leiden, The Netherlands.

van der Lans, Ivo A. and Willem J. Heiser (1990). "Constrained Part-Worth Estimation in Conjoint Analysis using the Self-Explicated Utility Model." Research Report RR 90-02, Department of Data Theory, University of Leiden, The Netherlands.

Wittink, Dick R., Joel C. Huber, John A. Fiedler and Richard L. Miller (1991). "The Magnitude of and an Explanation/Solution for the Number of Levels Effect in Conjoint Analysis." Working Paper, Cornell University at Ithaca, NY, August.

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 Effects of Differences in the Number of Attribute Levels on Conjoint Results." *Marketing Letters*, 1, 113-23.

Young, Forest W., Jan De Leeuw and Yoshio Takane (1976). "Regression with Qualitative and Quantitative Variables: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika*, 41 (December), 505-29.

Comment on van der Lans, Wittink, Huber, and Vriens

Robert Zimmermann
Maritz Marketing Research

This paper is an extension to the analysis of a paper presented at the last Sawtooth conference, which compared the validity of full profile and ACA conjoints, manipulating a number of additional dimensions. The original study was carefully designed and executed and warrants this extended analysis. The purpose of this paper is to evaluate the impact that prior constraints on utilities would have on the validity of conjoint in general and on the relative validity of ACA and full profile in particular.

The research method used introduces two potential sources of bias. First, the method used to present full profile involved rating profiles in terms of intent to buy, and so results might not generalize to studies in which the profiles are ranked or sorted. Second, the prior constraints are based on the self-explicated portion of the ACA data. Since these data are already incorporated in computing utilities in ACA, it might be expected that the prior constraints would have less impact on ACA than on full profile.

Also, while the study purports to deal with predictive validity, the criterion is collected in the same interview as is the conjoint data. Predictive validity should be limited to situations in which there is a real world criterion measure, in this case actual purchase behavior or observed market share. The current study is closer to construct validity than predictive validity.

I take two important points from these analyses. First, the validity of ACA shows virtually no improvement with the use of prior constraints. This increases my confidence in the robustness of the methods used in ACA to compute utilities. Second, full profile, as administered in this study, does show improved reliability with the use of prior constraints, and I will seriously consider using prior constraints when I use full profile with the rating method.

There is one aspect of the design of this study that I would like to discuss in detail. The paired validation profiles were selected in advance to be similar. The assumption was that very dissimilar profiles created easy choices which were unrealistic in a real setting, which inflated validity estimates, and most pertinently, which would not be sensitive to the different treatment effects.

However, it is useful to consider what would have happened if the researchers were actually capable of selecting profile pairs such that each subject attached equal utility to both members of the pair. These would also represent choices that were of no practical or theoretical interest, since there would be no basis for choosing. Validity in this case would be markedly depressed, in fact, not significantly different from chance. And lastly, these pairs would also be insensitive to treatment manipulations.

Fortunately, such prescience is very unlikely, since individuals vary widely in their valuative structures. This variation, however, presents a useful opportunity for extending the analysis. The two paired- and three triple-validation profiles can be analyzed as 8 paired choices, each associated with a utility difference estimated from the twin conjoint exercises. This yields for the

total sample over 3000 choice pairs, most likely providing a wide range of similarity or dissimilarity in utility value between pairs.

It should be possible to separate the 3000+ individual pairs into levels of difficulty, based on individual utility similarities. Easy choices should show high levels of validity and little difference in validity between full profile and ACA. Difficult choices should show low validity and again little difference between the two conjoint methods. Choices of intermediate difficulty would show intermediate levels of validity, and most importantly, the observed differences between ACA and full profile would be focused here, and would be quite large. In addition, it would be for these choices that prior constraints would have the most impact.

I would contend that it is in this intermediate range choice difficulty that conjoint is most powerful and most useful, and that it is in this range that the effects of this study should be most pronounced.

This raises an interesting side issue. In the paired comparison portion of ACA, an attempt is made to avoid profiles of very different utility levels. It is possible that the pairing process would be more efficient if profiles of very similar utilities were also avoided. Very similar pairs not only do not produce very reliable information, but respondents find them frustrating, which can lead to arbitrary response patterns.

CROSS-TASK COMPARISON OF RATINGS-BASED AND CHOICE-BASED CONJOINT

Karen Oliphant
Apple Computer
Thomas C. Eagle
Decision Research
Jordan Louviere
University of Utah
Don Anderson
University of Wyoming

ABSTRACT

This paper discusses differences in ratings-based (RB) and choice-based (CB) conjoint procedures. An empirical comparison is made of the two procedures in a real marketing research environment. The comparison is based on aggregate parameters, predictions to holdout choice sets and tests for scale differences between data sets. Results indicate little difference between individual-level RB models adjusted to take non-choice into account and generalized multinomial logit CB models. Unfortunately, the comparison is weakened by virtue of the fact that the empirical problem involved a preponderance of attributes whose directional effects could be anticipated *a priori*. Thus, consistent with results in the linear models in decision-making literature, both RB and CB models predicted well to holdout choice sets. As a consequence of this experimental weakness, we make suggestions for future research comparisons that can avoid these guaranteed results.

INTRODUCTION

Although ratings-based (RB) conjoint has been widely studied and applied in marketing for well over a decade, choice-based (CB) conjoint is relatively new, and has seen less academic and practical application. Recently, Green and Srinavasan (1990) and Batsell and Louviere (1991) have called for more research comparing RB and CB conjoint. The purpose of this paper is to undertake a limited comparison of traditional, full-profile RB conjoint with CB conjoint. By "limited" we mean that the comparison itself is limited to one context and holdout choices. We view such a comparison as limited because there are many other possible bases for comparison that are not explored in this paper, such as external or predictive validity comparisons. Our purpose is to compare the cross-task validity of both techniques in circumstances in which there are common and unique holdout choice sets.

STATEMENT OF THE PROBLEM

A division of the Allstate Insurance Company was considering the introduction of a new emergency road service package involving specific emergency and camping benefits. A series of focus group discussions were used to explore the concept and develop a list of potentially salient benefits for such a new service. The focus groups were followed by a quantitative survey to further reduce the list of benefits. These preliminary research efforts were followed by a second quantitative study whose objective was to identify potentially salient configurations of benefits, and price the package at a level that would maximize share in a relatively short time frame. An important strategic issue in the project was not only to predict likely market share consequences for a new Allstate entry, but also to examine the effects of the new entrant on existing offerings and to forecast how competitive reactions on the part of those existing offerings would impact on the likely market shares of any proposed new Allstate offerings.

This problem is typical of many that might be addressed by conjoint analysis techniques. Allstate was interested in comparing RB and CB conjoint; hence, a comparison was designed into the research effort. It is not clear that RB and CB conjoint are perfect substitutes. Our review of the academic and practical marketing literature suggests that there is considerable misunderstanding about CB conjoint; hence, we briefly compare both approaches.

Design Issues. It is probably fair to say that RB conjoint studies are easier to design and analyze than CB studies. RB studies are based on experimental designs for general linear models, which are widely available in design catalogs and computerized design generators such as Conjoint Designer by Bretton-Clark or CONSURV by Intelligent Market Systems. Because ratings are assumed to be approximately equal interval in measurement level, OLS regression is the primary analytical method applied to RB conjoint. OLS software is widely available in many statistical packages and in computerized conjoint software systems like Conjoint Designer or CONSURV. In contrast, CB conjoint studies rely on the family of Multinomial Logit (MNL) models and not general linear models; hence, design strategies for such problems are not widely cataloged or computerized. However, Louviere and Woodworth (1983), Louviere (1988a,b) and Bunch, Louviere, and Anderson (1991) have discussed the construction of such designs from designs for general linear models. The design problem for MNL models is complicated by the fact that one must simultaneously or sequentially design both conjoint profiles and the choice sets into which to put them. As well, choice data are discrete and not amenable to analysis by OLS estimation software; hence, they require special purpose software which has only recently become widely available. Examples of such software include Systat Logit or Ntelogit from Intelligent Marketing Systems, Inc.

Level of Analysis. Another potential difference is that RB conjoint commonly relies on individual-level analyses, although problems involving interactions often require aggregation of respondents. Although CB conjoint is most often associated with sample- or segment-level analyses, individual-level analysis is possible if one uses resource allocation tasks instead of discrete choice responses (Louviere and Woodworth; Louviere 1988a,b). Similarly, it is often said that RB conjoint permits analysts to compare and cluster respondents into benefit segments, but CB conjoint does not. This is not strictly true because CB conjoint studies can be designed such that all respondents receive common choice sets, permitting comparison and clustering. Also, depending on the magnitude of the task, it is possible to calculate individual-level marginal attribute level totals which are analogous to partworths (marginal means) in RB studies (Kaciak and Louviere 1990). However, it is fair to say

that CB conjoint often does not involve individual-level analyses, but rather relies on the incorporation of individual difference factors to account for differences in individuals.

Is "brand" an attribute or an outcome? Another difference lies in the treatment of "brands." In RB conjoint, brands are often varied as the level of a factor called "brand name." If, as is commonly done, main effects only designs are used to implement the study, one is forced to assume that the attribute effects are constant across brands. This may be highly unrealistic in many situations because respondents will associate unobserved attributes with brands, resulting in differential sensitivity to the levels of attributes which were varied. Thus, it may often be the case that attributes *interact* with brand names. CB conjoint, on the other hand, typically treats brand names as choice outcomes rather than attributes of profiles, which allows one to estimate separate, "alternative-specific" attribute effects in CB experiments.

Estimation error. All utility parameters estimated from response data, regardless of estimation method, contain sampling and other errors. These errors are explicitly incorporated into the theory underlying CB conjoint, but are often ignored in applications of RB conjoint. To date no methods for taking errors in RB partworths into account in choice forecasts have been proposed that have a sound basis in a theory of choice behavior. Thus, CB conjoint has a theoretical basis for incorporating errors in choice in forecasts, while RB forecasting methods remain largely *ad hoc*.

Task differences. CB and RB tasks differ. RB tasks involve assignment of ratings to single profiles presented one at a time, or sorted into ordered piles and then compared. Thus, at best, RB conjoint tasks can be thought of as mimicking the evaluation and comparison of a single set of offerings. In contrast, CB conjoint presents respondents with multiple sets of offerings, and asks them to make a single choice from each set or to allocate resources appropriate to the problem across the alternatives in each set.

Handling non-choice. Unlike most applications of RB conjoint, CB conjoint typically includes a "no choice" option in the choice sets evaluated by the respondents. As discussed by Louviere and Woodworth, Batsell and Louviere, and Louviere (1988a,b), "no choice" is essential to estimating demand as opposed to merely share. Thus, without "no choice," conjoint models forecast the probability of choosing a profile given that a choice will be made. The probability of choosing from a product category obviously varies considerably by product category, typically being low for new categories and relatively higher for mature categories.

Handling existing brands. Although RB conjoint can incorporate existing brands by having respondents rate such brands on the same scale and in the same context and task as the designed profiles, this is rarely reported. CB conjoint easily accommodates existing brands in choice studies because they can be included in choice sets as constantly available alternatives or varied as part of the choice design. The advantage of incorporating existing brands is that one doesn't have to assume that all respondents agree on particular profiles for certain brands, but can simply represent the brand by its own utility function.

Number of profiles observed. Finally, but not exhaustively, CB conjoint studies typically observe many more profiles than RB studies. Indeed, in the case of main effects only RB designs, a typical CB study will include at least — and usually more than — twice as many profiles. Thus, respondents evaluate and consider many more profiles, making it possible to explore much more of the response

surface. The advantage of this is that CB studies often make it possible to incorporate many attribute interaction effects and/or effects that capture differential cannibalization and switching effects among brands.

Given the similarities and differences in RB and CB conjoint, it is of interest to compare them in an empirical setting. Recently, Elrod, Louviere, and Davey (1992) provided a comparison, which was limited to a small number of attributes and student subjects. It is of interest to compare the techniques in a more realistic problem setting with actual consumers. Also, as noted by Elrod *et al.*, the problem they studied was one in which few differences were expected between the techniques, and the results were consistent with this expectation. Thus, we wished to compare the techniques in a richer problem environment in which we might expect some differences.

RESEARCH APPROACH

As discussed by Elrod and Kumar (1989), RB conjoint should predict choice well when there is little error in both the conjoint utilities and the observed choices to be predicted. Elrod, Louviere, and Davey (1992) studied only four attributes, three of which had obvious directional effects prior to the study (for example, the utility of price should decline with increasing price). The present study involves, respectively for RB and CB conjoint, ten or nine attributes. In the case of RB, the extra attribute has four brand name levels, while in the case of CB the four brands are choice outcomes. Thus, the respondent sees all ten attributes in both tasks, but in the CB tasks chooses one of the brand profiles or chooses none of the alternatives. Unfortunately, all remaining attributes have directional effects which can be assumed *a priori*. The problem with this state of affairs is that, as noted by Dawes and Corrigan (1974), Wainer (1976), Anderson and Shanteau (1977) and others, linear models are particularly robust predictive devices even when grossly misspecified, as long as directional monotonicity is satisfied and the respondents' decision rule is such that more good things are better and more bad things are worse. Due to the nature of the client's problem, we could not avoid this design limitation in our comparison. Thus, our comparison is weaker than we would like, and it must be recognized that we are trading off task and sample richness for this property.

The RB design consisted of two 4-level attributes (price and brand) and eight 2-level attributes, which are service features that are either offered or not. We constructed the RB design by first selecting a 16 profile main effects only plan and combining it with its statistically equivalent foldover, for a total of 32 profiles. Each respondent evaluated all 32 profiles using a 0 to 10 likelihood of purchase rating scale. The purpose of using a main effects and foldover design was to test the effect of forecasting choice based only on the first 16 profiles (the main effects design), compared with forecasts based on all 32 profiles. In the latter case, the combined designs have the property that all main effects can be estimated independently of unobserved but significant linear-by-linear two-way interaction effects. This latter property should reduce bias in the main effects significantly as the second order effects will cause the most bias in main effects if significant. An additional feature of the RB task was the inclusion of an *ad hoc* rating manipulation to allow us to estimate the likelihood of not choosing any of the alternatives. This latter manipulation was necessary to compare the two methods because the CB tasks included a "no choice" option.

The CB design consisted of 80 choice sets blocked into four sets of 20 choice sets. The design was constructed to permit us to estimate not only the comparable main effects to the RB design, but also alternative-specific attribute effects within brand, plus differential cannibalization effects (violations of

the IIA property of MNL models; see Batsell and Louviere). As previously mentioned, choice alternatives in the CB task were the four brands plus the option of not choosing. Blocking the design is a common strategy in CB applications, and 80 choice sets were considered to be too many to reasonably expect respondents to complete in a mail survey.

Respondents were 300 Recreational Vehicle owners, the majority of whom were recruited from a previous study by Allstate. Respondents were randomly assigned into either the RB or CB conditions, which included respectively 151 and 149 participants. Respondents in both conditions responded to a phone-mail-phone survey using a locked box that contain the conjoint materials. Within both RB and CB conditions, respondents were randomly assigned to four sub conditions that represented different holdout choice sets. All respondents in both RB and CB conditions completed a common choice set; the four conditions represented an additional choice set that systematically varied the brand composition of the choice set. Additionally, all respondents in either RB or CB conditions received an additional profile or choice set that was an exact duplicate of one contained in their respective design. The purpose of the latter manipulation was to measure the magnitude of test-retest variability among profiles and choice sets, and is not considered in this paper.

The conditions and sample sizes are shown below:

RB Conjoint	RB N Size	CB Conjoint	CB N Size
Subcond 1	35	Subcond 5	35
Subcond 2	40	Subcond 6	39
Subcond 3	38	Subcond 7	37
Subcond 4	38	Subcond 8	38
Total	151	Total	149

In addition to either the CB or RB conjoint tasks, respondents answered a number of other questions pertaining to current emergency road service, seasonal preferences, name preferences and other measures. Responses to these questions are not considered in this paper, but are the object of ongoing research investigating the effects of individual differences.

A problem immediately confronted in our research is how to base the comparison on a level playing field. Despite many years of research and applications experience with conjoint techniques, little attention has been devoted to ways of comparing different tasks and response scales in order to not favor one approach over another. Despite giving considerable attention to this issue in planning our research project, it is not clear whether we succeeded in leveling the field. The issues involved in such a leveling are as follows:

1. Using choices from sets of competing product descriptions as the holdout samples may have biased the results in favor of CB. To control for this possibility, all respondents completed the holdout sets first. In this way, any learning biases should be equal across conditions, and we avoid the problem of respondents in the CB condition learning rules in their tasks and simply

applying them to "another choice set." Nonetheless, it is clear that the holdout tasks were choice sets similar in nature to the CB task.

2. Commensurability of measurements poses a problem for any conjoint comparison. We favor comparisons based on actual observed marketplace choices, or at least on respondents' reported most recent choices. Unfortunately, it is not clear for psychophysical reasons that attribute levels manipulated in conjoint tasks have a one-to-one relationship with levels observed for real products. That is, the attribute levels that the analyst or client organization believe or know apply to a product may not be the same as the attribute levels perceived by the respondent. Thus, comparisons of this type beg the question of developing commensurable units.

3. There are obvious differences in utility scales between tasks, which are due to differences in variability induced by the tasks themselves. That is, the unit of the scale in most conjoint models is inversely related to the error variability in the response data. Thus, even if the utility formation process is the same in two tasks, model parameters may differ in magnitude because of differing amounts of error. To take these differences into account requires one to homogenize the variance between conditions and tasks.

RESULTS

We first compare the sample-level utility functions for each of the conditions and tasks. To do this we first factor analyze the vectors of parameters to obtain the eigenvalues, and then plot the parameters against the derived factor scores from the first eigenvector. These results are contained in Figures 1 (Factor Analysis Results), 2 (Graphical Result). As can be noted in Figure 1, a single factor accounts for over 93% of the variance in the six vectors of utility parameters, and the graphical results indicate that the various utility vectors are approximately proportional to one another. This latter result can be seen in Figure 2 by the fact that the lines representing each utility vector plotted against the overall factor scores for all utility vectors pass approximately through the 0, 0 coordinate (origin), the condition for proportionality to hold. We plan to conduct formal tests of this condition in the near future.

The proportionality condition derives from the fact that the scale of the utilities in an MNL model is inversely proportional to the variance. The scale cannot be identified for any particular data set (in this case, each condition represents one data set), but ratios of scales can be identified for pairs of data sets (Swait and Louviere 1991, 1992). Thus, the magnitude of the utilities is inversely proportional to the error variability in the data; and even if two vectors of utilities were generated by the same decision process, the magnitudes of the coefficients would differ if their error components differed. The present results suggest that all conditions measured the same utilities in the aggregate, albeit with differing magnitudes of error. If true, this suggests that all were valid measurement instruments with different degrees of reliability.

The second comparison involves tests against the holdout choice sets. Prior to discussing these results, we wish to make the following limitations clear: a) All holdout sets were completed before the conjoint or choice tasks; hence learning effects can be expected in these data, but they should be the same in all conditions. b) The holdout sets were constructed such that the brands competing were systematically varied. Brand emerged as a very powerful explanatory attribute; hence, the predictive validity results can be expected to be better than might have otherwise been the case. It is not clear

what is learned from predictive validity comparisons in which brand name plays a very powerful role in explaining preference because the effects captured by brand name are due to omitted attributes which are correlated with brand name. c) Except for brand name, all attribute effects should have signs that are obvious *a priori* ("yes" is positive, "no" is negative) and price is negatively signed. Unfortunately, the vast majority of respondents preferred one of the brand names, creating a similar situation for brand name. As demonstrated by Dawes and Corrigan, Wainer and many others, in this situation one expects simple linear conjoint models to predict preferences very well. The predictive ability of the models is enhanced by varying the brand composition of the holdouts.

Having said that, we compared the predictive validity of the RB and CB conditions with respect to the following criteria: a) The total chi-square (sum) calculated with respect to observed and predicted choice frequencies; b) the mean square calculated with respect to observed and predicted choice frequencies (The latter is simply the chi-square numerator); and c) the slope and intercept obtained by regressing the natural log of the observed choice frequencies against the natural log of the predicted choice frequencies. The latter comparison is useful because it tells us whether the differences observed in the other two comparisons are due to scale differences or real differences. That is, in the case of the choice models, the variability in the holdout choices might differ significantly from that in the choice experiment. For example, because of learning effects, the early choice set results might be more variable than the choice results observed later. If true, this implies a difference in the scale of the utilities between the model and the holdouts, which should be taken into account in the test.

There are two sets of holdout results: 1) Choices including non-choice, and 2) choices excluding non-choice. The former was obtained after asking respondents to make a forced choice among the alternatives in each set, then asking respondents whether they actually would purchase their preferred profile.

It should be noted that this is not the usual way of obtaining non-choice responses in CB conjoint experiments. The usual way of obtaining these responses is to allow the respondent to make a choice among the competing profiles and a non-choice alternative. Whether this manipulation would have made a difference in the present case cannot be determined from the results of this study.

The non-choice condition required us to develop a special procedure to make holdout choice predictions for RB conjoint. To predict non-choice, we developed an approach to determining non-choice threshold values for each individual as follows: a) after rating each of the profiles, respondents were asked to assume that the profiles that they had just rated were representative of the types of services currently available, and were asked to rate the likelihood that they would actually purchase a service in the next 12 months; b) this rating was converted to a non-choice rating by subtracting it from 10 (the top of the scale); c) we used the highest predicted utility equals first choice rule for the RB choice simulations, such that a profile's predicted rating value had to be the highest among the profiles in a particular set and beat the non-choice value, to be chosen. If no profiles beat the non-choice value, then a non-choice response was predicted.

It should be noted that we are unaware of any other applications of this *ad hoc* non-choice prediction device in RB conjoint. To fully compare RB and CB conjoint predictive validity for holdout sets that permitted non-choice, we were forced to develop a method for estimating the utility of non-choice. After considerable discussion among the members of the research team, we arrived at the procedure previously described. Thus, the fact that the predictive validity results for traditional conjoint turned

out well for this *ad hoc* adjustment does not reflect well on traditional applications of RB conjoint, but rather reflects well on the success of this *ad hoc* adjustment in this particular research problem.

Actual applications experience by one co-author and anecdotal evidence suggest that RB conjoint sometimes is used to predict non-choice by using *ad hoc* rules that apply to all respondents, rather than using individual-specific non-choice adjustments. For example, an assumption that profiles need a predicted rating of at least eight on a ten-point scale might constitute such a sample-wide adjustment. To compare the predictive validity of this type of adjustment, we computed the mean non-choice rating across all respondents and used it as a single, sample-wide non-choice threshold value. In the latter case, individuals' predicted ratings were compared to the mean non-choice rating, and a profile choice was recorded only if the predicted rating for the best profile exceeded the non-choice mean.

It should be noted that given the surprising success of the individual-level non-choice adjustments, using the mean of the individual adjustments can be expected to produce superior predictions to most arbitrary rating category cutoff values (for example, eight on a ten-point scale). Indeed, we compared the predictive validity results for the mean adjustment with an arbitrary "eight or better" cutoff and found the latter to be much worse than the former. In the interest of space, we do not report the "eight or better" results, but note that they were worse than the mean threshold adjusted results, which were considerably worse than chance for most comparisons.

RB models were estimated from the first 16 responses (a main effects plan); and the predictive validity of these models was compared with models based on all 32 responses. Few differences were found; hence, we primarily discuss results for the first 16 profiles.

In the case of CB conjoint, three aggregate-level choice models were estimated from the 80 choice sets represented by the experiment. These models were, respectively, a) a generic main effects only model identical to the RB conjoint models, except that brands were choice outcomes, not attributes; b) a brand-specific main effects model that allowed attribute effects to differ by brand; and c) a generalized extension of the brand-specific attribute effects model that allowed for aggregate violations of the IIA property of MNL models by incorporating cross-effects terms to allow for the attributes of one brand to influence the utility of a second brand.

It should be noted that these aggregate-level models do not necessarily reflect CB conjoint practice because it is possible to disaggregate choice data and estimate more complicated models that take individual differences into account. Due to the preliminary nature of this paper, we do not present such a comparison. Moreover, because of the limitations of the present study described earlier, we would not expect much gain over the generalized brand-specific attribute model.

The observed and predicted choice proportions are contained in Table 1 (a) and (b), which are respectively, the holdouts including and excluding non-choice. Chi-square and mean-square results for observed and predicted choices are contained in Table 2(a) to (d). Both measures strongly agree on their ordering of the fits of the models to the holdout choice sets for sets including and excluding non-choice. The results reveal that the individual-level non-choice threshold adjustment predicted the non-choice responses exceptionally well. Indeed, our results suggest that the individual-level non-choice adjusted RB models predicted the holdout choices near the limits of error in the data. On the other hand, the mean adjusted RB conjoint models predicted very poorly, in some cases doing considerably worse than the chance criterion of equal probability.

The most general MNL choice model also predicted very well and near the limits of error in the holdout data. Indeed, the difference in predictive validity is probably not statistically significant, and certainly is not managerially significant. (It is unclear how to test for statistical differences in the summary statistics because of differences in model specification, numbers of parameters and error structures.) The more general CB model should predict better than the less general models because the cross-effects terms should account for some of the individual differences. As can be seen, this expectation was fulfilled, as the more general MNL model predicted rather better than the other two choice models. All MNL models beat the equal probability model by a considerable margin. These results were consistent across both non-choice excluded and included holdout comparisons.

Thus, on the basis of chi-square and mean-square fit comparisons, the best CB model predicts the holdout data indistinguishably well compared to the RB models. It is worth emphasizing, however, that in the case of non-choice, RB predictions that were not individually adjusted to take non-choice into account predicted considerably worse than CB models. As the individual non-choice adjustment used in this project was *ad hoc* and unique to this project, it is fair to say that RB conjoint with a sample-wide non-choice adjustment was consistently quite inferior to CB conjoint.

To assess whether the chi-square and mean-square results were due to exceptional fits in some sets and not others, as well as whether there were scale differences between holdout and experimental choices, we regressed log-predicted on log-observed choice frequencies to test for possible scale differences. Log-log regressions are appropriate because differences in scale values give rise to differences in slopes, although intercepts will differ significantly from zero unless the two scales are the same in both data sets. In particular, if the scale is the same in both observed and predicted data sets, the slope should not differ significantly from one and the intercept from zero. Slopes that differ from one indicate differences in scale. It should be noted, however, that these tests do not take into account sampling error in both observed and predicted data sets. We plan to conduct tests that take sampling errors into account in future research with these data; hence, these results should be considered preliminary.

The scale results reveal that RB and CB conjoint produced approximately the same levels of fit. Both 16 and 32 profile RB predictions made with individual-level thresholds fit the holdout data very well, and cannot be distinguished from intercept zero and slope one in the non-choice holdout condition. In the condition that excludes non-choice, there appears to be a small but reliable tendency for the CB models to differ in scale, especially in the case of the CB holdout choices (Observed Choice). Thus, it is probably the case that the CB model requires a scale adjustment to fit the holdouts that exclude non-choice. This merely suggests that the CB utility parameters differ from the holdout parameters by a constant factor. That is, the CB parameters are proportional to the holdout parameters, but the constant of proportionality is not equal to one.

An additional test was conducted to determine whether there were differences in the utility functions between the RB and CB models. The utilities derived in multinomial logit models (CB) are affected by a scaling constant. This constant is inversely proportional to the error variance in the data set. While the scaling constant for any one data set cannot be uniquely identified, the ratio of scales between two data sets can be determined and tests conducted. Differences in the coefficients of the same utility function derived from two separate data sets may be due to: (1) different measurement techniques, (2) different variances in the known attributes, or (3) different random error components. The idea is to determine whether the utility functions differ only by a scaling constant or because their parameters are actually different.

The test, developed by Swait and Louviere (1991, 1992), is a modified Chow test of scale parameters. The basic hypothesis is that $\beta_1 = \beta_2$ and $\mu_1 = \mu_2$. However, this hypothesis cannot be directly tested. A sequential test of two hypotheses is conducted. First we test $\beta_1 = \beta_2$ allowing the scaling constants to vary. If this is not rejected, then we test $\mu_1 = \mu_2$ allowing the β's to vary. If both are not rejected, we can say the hypothesis is supported. If the first hypothesis is rejected, then the utility functions must differ by more than a scaling constant. Details of the test may be found in Swait and Louviere (1991, 1992).

The original 80 set CB data set comprised data set one. The RB simulators were used to estimate frequencies for the 80 CB sets. These made up data set two. MNL models using exactly the same utility function were estimated on all data sets. First we tested the sets without the non-choice option, and then we tested the sets using the non-choice option.

The test of sets without the non-choice alternative supports the main hypothesis. Namely, the difference between the estimated utility functions is due solely to a scaling constant. This result is not surprising given our previous results and discussion. The test of sets including the non-choice alternative yield different results, however, There is a statistically significant difference between the utility functions of these two data sets. The hypothesis of $\beta_1 = \beta_2$ is rejected at the .05 level. This suggests a major difference in the two techniques' measurement of the non-choice alternative in the decision making process — even though their predictions of choice are similar in the holdout sets.

DISCUSSION AND CONCLUSIONS

We compared RB and CB conjoint in a real client research situation. Because of the dictates of the research we were unable to control for the fact that all the attributes except brand would have known directional effects on preference *a priori*. This resulted in a weaker comparison than we would have wished, and in fact reflects the ideal case for conjoint choice simulation, as discussed by Elrod and Kumar. That is, there probably was relatively little error in the RB conjoint equations because it was easy for respondents to make judgments. Similarly, because brand composition was systematically varied in the holdout sets, choice alternatives were fairly differentiated, making it easier to predict the best alternatives. The latter result was not expected before the project, and we were surprised to find a rather strong brand preference unrelated to the composition of the sample.

Despite these limitations, the results suggest that one may be able to adjust full-profile conjoint experiments to take non-choice into account. The approach used in this research was *ad hoc*, and a contribution could be made if a more theoretically appealing method could be developed that could be applied in a wide variety of situations. Similarly, the results suggest that, at least in this research context, RB and CB conjoint are both valid measurement models of aggregate utilities. The primary difference between the two models lies in reliability and not validity. If this can be shown to hold in a variety of circumstances, it would suggest that research attention should be focused on developing a theoretical explanation for differences in reliability (for example, differences in task demands and attentional effects), as well as identifying which approach can be expected to be more reliable under what circumstances.

The predictive validity results for CB conjoint reveal once again that an approach based on aggregate data can predict choice in holdout sets as well as or better than (in the case of RB not adjusted to take account of individual difference in non-choice utility) individual-level RB conjoint. Thus, our

results provide additional empirical evidence that individual-level RB conjoint is not superior to aggregate-level CB conjoint in terms of predictive validity. Furthermore, as discussed in the paper, the CB results can be disaggregated to take individual differences into account, which would further reduce the perceived advantages of individual-level RB conjoint. Because product markets differ in terms of aggregate demand, non-choice is often a necessary and important component of market choices. CB conjoint easily accommodates non-choice, but a theoretically acceptable method for incorporating non-choice in RB conjoint has yet to be proposed. The scaling tests on the 80 set estimated frequencies indicate that the two approaches to modeling non-choice differ. Clearly, CB models specifically model this component of the decision making process. However, the predictive success of the *ad hoc* adjustment used in this study suggests that it may be profitable to develop such a theoretical rationale for non-choice in RB studies.

Finally, we were struck by the lack of a standard for comparing these two conjoint methods. The field could benefit from the establishment of a well-thought out, theoretically acceptable set of guidelines for comparing conjoint methods. Our review of the literature suggests that there are almost as many ways of implementing conjoint as there are conjoint researchers. Thus, one person's RB conjoint is not the same as another's. For example, we see little to be gained by continuing to compare conjoint techniques in situations in which the directionality of attribute effects is known *a priori*. Results from the linear models in decision making literature, as well as empirical conjoint studies have repeatedly shown that seriously incorrect linear models will predict well in these circumstances; hence, one learns little from such comparisons. Thus, future comparisons should be directed at situations in which the attributes are largely qualitative or quantitative and non-monotonic. The larger the degree of individual differences in preferences for attribute levels the better. It would also be useful to consider Elrod and Kumar's (1989) discussion of conditions under which conjoint models can be expected to predict choice well, and research should be directed toward conditions in which one expects poorer predictions. We need to understand when conjoint models will and will not predict well.

The assistance of Anne Roth of Allstate Research and Planning Center, and John White and Joffre Swait of Decision Research are gratefully acknowledged.

REFERENCES

Anderson, N.H. and J. Shanteau (1977). "Weak Inference With Linear Models." *Psychological Bulletin*, 84, 6, 1155-1170.

Batsell, R.R. and J.J. Louviere (1991). "Experimental Choice Analysis." *Marketing Letters*, 2.

Bunch, D.S., Louviere, J.J. and D.A. Anderson (1991). "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes." Unpublished working paper, Graduate School of Management, University of California, Davis, October.

Dawes, R.M. and B. Corrigan (1974). "Linear Models in Decision Making." *Psychological Bulletin*, 81, 2, 95-106.

Elrod, T.E. and S. Kumar (1989). "Bias in the First Choice Rule for Predicting Share." *Sawtooth Software Conference Proceedings*, Ketchum, ID, 259-271.

Elrod, T.E. and S.K. Kumar (1989). "Bias in the First Choice Rule for Predicting Share." *Sawtooth Software Conference Proceedings*, 259-271.

Elrod, T.E., J.J. Louviere and K.K. Davey (1992). "A Comparison of Ratings-Based and Choice-Based Conjoint Models." *Journal of Marketing Research*, Forthcoming.

Green, P.E. and V. Srinavasan (1990). "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice." *Journal of Marketing*, 54, 3-19.

Kaciak, E. and J.J. Louviere (1990). "Multiple Correspondence Analysis of Multiple Choice Experiment Data." *Journal of Marketing Research*, 23, 455-465.

Louviere, J.J. (1988a). *Modeling Individual Decisions: Metric Conjoint Analysis*. Sage University Series on Quantitative Applications in the Social Sciences No. 67. Newbury Park, Ca: Sage Publications, Inc.

Louviere, J.J. (1988b). "Conjoint Analysis Modeling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity." *Journal of Transport Economics and Policy*, 10, 93-119.

Louviere, J.J. and G.G. Woodworth (1983). "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregated Data." *Journal of Marketing Research*, 20, 350-367.

Swait, J. and J.J. Louviere (1991). "The Role of the Scale Parameter in Parameter Comparisons Involving Multinomial Logit Choice Models." Paper presented to the TIMS Marketing Science Conference, March.

Swait, J. and J.J. Louviere (1992). "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models." Unpublished Working Paper, Department of Marketing, Eccles School of Business, University of Utah, Salt Lake City, May.

Wainer, H. (1976). "Estimating Coefficients in Linear Models: It Don't Make No Nevermind." *Psychological Bulletin*, 83, 2, 213-217.

Figure 1: Factor Analysis of RB and CB Aggregate Utility Parameters

LATENT ROOTS (EIGENVALUES)

1	2	3	4	5	6
5.597	0.304	0.054	0.031	0.010	0.004

COMPONENT LOADINGS: 1st Component

CJ Main Effects Only Cond1	0.982
CJ Main Effects Only Cond2	0.991
CJ Main Effects Only Cond3	0.970
CJ Main Effects Only Cond4	0.964
CHGeneric, Main Effects Only	0.935*
CHGeneric, Plus Cross Effects	0.952*

PERCENT OF TOTAL VARIANCE EXPLAINED BY 1ST COMPONENT

93.289%

* Only generic main effects and generic main effects with cross-effects MNL model specifications produce partworth estimates that are comparable to the conjoint specifications.

Table 1(a): Observed and Predicted Data, Non-Choice Included Condition

Set	Alt	Nch	Ncj	Och	Ocj	Pl 16	Pl 32	Pn 16	Pn 32	P asm	P asc	P gen
90	1	149	150	.148	.207	.149	.205	.255	.273	.203	.213	.194
90	2	149	150	.396	.393	.411	.434	.511	.510	.350	.366	.326
90	3	149	150	.128	.087	.128	.112	.170	.126	.060	.088	.097
90	0	149	150	.329	.313	.255	.252	.064	.091	.377	.352	.372
91	2	35	34	.429	.588	.596	.580	.702	.720	.552	.537	.543
91	4	35	34	.200	.088	.170	.168	.213	.189	.117	.147	.136
91	0	35	34	.371	.324	.234	.252	.085	.091	.331	.317	.321
92	1	39	40	.154	.300	.262	.245	.312	.329	.227	.207	.227
92	2	39	40	.385	.500	.482	.503	.610	.573	.373	.368	.361
92	0	39	40	.462	.200	.255	.252	.078	.098	.401	.426	.412
93	2	37	35	.703	.657	.525	.531	.652	.650	.445	.455	.410
93	3	37	35	.108	.143	.163	.161	.220	.189	.076	.109	.122
93	0	37	35	.189	.200	.312	.308	.128	.161	.479	.437	.468
94	1	37	38	.189	.158	.149	.140	.206	.182	.178	.16	1.176
94	2	37	38	.460	.316	.326	.357	.390	.413	.292	.303	.281
94	3	37	38	.000	.105	.078	.056	.092	.056	.050	.073	.083
94	4	37	38	.162	.132	.220	.210	.282	.287	.166	.172	.139
94	0	37	38	.189	.290	.227	.238	.050	.063	.314	.291	.320

Table 1(b): Observed and Predicted Data, Non-Choice Excluded Condition

Set	Alt	Nch	Ncj	Och	Ocj	Pi 16	Pi 32	P asm	P asc	P gen
90	1	149	150	.255	.307	.277	.308	.374	.354	.349
90	2	149	150	.571	.580	.553	.552	.471	.466	.498
90	3	149	150	.175	.113	.170	.140	.155	.181	.153
91	2	35	32	.714	.781	.773	.783	.748	.705	.726
91	4	35	32	.286	.219	.227	.217	.252	.295	.274
92	1	39	40	.436	.350	.340	.364	.442	.432	.412
92	2	39	40	.564	.650	.660	.636	.558	.568	.588
93	2	37	35	.838	.829	.745	.755	.753	.720	.765
93	3	37	35	.162	.171	.255	.245	.247	.280	.235
94	1	37	38	.297	.316	.213	.189	.274	.250	.263
94	2	37	38	.514	.421	.418	.441	.346	.329	.376
94	3	37	38	.027	.105	.092	.063	.113	.128	.116
94	4	37	38	.162	.158	.277	.308	.267	.294	.245

Legend:
Set identifies holdout sets. **Alt** identifies a brand. **Nch** is the No. of respondents in the choice condition. **Ncj** is the No. of respondents in the conjoint condition. **Och** is the observed choice condition holdout shares. **Ocj** is the observed conjoint condition holdout shares. **Pi16** is the predicted share using the 1st 16 profiles and an individual-level no-choice threshold. **Pi32** is the prediction for 32 profiles. **Pn16** is the predicted share using the 1st 16 profiles and a mean no-choice threshold. **Pn32** is the prediction for 32 profiles. **Pasm** is the predicted share based, respectively on (asm) the alternative-specific main effects only MNL model; (**Pasc**) the alt.-sp. main and cross-effects MNL model; and (**Pgen**) the generic main effects MNL model.

LEGEND FOR TABLES 2(A) - 2(D)

The Conjoint Sample columns are results comparing model predicted frequencies to actual frequencies derived from the conjoint task respondents. The Choice Sample columns are results comparing model predicted frequencies to actual frequencies derived from the choice task respondents. The last column, Total Sample, compares model predicted frequencies to actual frequencies across all respondents.

There were 5 holdout choice sets. One was shown to all respondents. The other four were randomly assigned to respondents in such a way that each respondent saw 2 additional holdout choice sets (see text).

The conjoint choice simulators and choice models were used to predict choice proportions for each of the 5 holdout choice sets. The table values depict the measure comparing these predictions to the actual frequencies derived from each sample of respondents, or, in the case of the last column, the sample as a whole.

For example, for the rows labeled "Conjoint Profiles" and columns labeled "Conjoint Sample," the table values are the result of comparing the conjoint simulator predicted frequencies to the actual frequencies of choice for the respondents who performed the conjoint task. The "Conjoint Profiles" - "Choice Sample" rows and columns present the values for comparing the conjoint simulator predicted frequencies to the actual holdout set frequencies for the choice task respondents. In the rows labeled "80 Choice Sets" we have predicted frequencies for the holdout sets using the named choice model and compared them to the actual frequencies of choice for the respective sub-samples.

Table 2(a): Chi-Square Results For Choice Simulations

Holdout Sets Include None As An Alternative
(5 Holdout Sets, 18 Total Profiles)

	Conjoint Sample	Choice Sample	Total Sample
1st 16 Conjoint Profiles (OLS) Indiv. Thresh. Chi-Square Sum	15.745	27.800	26.012
1st 16 Conjoint Profiles (OLS) Mean Thresh. Chi-Square Sum	241.451	316.535	527.295
32 Conjoint Profiles (OLS) Indiv. Thresh. Chi-Square Sum	12.429	28.476	22.622
32 Conjoint Profiles (OLS) Indiv. Thresh. Chi-Square Sum	154.733	213.376	340.197
80 Choice Sets Generic Main Effects MNL Chi-Square Sum	23.185	32.872	40.014
80 Choice Sets Alter.-Specific Main Effects MNL Chi-Square Sum	25.368	39.822	46.416
80 Choice Sets Alter.-Specific Cross-Effects MNL Chi-Square Sum	19.229	26.129	29.051
Equal Probability Model (Null Ho) Chi-Square Sum	73.795	83.851	143.536

Table 2(b): Chi-Square Results For Choice Simulations

Holdout Sets Exclude None As An Alternative
(5 Holdout Sets, 18 Total Profiles)

	Conjoint Sample	Choice Sample	Total Sample
1st 16 Conjoint Profiles (OLS) Chi-Square Sum	8.787	9.857	13.391
32 Conjoint Profiles (OLS) Chi-Square Sum	9.183	12.044	15.197
80 Choice Sets Generic Main Effects MNL Chi-Square Sum	8.095	12.657	15.765
80 Choice Sets Alter.-Specific Main Effects MNL Chi-Square Sum	12.511	17.886	25.386
80 Choice Sets Alter.-Specific Cross-Effects MNL Chi-Square Sum	17.186	19.569	32.080
Equal Probability Model (Null Ho) Chi-Square Sum	88.038	82.344	166.504

Table 2(c): Mean Square Results For Allstate Choice Simulations

Holdout Sets Include None As An Alternative
(5 Holdout Sets, 18 Total Profiles)

	Conjoint Sample	Choice Sample	Total Sample
1st 16 Conjoint Profiles (OLS) Indiv. Thresh. Mean Square	15.378	21.777	52.297
1st 16 Conjoint Profiles (OLS) Mean Thresh. Mean Square	122.137	153.759	530.040
32 Conjoint Profiles (OLS) Indiv. Thresh. Mean Square	11.752	26.731	55.016
32 Conjoint Profiles (OLS) Indiv. Thresh. Mean Square	100.102	136.625	451.680
80 Choice Sets Generic Main Effects MNL Mean Square	26.073	30.946	91.840
80 Choice Sets Alter.-Specific Main Effects MNL Mean Square	23.123	32.342	88.770
80 Choice Sets Alter.-Specific Cross-Effects MNL Mean Square	17.013	22.996	57.830
Equal Probability Model (Null Ho) Mean Square	92.000	95.601	353.296

Table 2(d): Mean Square Results For Allstate Choice Simulations

Holdout Sets Exclude None As An Alternative
(5 Holdout Sets, 18 Total Profiles)

	Conjoint Sample	Choice Sample	Total Sample
1st 16 Conjoint Profiles (OLS) Mean Square	12.578	9.608	28.419
32 Conjoint Profiles (OLS) Mean Square	8.232	15.238	31.114
80 Choice Sets Generic Main Effects MNL Mean Square	21.046	29.988	86.064
80 Choice Sets Alter.-Specific Main Effects MNL Mean Square	36.877	48.660	154.876
80 Choice Sets Alter.-Specific Cross-Effects MNL Mean Square	42.661	45.325	159.674
Equal Probability Model (Null Ho) Mean Square	235.916	197.033	848.200

Table 3(a): Test For Rescaling Factor Equal To One, Holdouts That Include Non-choice

		Observed Total	Observed Conjoint	Observed Choice
Conjoint Indiv. Thresh. 16				
	Inter.	-0.353	-0.154	-0.044
	Slope	1.100	1.048	1.013
Conjoint Indiv. Thresh. 32				
	Inter.	-0.089	0.068	0.045
	Slope	1.019	0.961	0.973
Conjoint Aggreg. Thresh. 16				
	Inter.	1.512*	1.199*	1.556*
	Slope	0.542*	0.539*	0.421*
Conjoint Aggreg. Thresh. 32				
	Inter.	1.007*	1.331*	1.210*
	Slope	0.620*	0.507*	0.640*
Choice Generic Main Effects				
	Inter.	-0.084	0.043	0.213
	Slope	1.016	0.969	0.914
Choice Alt-Sp. Main Effects				
	Inter.	0.397	0.446	0.500
	Slope	0.882	0.828	0.814
Choice Alt-Sp. Cross Effects				
	Inter.	-0.029	0.124	0.177
	Slope	0.999	0.937	0.926*

Significantly Different from zero or one at alpha=.05

Table 3(b): Test For Rescaling Factor Equal To One, Holdouts That Exclude Non-choice

		Observed Total	Observed Conjoint	Observed Choice
Conjoint 16 No Thresh.				
	Inter.	-0.314	-0.069	-0.734
	Slope	1.075	1.009	1.225
Conjoint 32 No Thresh.				
	Inter.	0.051	0.233	-0.442
	Slope	0.980	0.912	1.135
Choice Generic Main Effects				
	Inter.	-0.651*	-0.358	-1.001*
	Slope	1.161	1.100	1.306*
Choice Alt-Sp. Main Effects				
	Inter.	-0.613*	-0.327	-0.978*
	Slope	1.150	1.088	1.297*
Choice Alt-Sp. Cross Effects				
	Inter.	-0.795	-0.457	-1.159*
	Slope	1.192	1.124	1.348**

Significantly Different from zero or one at alpha=.05

APPLICATIONS OF LOGIT MODELS IN MARKET RESEARCH

Dan Steinberg
San Diego State University

1. INTRODUCTION

Logit models have long been a staple in biostatistics and the social sciences, and over the past few years they have received increasing attention from market researchers. Nonetheless, logistic regression is a topic that is rarely covered in detail in statistics courses, and it is difficult to find comprehensive expositions accessible to the applied researcher. One reason for this scarcity of tutorial materials is that the term logit actually encompasses a family of diverse models and techniques. The logit of the biostatistician is somewhat different than that of the sociologist, and the economics-inspired discrete choice logit appears to be yet a third variant distinct from the other two. Textbooks and articles on logit usually focus on a single variant, inspired by a particular discipline. This leaves the reader with only a partial coverage of the topic and promotes a sense of mystery and confusion when other logit variants are encountered.

The multiplicity of versions and interpretations of logit has tended to cast logit as a specialty topic and has generated the perception that logit analysis is a black art. This is somewhat unfortunate, as logit is not more difficult to master than other advanced techniques and it is an effective tool for revealing data structure.

This paper is an attempt to provide an overview of logit modeling with emphasis on real world examples. In section two I review the core logit models, exhibit the main equations, and suggest the types of applications they might receive in market research. In section three I focus on the discrete choice model and provide some examples from designed choice experiments. Section four briefly moves on to the nested logit model, and section five highlights some important new developments in the field which are especially relevant to market research.

2. THE FAMILY OF LOGIT MODELS

There are four distinct variants of the logit model with wide application in applied research: (1) binary logit, (2) polychotomous logit, (3) multinomial discrete-choice logit, and (4) matched-sample case-control conditional logistic regression. All are tied together by a common mathematical core (Luce, 1959), but they have been specialized in their style of output, statistics reported, tools of interpretation, vocabulary, and preferred areas of application.

Binary Logit

The binary logit is best known in biostatistics although it is also common in the social sciences. Here the dependent variable is a zero/one indicator variable and the independent variables commonly include demographics and contextual variables. Examples of dependent variables are whether a consumer is planning to buy a new car, whether a household would switch long distance telephone companies in response to an offered incentive, and whether a person uses a particular

brand of soap. The binary logit model has compelling advantages over other modeling techniques such as linear probability models and probit: it provides efficient estimates of the response function coefficients, it is easy to compute and converges quickly, and it allows for a higher percentage of outliers than the normal distribution model (probit). The biostatisticians have introduced a particularly useful way of reporting the results by converting the raw logit coefficients into odds ratios. The odds ratios reveal clearly the effect each independent variable has on the outcome.

The equations for the binary logit probability are:

$$\text{prob}(Y=1) = \frac{\exp(X\,\beta)}{1 + \exp(X\,\beta)} \qquad (1)$$

$$\text{prob}(Y=0) = \frac{1}{1 + \exp(X\,\beta)}$$

These are rather simple equations but difficult to absorb intuitively. Transforming to log odds gives a familiar linear format:

$$\log [\, \text{Prob}(Y=1) / \text{Prob}(Y=0) \,] = \log (\exp(X\,\beta)) = X\,\beta \qquad (2)$$

We used binary logit to analyze data from a choice experiment involving a new information service. The new service was offered to current users of a competing information service and differed in range of information available, base monthly fees, and usage charges. The dependent variable was whether the customer would switch services, and the model included fee reduction, usage rate reduction, and an estimate of total cost savings on the customer's current level of service. The edited results of a simple model were:

```
=====================
BINARY LOGIT ANALYSIS
=====================

DEPENDENT VARIABLE: DPV0

CATEGORY  CHOICES
-----------------
 RESP   |     302
 REF    |     794
--------+--------
        |    1096


 CONVERGENCE ACHIEVED

RESULTS OF ESTIMATION
=====================

LOG LIKELIHOOD: -618.996

   PARAMETER                    ESTIMATE         S.E.      T-RATIO      P-VALUE
-------------------------------------------- ---------- ---------- ----------
 1 CONSTANT                |       -0.842        0.242       -3.475        0.001
 2 DFEE                    |        0.262        0.057        4.605        0.000
 3 DUSE                    |       41.140        9.060        4.541        0.000
 4 DTOTCOST                |        0.011        0.007        1.751        0.080
-------------------------------------------------------------------------------

                                                    95.0% BOUNDS
   PARAMETER                 ODDS RATIO         UPPER         LOWER
---------------------------------------------------------- ----------
 2 DFEE                    |        1.299        1.452        1.162
 3 DUSE                    |  .736115E+18  .378912E+26  .143005E+11
 4 DTOTCOST                |        1.012        1.025        0.999
---------------------------------- ----- ------------------------
LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -645.214
2*[LL(N)-LL(0)] = 52.437 WITH 3 DOF, CHI-SQ P-VALUE = 0.000
MCFADDEN'S RHO-SQUARED = 0.041
-------------------------------------------- ------------------------------
```

The results begin with estimated coefficients much like linear regression, but are based on a transformed scale. The constant tells us there is some inertia in the market with decision makers unwilling to change in the absence of any cost savings, even though the new service offers advantages. Just looking at the size of the coefficients it is plain that savings in the usage rate is the major driver in the decision to switch.

The odds ratios listed below the regression coefficients give the ratio of the probability of switching to the probability of not switching. An odds ratio of one means that the probabilities are 1/2 each; larger odds ratios indicate a higher likelihood of switching. In the panel above, a one dollar increase in monthly fee savings increases the relative probability of switching by 29%, while a similar savings in usage rate effectively captures the entire market.

A useful way of looking at these results is to trace out the response function. Holding all variables but one constant, we vary monthly fee savings from -$10 (a fee increase) to $10 in increments of $.10 and evaluate the probability of switching and its confidence interval. This simulation is

computed with a single command in SYSTAT LOGIT, and the results are plotted with SYGRAPH. From the graph, it is easy to read off the savings required to induce any target market share. (See figures 1 and 2)

Figure 1

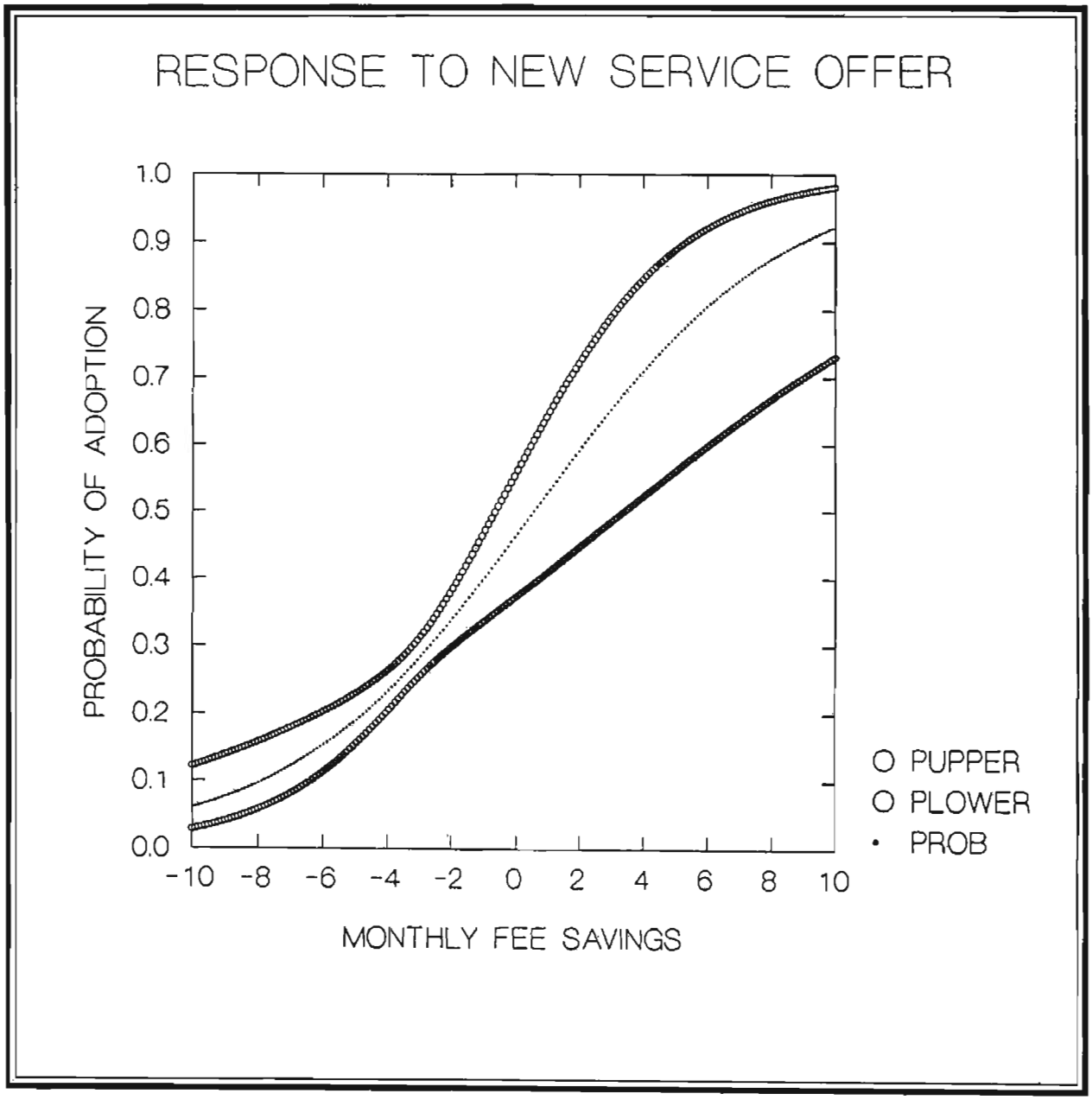
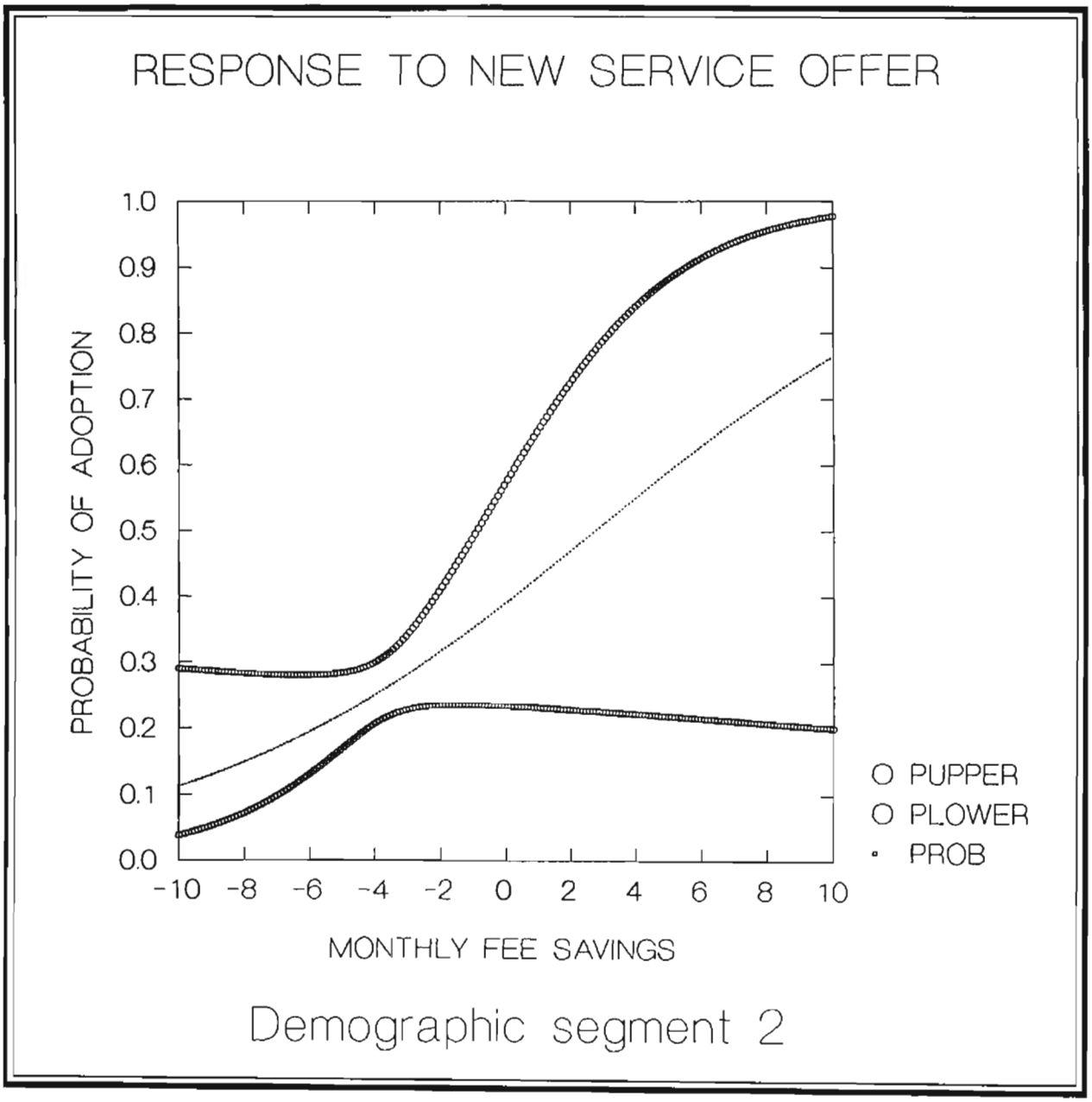


Figure 2


RESPONSE TO NEW SERVICE OFFER
PROBABILITY OF ADOPTION
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-10 -8 -6 -4 -2 0 2 4 6 8 10
MONTHLY FEE SAVINGS
O PUPPER
O PLOWER
PROB
Demographic segment 2

The complications of binary logit modeling relate to model evaluation, comparing alternative models, model diagnostics, useful representation of the output, and the interpretation of interactions. Most of these topics are explained in the excellent textbook of Hosmer and Lemeshow (1989).

Polychotomous Logit

The sociologists' version of the logit model (Nerlove and Press, 1973), sometimes referred to as the polytomous or polychotomous logit, extends the dependent variable from a binary indicator variable to a multi-level categorical variable. For example, the dependent variable might be country of origin of primary vehicle (US, Europe, Japan, or other) and the independent variables could be background characteristics of the study subject such as age, education, race, region of residence, marital status, household income, and a battery of attitude measures. Such models are typically estimated on "found" data, that is, data available in general social surveys of a population of interest.

The most important aspect of polychotomous logit is that the model generates a complete binary sub-model for each possible outcome of the dependent variable versus a reference level (Begg and Gray, 1984). In the following example, involving a series of choice experiments, individuals were faced with four alternatives concerning a new product: (1) take the product with premium A (2) take the product with premium B (3) take the product with premium C or (4) do not take the product. If the response is estimated as a polychotomous logit model estimated, we obtain the equivalent of these three binary sub-models:

Sub-Model	Dependent Variable	Cases Included
1 vs 4	DPV=1 if choice=1 DPV=0 if choice=4	Respondents choosing 1 or 4
2 vs 4	DPV=1 if choice=2 DPV=0 if choice=4	Respondents choosing 2 or 4
3 vs 4	DPV=1 if choice=3 DPV=0 if choice=4	Respondents choosing 3 or 4

In each sub-model the reference group is the same, and only a subset of the data is used. This feature is particularly difficult for novices to grasp and makes comprehensive interpretation of the output close to impossible. For example, the raw output of a typical polychotomous logit model could look like this:

```
==========================
MULTINOMIAL LOGIT ANALYSIS
==========================

DEPENDENT VARIABLE: CHOICE

INPUT RECORDS: 4581

RECORDS FOR ANALYSIS: 3643
RECORDS DELETED FOR MISSING DATA: 938

SAMPLE SPLIT
============

CATEGORY  CHOICES
-----------------
      1 |     499
      2 |     947
      3 |     583
      4 |    1614
--------+--------
        |    3643


CONVERGENCE ACHIEVED

RESULTS OF ESTIMATION
=====================

LOG LIKELIHOOD: -4640.4137

   PARAMETER                  ESTIMATE         S.E.      T-RATIO      P-VALUE
-------------------------------------------------------------------------------
CHOICE GROUP : 1          |
 1 CONSTANT               |      -1.2197       0.1677      -7.2731       0.0000
 2 INCLT20                |      -0.0415       0.1267      -0.3274       0.7433
 3 INC2040                |      -0.0741       0.1215      -0.6097       0.5420
 4 OFTEN                  |       0.0119       0.0221       0.5391       0.5898
CHOICE GROUP : 2          |
 1 CONSTANT               |      -0.9031       0.1399      -6.4531       0.0000
 2 INCLT20                |       0.0259       0.1045       0.2475       0.8046
 3 INC2040                |       0.2904       0.0951       3.0527       0.0023
 4 OFTEN                  |       0.0397       0.0182       2.1826       0.0291
CHOICE GROUP : 3          |
 1 CONSTANT               |      -1.1987       0.1618      -7.4074       0.0000
 2 INCLT20                |       0.0182       0.1193       0.1525       0.8788
 3 INC2040                |      -0.0226       0.1147      -0.1970       0.8438
 4 OFTEN                  |       0.0274       0.0212       1.2920       0.1964
-------------------------------------------------------------------------------
LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -4650.0685
2*[LL(N)-LL(0)] = 19.3096 WITH 9 DOF, CHI-SQ P-VALUE = 0.0227
MCFADDEN'S RHO-SQUARED = 0.0021
-------------------------------------------------------------------------------
```

The core notion is quite similar to that of dummy variables in linear regression. There, if we have a categorical independent variable with say 4 levels (such as race) coded into indicator variables, we expect to estimate three coefficients, each of which represents a difference in the intercept from a common reference level (the left out dummy). In polychotomous logit models, one level of the dependent variable is selected as the reference group (usually the highest level), and the equivalent of a binary model involving the reference group and one other level of the dependent variable is estimated. Instead of just estimating a single coefficient, though, as in the dummy variable case, a complete model is estimated with coefficients for all independent variables in each.

Why bother with polychotomous logit then, if the same results can be obtained by estimating a series of binary logits instead? The results of the combined polychotomous model are preferred to separate binary models because the coefficients are estimated more efficiently, cross response function hypotheses can be tested, and at least in the case of SYSTAT's LOGIT module, a comprehensive overview of the entire model can be generated.

In SYSTAT LOGIT the derivative summary table combines all the information from each sub-model into a coherent whole:

```
 INDIVIDUAL VARIABLE DERIVATIVES
 AVERAGED OVER ALL OBSERVATIONS
================================

    PARAMETER                    1           2           3           4
-------------------------------------------------------------------------
 2 INCLT20              |    -0.0062      0.0057      0.0023     -0.0017
 3 INC2040              |    -0.0185      0.0592     -0.0135     -0.0272
 4 OFTEN                |    -0.0006      0.0060      0.0018     -0.0072
-------------------------------------------------------------------------
```

Reading the second row of the table we see that if more respondents were in the lowest income group, choices and 1 and 4 would become less likely while choices 2 and 3 would increase in probability. Each element in the table is a probability change, and all rows show how probability is reallocated to sum to zero. As in the binary logit, predicted probabilities can be saved, and simulations can be generated to produce 3-dimensional response surfaces.

Before going on to discrete choice models it is useful to review the probability equations for this logit:

$$\text{prob}(Y=i) = \frac{\exp(X\beta_i)}{1 + \sum_{j=1}^{k-1} \exp(X\beta_j)} \qquad (3)$$

for $i=1,\ldots,k-1$, and

$$\text{prob}(Y=k) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(X\beta_j)} \qquad (4)$$

where the dependent variable Y ranges from 1 to k , β_i is the vector of coefficients corresponding to

the i'th level of the dependent variable, and the coefficients for β_k have been normalized to 0. We sometimes need to refer to a particular element in the β_i vector; for the q'th independent variable the coefficient corresponding to the i'th outcome is β_{iq} .

$$\log [\text{Prob}(Y=i) / \text{Prob}(Y=k)] = \log (\exp(X \beta_i) = X \beta_i \quad (5)$$

Similarly, the log odds of the i'th alternative relative to the j'th alternative are:

$$\log [\text{Prob}(Y=i) / \text{Prob}(Y=j)] = \log \frac{\exp(X \beta_i)}{\exp(X \beta_j)} = X(\beta_i - \beta_j) \quad (6)$$

Note that is the same as formula (5) with β_j replacing β_k; that is, the log odds can always be calculated using differences of coefficients. Although any one set of coefficients can be normalized to zero, the differences between coefficients are invariant across normalizations. Thus, it is easy to calculate the new coefficients that would arise when normalizing on the q'th outcome instead of the j'th outcome:

$$\text{new coefficient for } X_i = \beta_i - \beta_q \quad (7)$$

that is, replace the original coefficient (which is $\beta_i - \beta_k = \beta_i - 0 = \beta_i$) with a new difference.

In these equations (3) and (4) it is important to note that the X vector is constant for each case in the data, and a separate coefficient vector β_i is estimated for each alternative outcome but the reference level.

Discrete Choice

The discrete-choice variant of logit was introduced into economics by Daniel McFadden (1973) of MIT in the context of transportation research. McFadden termed the model conditional logit and within economics it is often referred to simply as multinomial logit (MNL). Like polytomous logit, the model takes a categorical variable as dependent variable but that is where the similarity with the previous logit models ends. What is distinct about the discrete choice model is how independent variables are handled and the interpretation of the corresponding coefficients.

The simplest way to explain this is to start with the theoretical basis of the discrete choice model. The model states that rational decision makers can be modeled as making choices from a well-defined set of alternatives, described as a choice set. Each alternative can be modeled as having a utility which is a function of the attributes of that alternative alone. Thus, for example, a car's utility might be a function of its price, horsepower (HP), internal carrying capacity (VOL), gas mileage (MPG), and perceived reliability (RELI). Mathematically, each alternative in a choice set can be described by the utility function:

$$U_1 = \alpha_1 + HP_1\beta_1 + VOL_1\beta_2 + MPG_1\beta_3 + RELI_1\beta_4 + \varepsilon_1$$

$$U_2 = \alpha_2 + HP_2\beta_1 + VOL_2\beta_2 + MPG_2\beta_3 + RELI_2\beta_4 + \varepsilon_2 \quad (8)$$

....

$$U_K = \alpha_K + HP_K\beta_1 + VOL_K\beta_K + MPG_K\beta_3 + RELI_K\beta_4 + \varepsilon_K$$

In the above set of equations, each utility is a function only of that alternative's own attributes, and in its simplest version, there are no demographics variables present. Note that in the above set of equations, the number of product attribute parameters is not a function of the number of alternatives in the choice set.

The error term is an important part of this model and its presence converts this representation into *a random utility model* (McFadden, 1984). Errors are introduced for all the reasons familiar from linear regression; errors in measurement, decision makers varying from time-to-time in their perceptions of attributes, mistakes, and modeling errors such as left out variables.

The choice model is completed by the assumption that the decision maker chooses the alternative with the highest utility. Since that utility has a random component, prediction will at best be probabilistic, and the model yields results in the form of a probability that a given alternative will be chosen. When the error terms have an extreme value type I distribution the probability statements take the form of logit models:

$$\text{prob}(Y=i) = \frac{\exp(X_i\beta)}{\sum_{j=1}^{k} \exp(X_i\beta)} \qquad (9)$$

In contrast to the equations listed for the polychotomous logit, the equation above contains only a single coefficient vector ß. The components which vary over the alternatives are not the coefficients but the product attributes, such as horsepower, price, and quality. The parallels to conjoint measurement will be evident to market researchers; the model is utility based, it is computed from product attributes, and the coefficients are similar to partworths. The differences between logit modeling and conjoint analysis are extremely important, however and will be discussed below.

Here is an example of discrete choice modeling output. The data come from the same choice experiment described above in the polychotomous logit example in which customers of an information service were offered premiums to switch to a competing service. In this analysis, we focused on the subset of customers who opted for the new service; they had a choice of three distinct premium packages, with product dimensions A, B, and C. The model included an alternative specific constant and the levels of the premium attributes. A very simple analysis yielded:

```
  ==========================
  CONDITIONAL LOGIT ANALYSIS
  ==========================

DEPENDENT VARIABLE: CHOICE

SAMPLE SPLIT
============

CATEGORY  CHOICES
-----------------
      1 |     513
      2 |    1091
      3 |     616
--------+--------
        |    2220


CONVERGENCE ACHIEVED

RESULTS OF ESTIMATION
=====================

LOG LIKELIHOOD: -2270.961

   PARAMETER                    ESTIMATE         S.E.      T-RATIO      P-VALUE
-------------------------------------------------------------------------------
 1 ATTRIBUTE-A          |         0.030        0.004        6.735        0.000
 2 ATTRIBUTE-B          |         0.031        0.009        3.468        0.001
 3 ATTRIBUTE-C          |         0.020        0.003        6.195        0.000
-------------------------------------------------------------------------------
 4 CONSTANT-ALT 1       |        -0.413        0.226       -1.828        0.068
 4 CONSTANT-ALT 2       |         1.056        0.184        5.740        0.000
-------------------------------------------------------------------------------
LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -2316.320
2*[LL(N)-LL(0)] = 90.717 WITH 3 DOF, CHI-SQ P-VALUE = 0.000
MCFADDEN'S RHO-SQUARED = 0.020
-------------------------------------------------------------------------------
```

The model gives us information on both the alternative specific constants and the relative desirability of the dimensions of the premium package. In the upper panel we see that attributes A and B are virtually identical in desirability and efficacy in drawing decision makers, each being about 1.5 times as desirable as attribute C. Clearly, the firm should focus on the attribute which gives the best draw per dollar cost. Also, in the second panel we see that, relative to alternative 3, alternative 1 appears less desirable, while alternative 2 is more desirable. The constant tells us that over and above measured package attributes, decision makers are exhibiting statistically significant preferences.

To evaluate this simple model we can look at the PREDICTION SUCCESS TABLE (McFadden, 1979):

MODEL PREDICTION SUCCESS TABLE

ACTUAL CHOICE	PREDICTED CHOICE 1	2	3	ACTUAL TOTAL
1	127.295	251.258	134.448	513.000
2	251.503	540.430	299.067	1091.000
3	134.203	299.312	182.485	616.000
PRED. TOT.	513.000	1091.000	616.000	2220.000
CORRECT	0.248	0.495	0.296	
SUCCESS IND.	0.017	0.004	0.019	
TOT. CORRECT	0.383			

The table is read first by rows. It shows that out of 513 cases in which alternative 1 was actually chosen, only 127 were predicted to make that choice. The model does a little better among the alternative 2 choosers and again poorly for alternative 3. Overall, this simple model manages to predict correctly 38.3 percent of the learning sample cases. This is not surprising as this model does not include any of the important demographic variables such as ethnicity and income; however, for this example we wanted to keep the model as simple as possible.

Discrete choice models and conjoint models both emerged in the early 1970's but largely evolved independently of each other. One reason for this is that discrete choice models originally had their application in transportation economics, and thus remained in a highly specialized sub-discipline even within economics. (See, for example, Ben-Akiva and Lerman, 1985; Train, 1986). Other important reasons are that the original expositions of the model were highly technical in nature, were quite difficult to understand, and they did not receive widespread attention even within economics. Finally, a major reason for the separate evolution of conjoint and logit models is that during the 1970's economists shied away from stated preference surveys, opting instead to base inference on observed market behavior.

A classic example is the analysis of the mode-of-travel-to-work choice. Urban transportation planners have long considered the viability of mass transit, car pooling, and other alternatives to the drive-alone mode, in an effort to alleviate traffic congestion in major cities. Discrete choice models have been frequently generated to characterize marketplace consumer choice from the available alternatives.

To conduct the study, a random sample of the population is selected, and respondents are asked to provide information on where they work, their travel mode, some related cost information, such as parking fees, and make, model, and year of the vehicle if owned, and common demographics. In most studies, it is easy to identify the entire set of options, and to measure the objective mode attributes of each. For example, in New York City, persons living and working in Manhattan can reasonably consider walking, train, bus, bicycle, taxi, private car, and car pool as alternative travel modes. The time of travel and the cost of each alternative can be calculated from bus and train schedules and traffic network models. Thus, each respondent needs only to provide information

about his or her chosen alternative, and the choice is revealed in market behavior rather in an artificial or hypothetical setting.

Once the model is estimated, implicit tradeoffs between price of travel and time of travel (among other attributes) can be estimated. Also, important economic quantities such as elasticity of demand can be calculated straightforwardly. Key features of the approach are: only choice information is used, ratings are not part of the process; the choice set need not be the same for each person, and it need not contain the same numbers of alternatives.

Another key component of the model is the alternative-specific constant. Each alternative has unmeasured aspects contributing to its utility, aspects that may be difficult or impossible to capture in generic attributes such as price, size, or convenience. For models in which all decision makers do face the same alternatives and the same choice set, a constant devoted to each alternative can be estimated, which captures that alternative's relative desirability, separate from any measured attributes. For example, in a choice between different laundry detergents, alternative specific constants could pick up brand specific preference that is not captured by attributes.

DISCRETE CHOICE AND DESIGNED CONSUMER PREFERENCE SURVEYS

The original application of discrete choice models was confined to observed market behavior. Consumers were surveyed to obtain their revealed choices concerning travel modes, use of gas or electric appliances, living situation (own, rent-head of household, rent-non-head of household), and so on. The resulting data were filled out to provide a complete objective description of the choice set facing each respondent (See, for example, Hensher and Johnson, 1981).

During the last decade a renewed interest in stated preference surveys has emerged in economics, driven by the need to value public (non-marketed) goods. Focused primarily on environmental issues such the valuation of clean air, an adequate water supply, or public sport fisheries, this line of research has used hypothetical choice scenarios to elicit consumer preference. A substantial portion of the public goods valuation literature still focuses on revealed market behavior, eliciting environmental goods values from data on how far consumers are willing to drive, for example, to see the Grand Canyon. For many related questions, however, there is general agreement that no observed behavior is adequate to solve the valuation problem. Consequently, environmental economists have resorted increasingly to the a survey style that is quite common in market research. (For a comprehensive survey, see Carson and Mitchell, 1990).

At the same time, a handful of pioneering market researchers, realizing the power of discrete choice models, attempted to take the best ideas emerging from the conjoint literature and merge them into econometric choice modeling. (Louviere and Hensher, 1982, 1983; Louviere, 1984). The result of both of these lines of research has been to gather data from hypothetical choice experiments, and then to analyze them using discrete choice models (Louviere, 1991). This had been the method used to price telephone services such as phone rental and calling options prior to the divestiture, and is apparently still at the heart of the ongoing consumer research in the communications industry.

Going yet a step further, research has also proceeded on the optimal design of those hypothetical scenarios. Alberini and Carson (1990) tackled the problem of binary choice response and the design points of a single independent variable, while Bunch, Louviere and Anderson (1991) have focused on the multinomial choice problem including the selection of choice sets.

In application, the ideas are very straightforward. The attributes which define actual and hypothetical products are used describe a choice scenario. Respondents are asked to either choose the most preferred item or to rank the alternatives, anchoring the choices to market behavior by always allowing a no-choice alternative ("none of the above"). The results are then analyzed using the McFadden model.

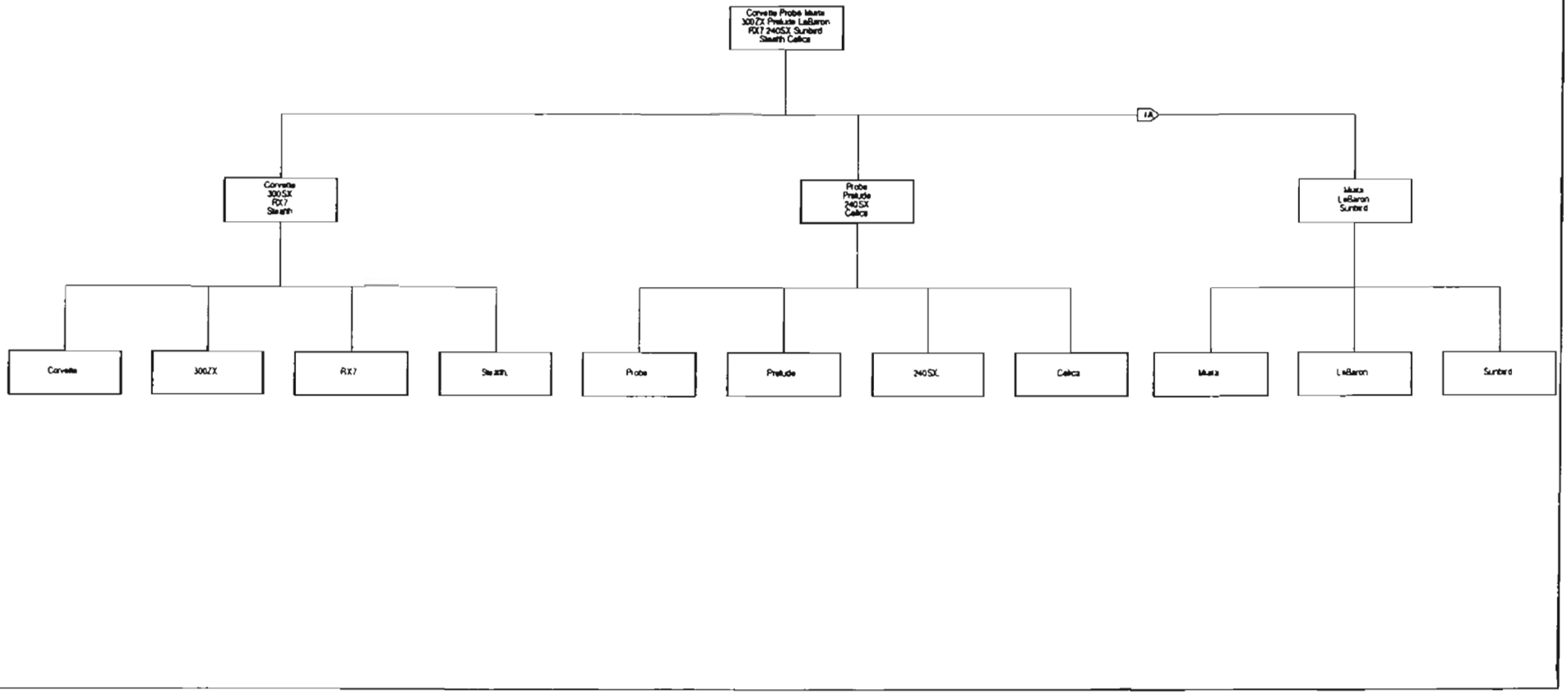
Nested Logit Models

The simple MNL model suffers from one serious design flaw. When a new alternative is added to an existing choice set the mathematics requires that its market share does not alter the relative standings of the original alternatives. This means, for example, that when Lexus added its line of cars to the consumer choice set, the simple MNL model would have predicted that the relative market shares of Mercedes-Benz and Cadillac would remain unchanged. So that if Mercedes and Cadillac had roughly equal market shares before the Lexus, they would still have roughly equal market shares after. However, it is quite possible that Lexus would draw more heavily from Mercedes-Benz than from Cadillac, thus altering the balance in favor of Cadillac. A similar prediction would be made by simple logit in presidential politics; the relative standing of George Bush and Bill Clinton would be unchanged by the entry or departure of Ross Perot.

This property, technically known as IIA (independence of irrelevant alternatives) is clearly a flaw in many choice contexts. The problem arises because some alternatives are more similar to each other than others, and thus compete more among each other for market share. When a new alternative is introduced into the choice set, the results will depend on which alternatives it is most "similar" to.

The nested logit model was introduced to deal with this problem by allowing for a similarity parameter to connect various alternatives. Nested logit models can be represented by sequential tree structured choices, where each dimension of similarity looks like another component of the choice process. In the following example, consumers are faced with a choice of renting from a selection of 8 sporty automobiles. In the non-nested logit, the decision process would be modeled as a straight selection from all the alternatives. It is unlikely, however, that if the LeBaron were withdrawn from the list, that the person who ranked it first would switch to the Dodge Stealth, since he is probably looking for a convertible. We thus structure the choice by grouping "similar" cars together in the following tree structure. There is not necessarily any compelling reason for the structure exhibited here to be superior to an alternative grouping. The grouping selected here was guided by the classifications made by car rental companies.

Nested Logit Model: Sporty Car Rental Choice



The importance of the nested logit lies in its ability to provide more accurate predictions of market share changes in response to changes in choice sets. Although the theory of the nested logit is too complicated to discuss in here, it can be applied fairly easily and it is a critical component in some modeling contexts.

NEW DEVELOPMENTS

Logit research is currently quite active, and a number of new developments have recently emerged, or are in the works. One very important arena is consumer heterogeneity. By and large, discrete choice surveys tend to collect fewer profiles than is common in a conjoint study, and no attempt is made to estimate individual level models. This is partly because the nonlinear logit model requires a large number of observations to yield satisfactory results, and partly because there is very good reason to be skeptical of models estimated with so few degrees of freedom. Nonetheless, it is evident that heterogeneity is the rule rather than the exception, and population-wide coefficient values (partworths) will not suffice to identify market niches that will appeal to the atypical consumer.

To address this problem, Scott Cardell (1988) introduced a mixture model in which the population distribution parameters of the logit coefficients are also estimated along with the logit model. Thus, for example, if a logit model determines that the value of an extra 10 horsepower is worth $600 to the average consumer, the heterogeneity model could tell us that for 17% of market, say, it is worth at least $800. Similar distributions can often be estimated for each coefficient in the model. At the present time Cardell and Steinberg are working on a software implementation of the heterogeneity logit; it is severely computer intensive, however, and typical market research data sets could easily require several days of running time on a fast 486.

Another area of new research involves the attempt to link and merge stated preference and revealed preference data. This work undertaken by Ben-Akiva and Morikawa (1990) and others combines observed market behavior with hypothetical scenarios in the same survey. The aim of the work is to use the behavioral data to adjust and correct the stated preferences, bringing them into a common scale that can yield reliable market forecasts.

A third innovation, by Steinberg and Cardell (1992), is a methodological procedure for making use of proprietary data sets that do not contain sufficient behavioral choice information to estimate a conventional discrete choice model. For example, a firm might have an extensive data set on its own customer base, yet might have no information on consumers dealing with other firms. The method allows the partial proprietary data to be merged with publicly available general population data to estimate a model. The method applies even when the public data do not contain any choice information.

REFERENCES

Alberini, A. and R. Carson (1990). "Choice of Threshholds for Efficient Binary Discrete Choice Estimation." Discussion Paper 90-34, Department of Economics, University of California, San Diego.

Begg, Colin B. and Robert Gray (1984). "Calculation of Polychotomous Logistic Regression Parameters Using Individualized Regressions." *Biometrika*, 71, 11-18.

Ben-Akiva, Moshe and Steven Lerman (1985). *Discrete Choice Analysis*. MIT Press, Cambridge.

Ben-Akiva, M. and T. Morikawa (1990). "Estimation of Travel Demand Models From Multiple Data Sources." In M. Koshi (ed.) *Transportation and Traffic Theory,* Elsevier Science Publishing Co.

Bunch, D., J.J. Louviere, and D. Anderson (1991). "A Comparison of experimental Design Strategies for Multinomial Logit Models: The case of Generic Attributes." Manuscript. University of California, Davis.

Cardell, Nicholas Scott (1988). "The Hedonic Demand Model and some Other Extensions of Multinomial Logit." Ph.D. Dissertation, Harvard University.

Hensher, David and Lester W. Johnson (1981). *Applied Discrete Choice Modelling*. London: Croom Helm.

Hosmer, D.W. and S. Lemeshow (1989). *Applied Logistic Regression*. New York: John Wiley and Sons.

Louviere, J.J. (1984). "Using Discrete Choice Experiments and Multinomial Logit Choice Models to Forecast Trials in a Competitive Environment." *Journal of Retailing,* 60, 81-107.

Louviere, Jordan, J. (1991). "Consumer Choice Models and the Design and Analysis of Choice Experiments." Tutorial presented at American Marketing Association Advanced Research Techniques Forum.

Louviere, J.J. and D.A. Hensher (1982). "On the Design and Analysis of Simulated Choice or Allocation Experiments inn Travel Choice Modeling." *Transportation Research Record,* 890: 11-17.

Louviere, J.J. and D.A. Hensher (1983). "Using Discrete Choice Models with Experimental Design Data to Forecast Consumer Demand for Unique Cultural Event." *Journal of Consumer Research,* 10, 348-361.

Luce, D. R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

McFadden, Daniel (1973). "Conditional Logit Analysis of Qualitative Choice Behavior." In P. Zarembka (ed) *Frontiers in Econometrics*. Academic Press.

McFadden, Daniel (1979). "Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments." In D.A. Hensher and P. R. Stopher (eds) *Behavioral Travel Modelling*. London: Croom Helm.

McFadden, Daniel (1984). "Econometric Analysis of Qualitative Response Models." In Zvi Griliches and M.D. Intrilligator (eds) *Handbook of Econometrics*, Volume III. Elsevier Science Publishers BV.

Mitchell, Robert Cameron and Richard T. Carson (1989). "Using Surveys to Value Public Goods: the Contingent Valuation Method." *Resources for the Future*, Washington, D.C.

Nerlove, Marc and S. James Press (1973). "Univariate and Multivariate Loglinear and Logistic Models." Rand Report No R-1306-EDA/NIH.

Steinberg, Dan and N. Scott Cardell (1992). "Estimating Logistic Regression Models When the Dependent Variable has no Variance." *Communications in Statistics*, 21, 423-450.

Train, Kenneth (1986). *Qualitative Choice Analysis.* MIT Press, Cambridge.

ADDITIONAL REFERENCES

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley and Sons.

Amemiya, Takeshi (1981). "Qualitative Response Models: A Survey." *Journal of Economic Literature*, December, 1483-1536.

Beggs, S. N.S. Cardell and J.A. Hausman (1981). "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics*, 16, 1-19.

Breslow, N. and N.E. Day (1980). *Statistical Methods in Cancer Research, vol.II: The Design and Analysis of Cohort Studies.* Lyon: IARC.

Cardell, Nicholas Scott and Dan Steinberg (1987). "Logistic Regression on Pooled Choice Based Samples and Samples Missing the Dependent Variable." In *American Statistical Association, 1987 Proceedings of the Social Statistics Section.* Alexandria, VA: American Statistical Association, 158-160.

Carson, Richard, Michael Hanemann, and Dan Steinberg (1990). "A Discrete Choice Contingent Valuation Estimate of the Value of Kenai King Salmon." *Journal of Behavioral Economics*, 19, 53-68.

Chamberlain, Gary (1980). "Analysis of Covariance with Qualitative Data." *Review of Economic Studies*, 47, 225-238.

Cox, D.R. and D. Oakes (1984). *Analysis of Survival Data.* New York: Chapman and Hall.

Engel, R. F. (1984). "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics." In Z. Griliches and M. Intrilligator (eds) *Handbook of Econometrics.* New York: North-Holland.

Hausman, Jerry (1978). "Specification Tests in Econometrics." *Econometrica,* 46, 1251-1271.

Hausman, Jerry and Dan McFadden (1984). "A Specification Test for the Multinomial Logit Model." *Econometrica,* 52, 1219-1240.

Louviere, J.J. (1988). "Conjoint Analysis Modelling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity." *Journal of Transport Economics and Policy,* 10, 93-119.

Louviere, J.J. and G.J. Gaeth (1988). "A Comparison of Rating and Choice Responses in Conjoint Tasks." In R.M. Johnson (Ed.). *Sawtooth Software Conference Proceedings.* 59-74.

Louviere, J.J. and G. Woodworth (1983). "Design and Analysis of Correlated Conjoint Experiments using Difference Designs." In Michael J. Houston (Ed.), *Advances in Consumer Research,* 15, 510-517. Provo, UT: Association for Consumer Research.

Luft, H., D. Garnick, D. Peltzman, C. Phibbs, E. Lichtenberg, and S. McPhee (1988). "The Sensitivity of Conditional Choice Models for Hospital Care to Estimation Technique." Draft, Institute for Health Policy Studies. University of California, San Francisco.

Maddala, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

McFadden, Daniel (1982). "Qualitative Response Models." In Werner Hildebrand (ed) *Advances in Econometrics.* Cambridge University Press.

McFadden, Daniel (1976). "Quantal Choice Analysis: A Survey." *Annals of Economic and Social Measurement,* 5, 363-390.

McFadden, Daniel (1986). "The Choice Theory Approach to Market Research." Marketing Science, Vol. 5, No. 4: 275-97.

Manski, Charles and S. Lerman, (1977). "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica,* 8, 1977-1988.

Manski, Charles and Daniel McFadden (eds) (1981). *Structural Analysis of Discrete Data with Econometric Applications.* MIT Press.

Manski, Charles and Daniel McFadden (1980). "Alternative Estimators and Sample Designs for Discrete Choice Analysis." In Manski, Charles and Daniel McFadden (eds) *Structural Analysis of Discrete Data with Econometric Applications.* The MIT Press, Cambridge.

Pregibon, D. (1981). "Logistic Regression Diagnostics. "*Annals of Statistics,* 9, No. 4, 705-724.

Santer, T.J. and D.E. Duffy (1989). *The Statistical Analysis of Discrete Data.* New York: Springer-Verlag.

Steinberg, Dan (1987). *PROC MLOGIT and PROC.* San Diego: Salford Systems.

Steinberg, Dan (1992). *LOGIT: A Supplementary Module for SYSTAT* (Version 2.01). Evanston: SYSTAT Inc.

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50, 1-25.

Wilkinson, Leland (1990). SYSTAT: The System for Statistics (Version 5). Evanston: SYSTAT Inc.

Wrigley, Neil (1985). *Categorical Data Analysis for Geographers and Environmental Scientists.* New York: Longman.

Comment on Steinberg

Carl Finkbeiner
National Analysts, Inc.

Steinberg has done a very creditable job of laying out and describing several variations of logit models. While I acknowledge that there are some applications in which logit may be perfectly appropriate, I have concerns about certain common uses of logit about which I can get quite worked up, and so I will use this opportunity to engage in a little recreational "logit bashing." I shall discuss what I perceive to be problems with logit and other discrete choice models by first contrasting logit with probit and then by comparing conjoint analysis to discrete choice experiments.

LOGIT VERSUS PROBIT

Except in certain special cases, logit is fundamentally a flawed model because of the Independence of Irrelevant Alternatives assumption, referred to by Steinberg. To illuminate this point, consider the chief alternative discrete choice model, probit.

Logit and probit models can both be specified as *utility models*, that is, models which hypothesize that underlying choice is some degree of attractiveness (or utility) to the decision-maker for the object being chosen. ("Choice" is only one application of these models, but I will limit this discussion to "choice" for illustrative convenience; all comments apply as well to other non-choice applications of these models.) Roughly speaking, if the choice is between choosing and not choosing a single object, greater utility is associated with greater likelihood of choosing the object; if the choice is among multiple objects, each with its own underlying utility to the decision maker, then objects with the greatest utility tend to be most likely to be selected.

The two models can be formulated so that the only difference between logit and probit is the set of assumptions about the population distribution of the utility for each choice object. Logit assumes an extreme value distribution — a positively skewed distribution — of utilities and zero correlation among objects on their utilities. This assumption of uncorrelatedness (actually, statistical independence) is called the Independence of Irrelevant Alternatives assumption, pointed out by Steinberg.

Probit assumes a multivariate normal distribution of utilities and allows for different objects having different intercorrelations with one another on their utilities. Interestingly, it has been shown by Bock (1975, pp. 521-522) that the logit model approximates the probit model when all intercorrelations are *constant*, not just zero, so, in effect, the logit assumption about correlations isn't confined to strict independence.

This difference in distributional assumptions raises three issues:

- The skewed distribution assumed by logit makes it better at accommodating positive outliers, though probit can be made to be insensitive to positive or negative outliers by appropriate choice of estimation procedure. Even if probit couldn't be made insensitive to outliers, it is not at all clear to me why the underlying utility distribution ought to be skewed.

- Since the multivariate normal distribution has more parameters for a given application than the extreme value distribution used by logit, probit is more computationally intensive. However, probit is not so intensive as to be impractical, and certainly the complexity of nested logit (designed to make it more realistic) substantially reduces, if not reverses, the difference.

- The most important difference between logit and probit is the assumptions about correlations among objects' utilities. When does it matter if the intercorrelations are constant or varying? Most of the time, in real-world applications. The similarity between objects affects the correlation between them and whenever an object is more like one object than it is like another, the correlations should be different. The most notable exceptions are commodity markets, where all products are alike, and markets with only two products, where there is only one correlation and hence there can't be any differences. In fact, in the case of two objects, probit and logit are remarkably similar in their predictions of probabilities, differing by less than .01 everywhere.

With the exception of special cases like those just mentioned, it is clear that probit's assumption of possibly varying correlations is more realistic than the constant correlation assumption of logit. That difference in assumption can make a very large difference in the predictions made by the two models. Consider a two-product market with products 1 and 2 having shares S_1 and S_2, respectively. We introduce a third product, 3, which is identical to product 1. Logit predicts that the products 1 and 2 lose share in proportion to their original shares, so that the two identical products, 1 and 3, will each have share $S_1/(2S_1+S_2)$ and product 2 will have share $S_2/(2S_1+S_2)$. Probit predicts that the identical products, 1 and 3, will each have share $S_1/2$ and product 2 will have unchanged share, S_2.

To make this concrete, consider that S_1 and S_2 are 40% and 60%, respectively. After introducing the third product identical to product 1, logit predicts the new shares will be 28.6% each for products 1 and 3 and 42.9% for product 2. Probit predicts shares of 20% each for products 1 and 3 and 60% for product 2. Note that logit now predicts that the least popular of the two original products, 1, will, in effect, dominate the market (with a combined share of 57.2% for products 1 and 3) simply by virtue of having a duplicate. This would imply that the way to dominate a market with a product, no matter how weak, is to simply produce more of it.

Steinberg points out this shortcoming of logit and refers to "nested logit" as a solution to the problem. Aside from being significantly more complex to apply than ordinary logit, this approach relies on your ability to accurately specify the tree structure for the product category. If you don't know for sure and guess wrong, your predictions will be distorted.

Other adjustments to eliminate the constant correlation assumption from the logit model have been suggested, notably Smallwood's *ad hoc* adjustment used in ACA (ACA System by Sawtooth Software) (Johnson, 1987), and the related contextual choice model of Lakshmi-Ratan *et al.* (1984). I have commented on the ACA approach previously (Finkbeiner, 1988). The Lakshmi-Ratan

approach does free up the correlation assumption, but appears to hold constant the variances of the utilities, making it still somewhat more specialized than probit. Both adjustments to logit have been limited in application to the conjoint analysis or discrete choice experiment contexts, to be described in the next section.

DISCRETE CHOICE EXPERIMENTS VERSUS CONJOINT ANALYSIS

Logit and probit models are typically applied in discrete choice research where the only data collected are choices among existing products. Choices are then modeled as a function of product and/or respondent characteristics. As such, these models are non-linear forms of multiple regression applied to "found" data, as Steinberg calls it. Regression (linear or non-linear) on "found" data has some important deficits which I have discussed elsewhere in these proceedings. Important among these deficits are:

- Multicollinearity, which leads to instability of the parameters, counter-intuitive results (such as coefficients with the wrong signs), and poor cross-validation.

- Over-reliance on the "world-as-it-is" for establishing the model, making extrapolation to changes not currently present in the market a risky proposition.

- Lack of accommodation of individual differences: for every respondent, the impact of a product (or respondent) attribute on choice is identical.

Logit models on "found" data are no exception as regards these deficits. As Steinberg notes, an attractive alternative is to apply logit to choice data in which the objects are systematically varied attribute bundles, as in conjoint analysis (Louviere & Woodworth, 1983). Since probit could be used to analyze the same choice data, I will follow Louviere in referring to these as discrete choice experiments. In such applications, logit (and discrete choice modeling in general) is no longer susceptible to the first two of the above three problems.

However, for reasons having to do with the weakness of choice data, discrete choice modeling in choice experiments remains a fixed parameter model which doesn't allow for individual differences. Conjoint analysis, as typically practiced, is an individual differences model in which every respondent may have a unique set of partworths to account for that respondent's ratings. Since we know there are usually differences of opinion on how valuable any attribute is, such a model is more realistic.

Steinberg notes logit developments in which the coefficients are allowed to have a distribution across a population. Hausman & Wise (1978) have also provided such developments for both logit and probit models. However, these enhancements may not make all that much difference in aggregate analyses: consider that aggregate partworths produced by conjoint analysis (where the partworths vary across respondents) have in the past usually been similar to the discrete choice experiment partworths, which do not vary across respondents (Louviere & Gaeth, 1988). In addition, a great deal of similarity is found in the predictions of preference shares for holdout configurations (Oliphant, *et al.*, 1992).

Where allowing for variation in coefficients (or partworths) would be helpful is at the *dis*aggregate level, where we wish to identify differences among individuals in which attributes are most influential,

either *a priori* (through crosstabulation) or *post hoc* (via cluster analysis). The Hausman & Wise model assumes that distributions on the coefficients are unimodal (in fact, normality is assumed), thus not accommodating the possibility of multimodality (that is, segmentation) on which attributes are most important, unless we plan the research so as to be able to estimate a separate model for every identifiable subgroup we wish to consider.

I would like to comment on an argument often used in favor of discrete choice experiments: namely, that, in contrast to the rating task of conjoint, the choice task of discrete choice experiments is more like purchase since one object out of many alternatives is selected. I would maintain that *neither* task is like actual purchase, in that both involve an artificial context with complete information about hypothetical products, and since neither task involves the exchange of money for goods/services.

Beyond that, I view conjoint ratings as being considerably stronger than choice judgments, without sacrificing any information. If the conjoint task is comparative, as it ought to be, then it may be thought of as providing information not only about which conjoint product profile the respondent likes the most (analogous to a choice decision), but also about how much each profile is liked in comparison to all other conjoint profiles. I don't see any theoretical advantage to choice tasks here.

However, I readily admit that a choice may be easier for respondents than a conjoint rating. Unfortunately, that advantage is eaten up in the need for many choice judgments from each respondent and is eliminated if the final model does not allow for individual differences.

CONCLUSION

As you can no doubt guess from the preceding, I see no advantage for discrete choice experiments in the constructed experiment context. Conjoint analysis has the advantage of explicitly modeling individual differences in a way that lends itself well to a segmentation view of markets. I have not seen any disadvantages for conjoint analysis that do not apply equally well to discrete choice experiments. In any event, logit need not be the model used with discrete choice data: probit can be used as well.

So, when is logit appropriate? Bear in mind that there are many contexts in which logit is an analytical option, including situations in which classification into categories is desired (where discriminatory analysis is an option), and analysis of relationships among categorical variables (where log-linear modeling is an alternative). I recognize that logit-type models may be appropriate in those other contexts, but I will confine my remarks here to the contexts described in this commentary: the prediction of group membership or choice, either based on "found" data or on constructed experiments.

Both logit and probit are frequently options. In the binary outcome case (for example, choosing or not choosing a product), although logit and probit are equivalent models, logit does provide simpler and more efficient computations. In the multiple outcome case, probit can always be formulated to apply to the same data to which logit applies and probit is generally a more reasonable model. In fairness, remember that there are some special circumstances (for example, commodity markets) where logit and probit may be equivalent.

The major limitation of probit is the lack of widely available and easily understood software for implementing it. In some instances, probit is computationally less practical than logit, although the enhancements required of logit to make it more realistic also add to its computation costs. I recognize that there are some situations in which it is simply not practical to invent and implement the appropriate probit model, and so the more restrictive assumptions of logit must sometimes be accepted for the sake of practicality.

REFERENCES

Bock, R.D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Finkbeiner, C.T. (1988). "Comparisons of Conjoint Choice Simulators." *Sawtooth Software Conference Proceedings*, 75-103.

Hausman, J.A. and D.A. Wise (1978). "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." *Econometrica*, 46 403 - 426.

Johnson, R.M. (1987). "Adaptive Conjoint Analysis," working paper, Sawtooth Software, Inc., Ketchum, Idaho, March.

Lakshmi-Ratan, R.A., S. Chaiy and J. May (1984). "Mathematical Modelling of Contextual Effects on Individual Choice Behavior: Axiom and Model of Contextual Choice." working paper, University of Wisconsin, Graduate School of Business, September.

Louviere, J.J. and G.J. Gaeth (1988). "A Comparison of Rating and Choice Responses in Conjoint Tasks." *Sawtooth Software Conference Proceedings*, 59-74.

Louviere, J.J. and G.G. Woodworth (1983). "Design and Analysis of Simulated Choice or Allocation Experiments: An Approach Based on Aggregate Data." *Journal of Marketing Research*, 20 (1983), 350-367.

Oliphant, K., T.C. Eagle, J. Louviere, and D. Anderson (1992). "Cross-Task Comparison of Ratings-Based and Choice-Based Conjoint." *Sawtooth Software Conference Proceedings*, (this volume).

TREE STRUCTURED DATA ANALYSIS: AID, CHAID, AND CART

Leland Wilkinson
SYSTAT, Inc. and Northwestern University

INTRODUCTION

Trees are directed graphs beginning with one node and branching to many. They are fundamental to computer science (data structures), biology (classification), psychology (decision theory), and many other fields. Classification and regression trees are used for prediction. In the last two decades, they have become popular as alternatives to regression, discriminant analysis, and other procedures based on algebraic models.

Tree fitting methods have become so popular that several commercial programs now compete for the attention of market researchers and others looking for software. Different commercial programs produce different results with the same data, however. Worse, some programs provide no documentation or supporting materials to explain their algorithms. The result is a marketplace of competing claims, jargon, and misrepresentation. Reviews of these packages (for example, Levine, 1991; Simon, 1991) have used words like "sorcerer," "magic formula," and "wizardry" to describe the algorithms and have expressed frustration at vendors' scant documentation. Some vendors, in turn, have represented tree programs as state-of-the-art "artificial intelligence" procedures capable of discovering hidden relationships and structures in databases.

Despite the marketing hyperbole, most of the now popular tree fitting algorithms have been around for decades. The modern commercial packages are mainly microcomputer ports (with attractive interfaces) of the mainframe programs which originally implemented these algorithms. Warnings of abuse of these techniques are not new either (for example, Einhorn, 1972; Bishop, Fienberg, and Holland, 1975). Originally proposed as automatic procedures for detecting interactions among variables, tree fitting methods are actually closely related to classical cluster analysis (Hartigan, 1975). This paper will attempt to sort out some of the differences between algorithms and illustrate their uses on real data. In addition, tree analyses will be compared to discriminant analysis and regression. This is *not* a product comparison. Compiling a chart comparing features among different tree programs would be almost impossible because the manuals are not always clear on whether a feature is present or correctly implemented and each manual uses idiosyncratic language. This tutorial instead should help you to ask the right questions.

THE TREE MODEL

Figure 1 shows a tree for predicting decisions by a medical school admissions committee (Milstein *et al.*, 1975). It was based on data for a sample of 727 applicants. Notice that the values of the predicted variable (admissions decision to reject or interview) are at the *bottom* of the tree and the predictors (Medical College Admissions Test and College Grades) come into the system at each node of the tree. The top node contains the entire sample. Each of the remaining nodes contains a subset of the sample in the node directly above it. Furthermore, any node contains the sum of the

samples in the nodes connected to and directly below it. The tree thus *splits* samples. Each node can be thought of as a cluster of objects (cases) which is to be split by further branches in the tree. The numbers with asterisks below the terminal nodes show how many cases are *incorrectly* classified by the tree. A similar tree data structure is used for representing the results of single and complete linkage and other forms of hierarchical cluster analysis (Hartigan). Tree prediction models add two ingredients: the predictor and predicted variables labeling the nodes and branches.

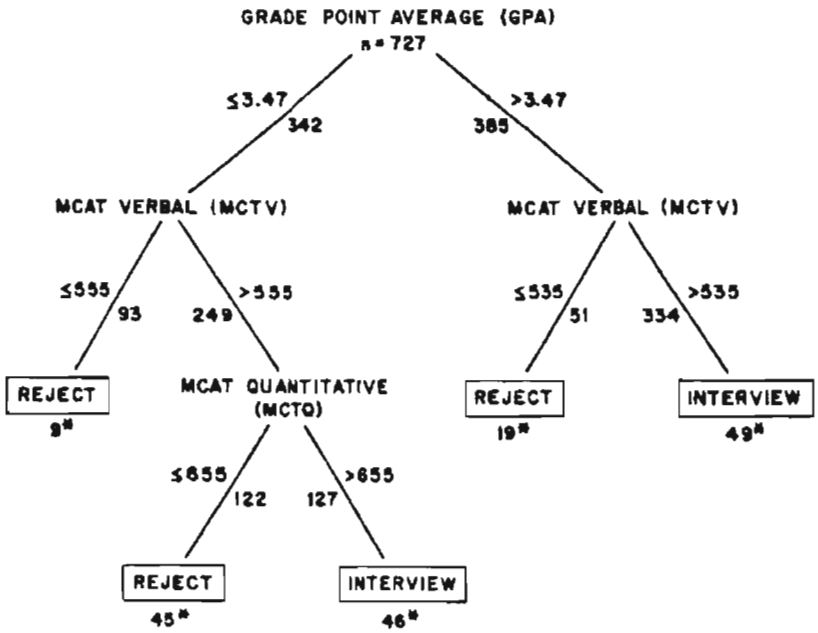


Figure 1. Tree for predicting admissions to medical school.

The tree in Figure 1 is *binary* because each node is split into only two subsamples. Classification or regression trees need not be binary, but most are. Despite the marketing claims of some vendors, non-binary, or multi-branch trees are not superior to binary trees. Each is a permutation of the other. Figure 2 shows this. The tree on the left in Figure 2 is not more parsimonious than that on the right. Both trees have the same number of parameters (split points), and any statistics associated with the tree on the left can be converted trivially to fit the one on the right. A computer program for scoring either tree (IF...THEN...ELSE) would look identical. For display purposes, however, it is often convenient to collapse binary trees into multi-branch trees, but this is not necessary.

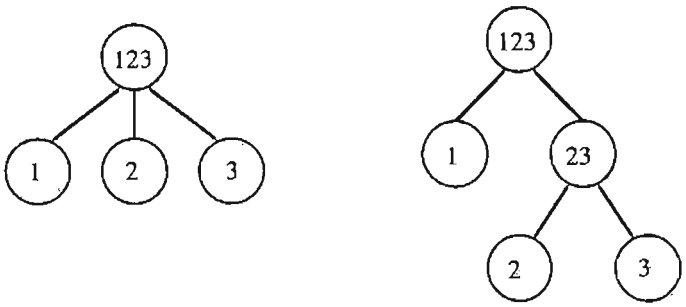


Figure 2. Ternary (left) and binary (right) trees.

Some programs which do multi-branch splits do not allow further splitting on a predictor once it has been used. This has an appealing simplicity but it can lead to unparsimonious trees. It is unnecessary to make this restriction before fitting a tree. Figure 3 shows an example of this problem. The upper right tree classifies objects on an attribute by splitting once on shape, then fill, then again on shape. This allows the algorithm to separate the objects into only four terminal nodes having common values. The upper left tree splits on shape, then only on fill. By not allowing any further splits on shape, the tree requires five terminal nodes to classify correctly. This problem cannot be solved by splitting first on fill, as the lower left tree shows. In either case, restricting the split to only one predictor per branch results in more terminal nodes.

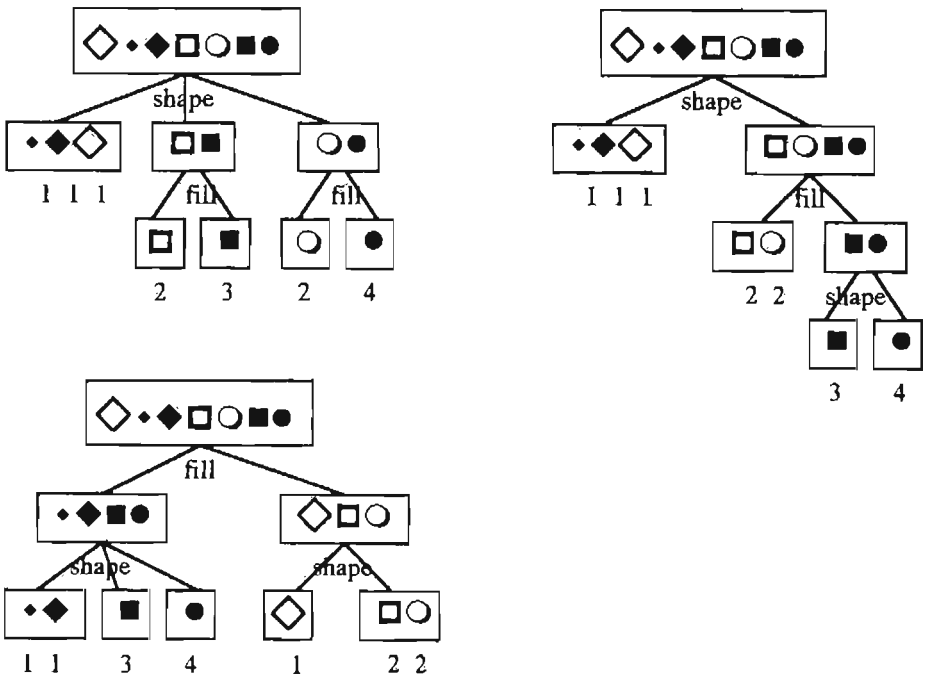


Figure 3. Multi-branch trees with only one split per predictor in each branch (left) and binary tree with multiple splits per predictor in each branch (right)

CATEGORICAL OR QUANTITATIVE PREDICTORS

The predictor variables in Figure 1 are quantitative, so splits are created by determining cut points on a scale. If predictor variables are categorical as in Figure 3, splits are made between categorical values. It is not necessary to categorize predictors before computing trees. This is as dubious as turning data well-suited for regression into categories in order to use chi-square tests. Those who recommend this practice are turning silk purses into sows' ears. In fact, if variables are categorized before doing tree computations, then poorer fits are likely to result. Algorithms are available for mixed quantitative and categorical predictors, analogous to analysis of covariance.

REGRESSION TREES

Morgan and Sonquist (1963) proposed a simple method for fitting trees to predict a quantitative variable. They called the method AID, for Automatic Interaction Detection. The algorithm performs stepwise splitting. It begins with a single cluster of cases and searches a candidate set of predictor variables for a way to split this cluster into two clusters. Each predictor is tested for splitting as follows: sort all the n cases on the predictor and examine all n-1 ways to split the cluster in two. For each possible split, compute the within cluster sum of squares about the mean of the cluster on the dependent variable. Choose the best of the n-1 splits to represent the predictor's contribution. Now do this for every other predictor. For the actual split, choose the predictor and its cut point that yields the smallest overall within cluster sum of squares.

Categorical predictors require a different approach. Since categories are unordered, all possible splits between categories must be considered. For deciding on one split of k categories into two groups, this means 2^k-1 possible splits must be considered. Once a split is found, its suitability is measured on the same within cluster sum of squares as for a quantitative predictor.

Morgan and Sonquist called their algorithm AID because it naturally incorporates interaction among predictors. Interaction is not correlation. It has to do instead with conditional discrepancies. In the analysis of variance, interaction means that a trend within one level of a variable is not parallel to a trend within another level of the same variable. In the ANOVA model, interaction is represented by cross-products between predictors. In the tree model, it is represented by branches from the same node which have different splitting predictors further down the tree. Figure 4 shows a tree without interactions on the left and with interactions on the right. Because interaction trees are a natural byproduct of the AID splitting algorithm, Morgan and Sonquist called the procedure "automatic." In fact, AID trees without interactions are quite rare for real data, so the procedure is indeed automatic. To search for interactions using stepwise regression/ANOVA linear modeling, we would have to generate 2^p interactions among p predictors and compute partial correlations for every one of them in order to decide which ones to include in our final model.

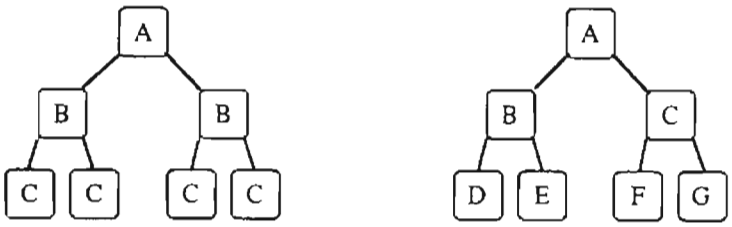


Figure 4. No interaction (left) and interaction (right) trees.

CLASSIFICATION TREES

Regression trees parallel regression/ANOVA modeling, where the dependent variable is quantitative. Classification trees parallel discriminant analysis and algebraic classification methods. Kass (1980) proposed a modification to AID called CHAID for categorized dependent and independent variables. His algorithm incorporated a sequential merge and split procedure based on a chi-square test statistic. Kass was concerned about computation time (although this has since proved an unnecessary worry), so he decided to settle for a suboptimal split on each predictor instead of searching for all possible combinations of the categories.

Kass's algorithm is like sequential crosstabulation. For each predictor: 1) crosstabulate the categories of the predictor with the categories of the dependent variable, 2) find the pair of categories of the predictor whose 2 x k sub-table is least significantly different and merge these categories if the chi-square test statistic is not significant according to a preset critical value; repeat this merging process until no non-significant chi-square is found for a sub-table, and 3) pick the splitting variable whose chi-square is largest and continue splitting as with AID.

The CHAID algorithm saves computer time, but like stepwise regression, it is not guaranteed to find the splits which predict best. Only all possible subsets regression or exhaustive search of category subsets can do that. It is also limited to categorical predictors, so it cannot be used for quantitative or mixed categorical-quantitative models, as in Figure 1. Nevertheless, it is an effective way to search heuristically through rather large tables quickly.

It is interesting that within the computer science community there is a categorical splitting literature which does not cite the statistical work and is, in turn, not cited by statisticians. Quinlan (1986), the best known of these researchers, developed a set of algorithms based on information theory. These methods, termed ID3, iteratively build induction trees based on training samples of attributes. They do not include much consideration of the statistical error issues.

PRUNING AND CROSS VALIDATION

AID, CHAID, and other forward sequential tree fitting methods share a problem with other tree clustering methods — where do we stop? If we keep splitting, a tree will end up with only one case or object at each terminal node. We need a method for producing a smaller tree than the exhaustive

one. One way is to use stepwise statistical tests, as in the *F*-to-enter rule for forward stepwise regression. We compute a test statistic (such as chi-square or *F*), choose a critical level for the test (sometimes modifying it with the Bonferroni inequality), and stop splitting any branch which fails to meet the test. Most programs work this way.

Breiman *et al.* (1984) showed that this method tends to yield trees with too many branches and can also fail to pursue branches which can add significantly to the overall fit. They advocate, instead, *pruning* the tree. After computing an exhaustive tree, their program eliminates nodes which do not contribute to the overall prediction. They add another essential ingredient, however: the *cost of complexity*. This measure is similar to other cost statistics, such as Mallows' C_p (for example, Neter, Wasserman, and Kutner, 1985), which add a penalty for increasing the number of parameters in a model. Regardless of how a tree is pruned, it is important to cross validate it. As with stepwise regression, the prediction error for a tree applied to a new sample can be considerably higher than for the training sample on which it was constructed. Whenever possible, data should be reserved for cross validation.

The Breiman *et al.* CART program is the only tree builder which incorporates a thoroughly developed probabilistic model. All the other programs approach the problem with *ad hoc* procedures adapted from forward selection subset regression methodology. They terminate splitting based on chi-square or *F* tests which are sometimes Bonferroni adjusted, but as Wilkinson (1979) and others showed for stepwise regression, these subset distributions are not generally close to the unconditional ones that authors wish to adopt.

AN APPLICATION

Let's look at how these issues apply to a real dataset. Figure 5 shows a scatterplot matrix (SPLOM) of number of deaths per 100,000 people by 7 different afflictions in the U.S. in 1985. We are going to try to predict region of the country from these death statistics. This will be a classification (as opposed to regression) problem, because region is already coded in four categories (1=Northeast, 2=Midwest, 3=South, 4=West).

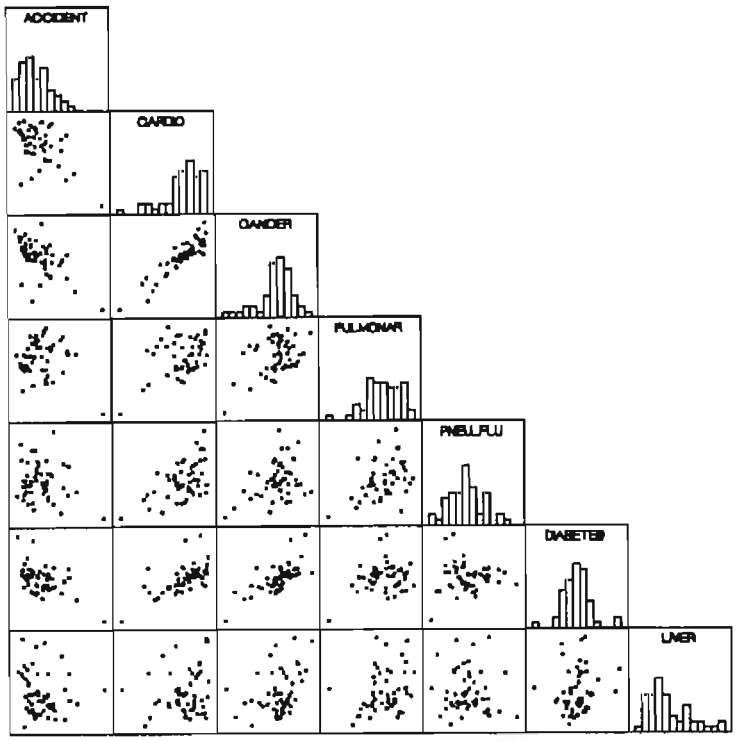


Figure 5. Scatterplot Matrix of U.S. Mortality Data

Fortunately, most of the predictor variables are not highly correlated. Tree fitting programs have the same problems with correlated predictors that linear models do. We might expect, for example, that a tree fit by CARDIO would predict about as well as one with CANCER substituted for CARDIO, because both variables show similar distributions throughout the SPLOM and throughout Figure 6 as well. Because of this substitutability, and the possibility of other influential variables outside the system, it would be misleading to conclude that the tree we fit will give us any insight into causation.

Figure 6 shows a box/dot plot of these data broken down by region of the country. We can see distinct regional differences in the death statistics. Western states have high accident rates, for example, though relatively few deaths from cardio-vascular disorders. Notice, also, that CARDIO and CANCER appear similar in these plots as they do in the SPLOM.

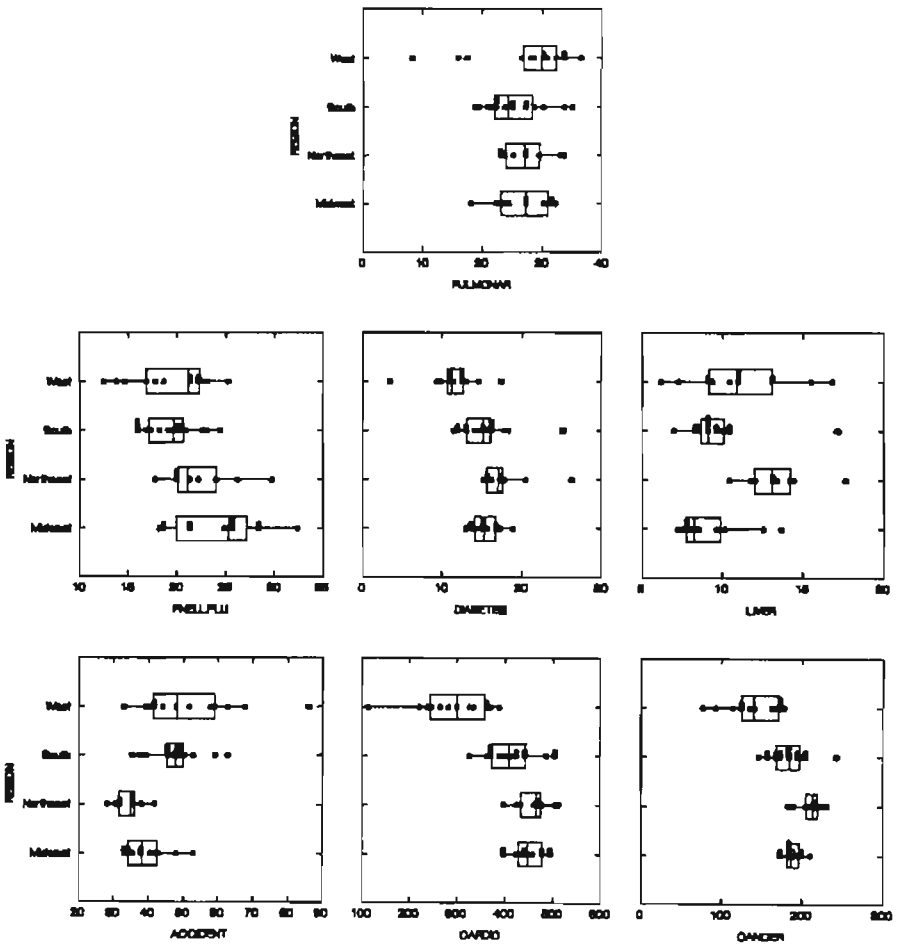


Figure 6. Box/Dot Plot of Mortality Data by Region of Country

One way of understanding how tree models classify data is to compare discriminant analysis and classification trees on the same data. Figure 7 shows a three dimensional scatterplot of ACCIDENT versus CARDIO versus LIVER. These three variables were selected from the seven predictors by a tree fitting algorithm based on a Gini-type index of badness of fit. The Gini index compares all possible pairs of cases to construct a single measure of diversity within nodes. The plotting symbols are taken from the first letter of each of the regions: Northeast, South, Midwest, West.

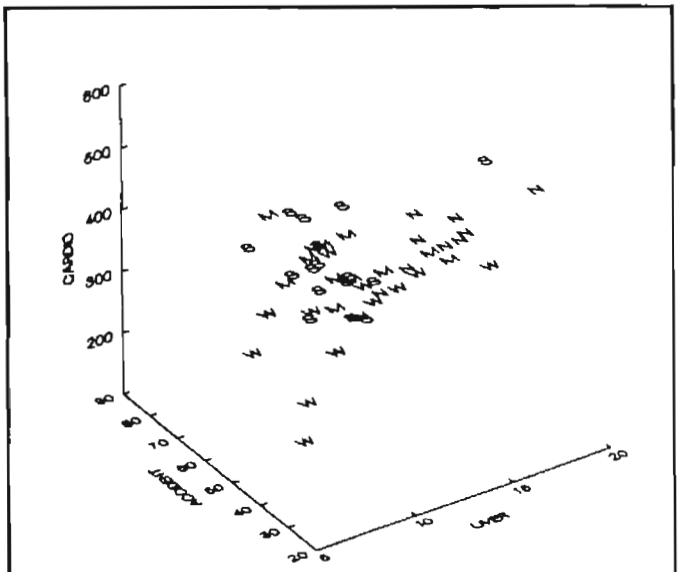


Figure 7. 3-D Plot of CARDIO versus ACCIDENT versus LIVER

Figure 8 shows how the states are split by a linear discriminant analysis on these variables. The cutting planes are positioned roughly halfway between each pair of region centroids. Their orientation was determined by the discriminant analysis. With three predictors and four groups, there are six cutting planes, although only four show in the figure. Only the states classified in the Midwest region are shown in the figure. Notice that there are three misclassifications: two North and one Southern states.

Figure 9 shows how the tree fitting algorithm cuts the data. Notice that the cutting planes are parallel to the axes. While this would seem to restrict the discrimination compared to the more flexible angles allowed by the discriminant planes, the tree model allows interactions between variables which do not ordinarily appear in the discriminant model. Notice, for example, that for ACCIDENT less than 45, the cases are split by LIVER into the Midwest and North. For ACCIDENT greater than 45, on the other hand, the split leaves the South classified.

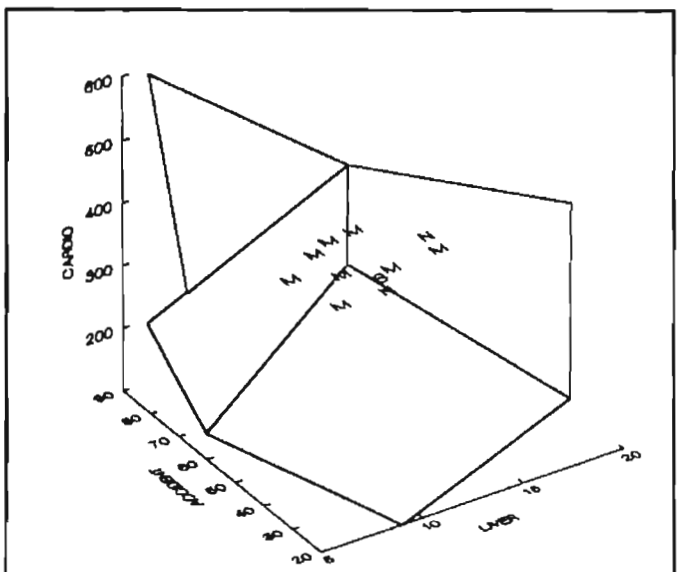


Figure 8. Cutting Planes for Discriminant Model. Only cases for Midwest prediction region are shown.

Tree models are not usually related to dimensional plots in this way, but it is helpful to see that tree methods have a geometric twin. Similarly, we can construct algebraic expressions for trees. They would require dummy variables for any categorical predictors and interaction (product) terms for every split whose descendants (lower nodes) did not involve the same variables on both sides.

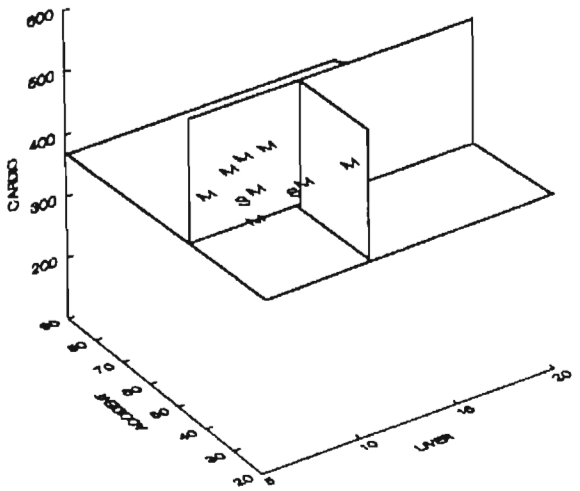


Figure 9. Cutting Planes for Tree Model.
Only cases for Midwest prediction region are shown.

Now let's look at the discriminant analysis results before seeing the tree for these data. Figure 10 shows the discriminant analysis output using the three predictors from the tree model. The codes used were Northeast (1), Midwest (2), South (3), and West (4). Notice that LIVER is the most important variable for discriminating the Midwest. This is true also in the tree model.

The classification table is at the bottom of the panel. The model correctly discriminates all but 11 cases in the sample. Using all the variables in a discriminant function correctly classifies an additional 5 cases, but with some cost in complexity.

CANONICAL LOADINGS (CORRELATIONS BETWEEN CONDITIONAL DEPENDENT VARIABLES AND DEPENDENT CANONICAL FACTORS)

	1	2	3
ACCIDENT	0.413	0.559	0.719
CARDIO	-0.903	-0.217	0.370
LIVER	0.083	-0.846	0.527

GROUP CLASSIFICATION FUNCTION COEFFICIENTS

	1	2	3	4
ACCIDENT	0.720	0.657	0.805	0.794
CARDIO	0.178	0.168	0.167	0.115
LIVER	0.566	1.367	0.768	1.348

GROUP CLASSIFICATION CONSTANTS

1	2	3	4
-58.217	-60.371	-58.804	-46.093

TABLE OF GROUP (ROWS) BY PREDICT (COLUMNS)

FREQUENCIES

	1.000	2.000	3.000	4.000	TOTAL
1.000	9	2	1	0	12
2.000	1	8	0	0	9
3.000	3	1	11	1	16
4.000	0	0	2	11	13
TOTAL	13	11	14	12	50

Figure 10. Discriminant analysis of U.S. mortality data.

Figure 11 shows the results of a tree model analysis of these data. Notice the asymmetry of the tree. The Western region is broken out immediately by splitting CARDIO<365.3. Of the higher cardiovascular death regions, higher accident rates are more common in the South (ACCIDENT>45). And more deaths by liver disorders are associated with the Northeast (LIVER>10.4).

The PRE statistic at the top of the figure is a "proportional reduction of error" measure. It behaves like an R^2 statistic, varying between 0 and 1, with 1 meaning perfect classification. The final PRE for this tree is .556, which is analogous to having more than 50 percent of the sample variance of the dependent variable accounted for by the predictors in a regression model.

Finally, let's look at CART's cross-validation output. Figure 12 shows the CART analysis for a ten-fold cross validation. These computations are done by drawing 10 samples each of 40 cases from the 50 cases (states) in the dataset. In general, CART uses a resampling scheme to get an estimate of cross-validation error by computing separate trees for the subsamples. This procedure is similar to the bootstrap, in which subsamples are used to derive estimates of standard errors based on a single dataset. CART's output shows that the tree fitted performs fairly stably across samples. The misclassification is not substantially worse for the cross-classified samples than for the learning sample.

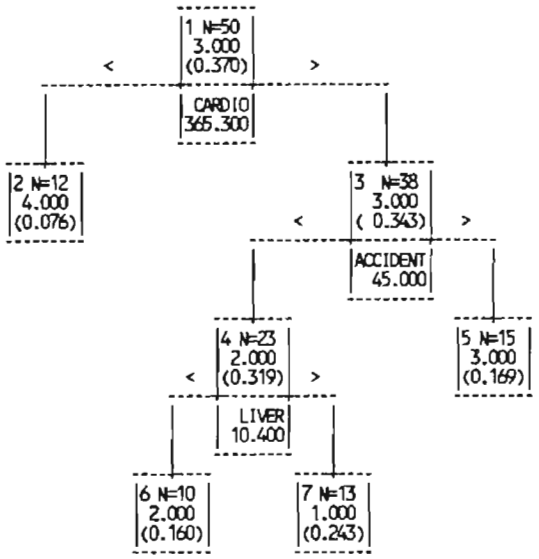
CONCLUSION

This paper cannot cover more than the basics of tree modeling. Its goal is to introduce the basic issues and allow you to distinguish heuristic, or suboptimal tree fitting algorithms from exhaustive search methods. In addition, the real data analysis should give some idea of the importance of resampling methods, particularly for samples as small as this one. For any datasets fewer than 900 or 1000 cases, fitting trees without a resampling estimate or cross validation can be risky.

```
Split Variable              PRE  Improvement
    1   CARDIO            0.246        0.246
    2 ACCIDENT            0.416        0.170
    3    LIVER            0.556        0.140

Fitting Method:              Gini Index
Predicted variable:  REGION
Minimum split index value:       0.100
Minimum improvement in PRE:      0.100
Maximum number of nodes allowed:  22
Minimum count allowed in each node:   5
The final tree contains   4 terminal nodes.
Proportional Reduction in Error (PRE):       0.556

Node From Count       Mode    Impurity   Split Var   Cut Value       Index
   1    0    50      3.000       0.370   CARDIO         365.299       0.246
   2    1    12      4.000       0.076
   3    1    38      3.000       0.342 ACCIDENT         45.000       0.241
   4    3    23      2.000       0.319    LIVER         10.399       0.353
   5    3    15      3.000       0.168
   6    4    10      2.000       0.160
   7    4    13      1.000       0.242
```




```
Tree for predicting REGION

IF   CARDIO     <      365.299 THEN FOR
  LET ESTIMATE =        4.000
NEXT
ELSE FOR
  IF ACCIDENT     <       45.000 THEN FOR
    IF    LIVER     <       10.399 THEN FOR
      LET ESTIMATE =        2.000
    NEXT
    ELSE FOR
      LET ESTIMATE =        1.000
    NEXT
  NEXT
  ELSE FOR
    LET ESTIMATE =        3.000
  NEXT
NEXT

TABLE OF   REGION     (ROWS) BY ESTIMATE     (COLUMNS)

FREQUENCIES
          1.000     2.000     3.000     4.000     TOTAL
  1.000       9         0         0         0         9
  2.000       2         8         2         0        12
  3.000       1         2        12         1        16
  4.000       1         0         1        11        13
TOTAL        13        10        15        12        50
```

Figure 11. Classification tree analysis for U.S. mortality data

****
MISCLASSIFICATION BY CLASS
****

Class	Prior Prob.	CROSS VALIDATION N	N Mis-Classified	Cost	LEARNING SAMPLE N	N Mis-Classified	Cost
1	0.180	9	2	0.222	9	0	0.000
2	0.240	12	7	0.583	12	4	0.333
3	0.320	16	6	0.375	16	4	0.250
4	0.260	13	2	0.154	13	2	0.154
Tot	1.000	50	17		50	10	

****
CROSS VALIDATION CLASSIFICATION TABLE
****

ACTUAL CLASS	PREDICTED CLASS 1	2	3	4	ACTUAL TOTAL
1	7.000	1.000	1.000	0.000	9.000
2	5.000	5.000	2.000	0.000	12.000
3	1.000	0.000	10.000	5.000	16.000
4	1.000	0.000	1.000	11.000	13.000
PRED. TOT.	14.000	6.000	14.000	16.000	50.000
CORRECT	0.778	0.417	0.625	0.846	
SUCCESS IND.	0.598	0.177	0.305	0.586	
TOT. CORRECT	0.660				

****
CROSS VALIDATION CLASSIFICATION PROBABILITY TABLE
****

ACTUAL CLASS	PREDICTED CLASS 1	2	3	4	ACTUAL TOTAL
1	0.778	0.111	0.111	0.000	1.000
2	0.417	0.417	0.167	0.000	1.000
3	0.063	0.000	0.625	0.313	1.000
4	0.077	0.000	0.077	0.846	1.000

****
LEARNING SAMPLE CLASSIFICATION TABLE
****

ACTUAL CLASS	PREDICTED CLASS 1	2	3	4	ACTUAL TOTAL
1	9.000	0.000	0.000	0.000	9.000
2	2.000	8.000	2.000	0.000	12.000
3	1.000	2.000	12.000	1.000	16.000
4	1.000	0.000	1.000	11.000	13.000
PRED. TOT.	13.000	10.000	15.000	12.000	50.000
CORRECT	1.000	0.667	0.750	0.846	
SUCCESS IND.	0.820	0.427	0.430	0.586	
TOT. CORRECT	0.800				

****
LEARNING SAMPLE CLASSIFICATION PROBABILITY TABLE
****

ACTUAL CLASS	PREDICTED CLASS 1	2	3	4	ACTUAL TOTAL
1	1.000	0.000	0.000	0.000	1.000
2	0.167	0.667	0.167	0.000	1.000
3	0.063	0.125	0.750	0.063	1.000
4	0.077	0.000	0.077	0.846	1.000

Figure 12. CART output for cross-validation.

REFERENCES

Bishop, Y.M., S.E. Fienberg, and P.W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Einhorn, H. (1972). "Alchemy in the Behavioral Sciences." *Public Opinion Quarterly*, 3, 367-378.

Comment on Wilkinson

Thomas L. Pilon
TRAC, Inc./University of North Texas

I like tree-structured data analysis. The first time that I used the technique, I became a fan "overnight." I was conducting a market segmentation analysis with a major client and we were having considerable difficulty developing a satisfactory segmentation scheme. I remembered that I had recently received a review copy of SPSS/PC+CHAID. Although I was not wild about the idea of learning and trying a new software product in the middle of the night (literally) with a client breathing down my neck (literally), we were desperate. After reading the promotional literature on CHAID and deciding that it had a reasonable chance of helping, we took the leap. A few hours later (very early morning) we had a segmentation scheme that worked and a nice looking laser tree diagram, as well.

For a fairly detailed example of using tree-structured data analysis (TSDA) for market segmentation see Pilon and Witt (1991). Allison and Forsyth (1990), Green (1978), and Magidson (1988a, 1988b) also include marketing examples of TSDA.

TO TREE OR NOT TO TREE

Although I have successfully utilized tree-structured data analysis in numerous studies since the above referenced night of desperation, my enthusiasm is not unbridled. I do not view TSDA as a marketing research data analysis panacea any more than I do any other technique.

One thing that disturbs me about TSDA is that often the technique must make some very close calls when choosing among two or more possible splits. Furthermore, each of the candidate splits may result in a very different looking tree. For example, a few years ago I was analyzing the data on a study in which there were nearly 10,000 respondents. One possible split was significant at the 1.0E-37 level while another was significant at the 1.1E-37 level. If one respondent had responded differently or if there was one data entry error (imagine that), the tree would have split differently and resulted in a considerably different final tree. As a result of this and other similar experiences, I prefer to use TSDA as an exploratory analysis tool to reduce a large number of variables down to a manageable number and then further examine the resulting variables with a detailed analysis technique such as discriminant analysis or logit analysis. See Pilon and Witt for a complete discussion.

AVOID THE DATA MINING TRAP: HOLDOUT, VALIDATE, CROSS-VALIDATE, AND/OR REPLICATE

As with all data analysis methods, a prudent researcher should always validate the results of TSDA. This canon is particularly true for those methods (such as TSDA) that search the data for structure (as opposed to those that are testing a pre-defined theory). At the very least, a portion of sample

should be held out of the analysis and later checked for fit against the final model. Cross-validation, in which numerous hold-out samples are taken one at a time, is better yet. Finally, if at all possible, the sample should be replicated and checked for fit on the final model.

TREE-STRUCTURED DATA ANALYSIS SOFTWARE

In addition to the three software packages mentioned in Leland Wilkinson's paper (AID, CHAID, and CART), a fourth package, Knowledge Seeker, should have been included. I had intended to include a comprehensive product feature matrix of the four major packages in this section. However, two of the four packages (CHAID and CART) will release significant upgrades in the next few months and detailed information is not yet available. Therefore, I decided against publishing information that would have a very short (if any) shelf life. See Struhl (1992) for a review of AID, CHAID, and CART that is based on the current releases of the software.

REFERENCES

Allison, Neil and John Forsyth (1990). "Describing Needs-Based Segments." *American Marketing Association Advanced Research Techniques Forum Proceedings*, 238-269.

Green, Paul E. (1978). "An AID/Logit Procedure for Analyzing Large Multiway Contingency Tables." *Jounal of Marketing Research*, 15 (February), 132-136.

Magidson, Jay (1988a). *CHAID, LOGIT, and Log-Linear Modeling*. Datapro Research Corporation, (May).

Magidson, Jay (1988b). "Improved Statistical Techniques for Response Modeling." *Journal of Direct Marketing*, 2 (Autumn), 6-18.

Pilon, Thomas L. and Karlan J. Witt (1991). "Making Cluster-Based Needs Segments Actionable." *Proceedings of the Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis, and Computer Interviewing*, 97-126.

Struhl, Steven (1992). "Classification Tree Methods: AID, CHAID, and CART." *Quirk's Marketing Research Review*, (February), 10-13.

SOFTWARE REFERENCES

AID
PC-MDS Multidimensional Statistics Package
Institute of Business Management
Brigham Young University
Provo, Utah 84602
(801) 378-5569

CART
SYSTAT
SYSTAT, Inc.
1800 Sherman AVenue
Evanston, Illinois 60201
(708) 864-5670

CHAID
SPSS/PC+
SPSS Inc.
444 N. Michigan
Chicago, Illinois 60611
(312) 329-3300

Knowledge Seeker
FirstMark Technologies Ltd
14 Concourse Gate, Suite 600
Ottawa, Ontario, Canada K2E 7S8
(800) 387-7335

PROCEEDINGS VOLUMES FROM PREVIOUS CONFERENCES

1991

Interviewer Error: A Comparison of Methods
Thomas L. Van Valey and Sue R. Crull, Kercher Center for Social Research, Western Michigan University

Perspectives on Data Quality
The Client and the Marketing Researcher
Rebecca Elmore-Yalch, President, MACS, Inc.
Jane Glascock, Metro Transit, Seattle

Quality in Computer Interviewing — Tricks of the Trade
Emily Wyatt, Burke Marketing Research

Controlling Non-Response Bias and Item Non-Response Bias Using Computer-Assisted Telephone Interviewing (CATI) Techniques
Michael Sullivan, Freeman, Sullivan & Company

Of Mice and Men, and Women
Ray Poynter, Sandpiper International Limited

A Comparison of Factor Analysis and Correspondence Analysis as Data Reduction Techniques with Clustering in Market Segmentation
Marsha A. Wilcox, Decision Research Corp.

Making Cluster-Based Needs Segments Actionable
Thomas L. Pilon and Karlan J. Witt, IntelliQuest, Inc.

CONJOINT ANALYSIS

Measuring Brand Equity with Conjoint Analysis
Douglas L. MacLachlan, University of Washington
Michael G. Mulhern, Mulhern & Associates

Using Adaptive Conjoint Analysis for the Development of Lottery Games — an Iowa Lottery Case Study
Philip S. Kopel, Kopel Research Group, Inc.
Dale Kever, Iowa Lottery

Using Conjoint Analysis for Price Optimization
Richard D. Smallwood, Applied Decision Analysis, Inc.

Using Conjoint Information: Organizational Factors
Sharon W. Salling, The Dun & Bradstreet Corporation
John Deegan, Receivable Management Services a company of The Dun & Bradstreet Corporation

Validation of Adaptive Conjoint Analysis (ACA) Versus Standard Concept Testing
James J. Tumbusch, MarketVision Research, Inc.

An Empirical Comparison of ACA and Full Profile Judgments
Joel C. Huber, Duke University
Dick R. Wittink, Cornell University
John Fiedler, POPULUS, Inc.
Richard Miller, Consumer Pulse, Inc.

Unreliable Respondents in Conjoint Analysis: Their Impact and Identification
Keith Chrzan, Walker: Research and Analysis

Modeling Preference in Conjoint Measurement
Paul F. Hase, Hase/Schannen Research

Scaling Prior Utilities in Sawtooth Software's Adaptive Conjoint Analysis
William G. McLauchlan, McLauchlan & Associates, Inc.

Including Interactions in Conjoint Models
Carl Finkbeiner and Pilar Lim, National Analysts Division of Booz·Allen & Hamilton, Inc.

A Validation Study of Sawtooth Software's Adaptive Conjoint Analysis (ACA)
Paul E. Green, The Wharton School, University of Pennsylvania
Catherine M. Schaffer, Marketing Department, University of Denver
Karen M. Patterson, Innovative Research, Incorporated

Conferences were held every 18 months after 1989; there are no 1990 Proceedings.

1989

THE DISK-BY-MAIL SURVEY AS A COMPETITIVE TOOL

Disk-By-Mail Surveys: Three Years' Experience
Brant Wilson, Compaq Computer Corporation

Customer Satisfaction Research Using Disks-By-Mail
Peter Zandan and Lucy Frost, IntelliQuest, Inc.

Context-Specific Choice Experiments for Multi-Featured Products: A Disk-By Mail Survey Application
Shari Gershenfeld and Terry Atherton, Cambridge Systematics, Inc.
Moshe Ben-Akiva, MIT
Larry Musetti, AT & T Business Markets Group

PRACTICAL ISSUES IN COMPUTER INTERVIEWING

Maintaining Quality in Large Computer-Interactive Interviewing Projects
Nancy L. Messinger, Burke Marketing Research

20,000 Computer Interviews a Year: Is This Insanity?
Diane L. Pyle, Hallmark Cards, Inc.

CONJOINT ANALYSIS AS A TOOL FOR COMPETITIVE PRICING

Optimal Pricing Strategies Through Conjoint Analysis
Mary Jane Tyner, Levi Strauss & Co.
Jonathan Weiner, MACRO

Simulated Purchase "Chip" Testing Versus Trade-Off (Conjoint) Analysis — Coca-Cola's Experience
N. Carroll Mohn, The Coca Cola Company

Using Ci2 and ACA to Obtain Complex Pricing Information
Greg S. Gum, U S WEST

EMERGING TECHNOLOGY

Handwritten Data Entry Into A Computer
Ralph Sklarew and Tim Bucholz, Linus Technologies, Inc.

The Missing Link
Peter A. Schleim, Intercept Network Corp.

COMPUTER INTERVIEWING APPLICATIONS

Large-Scale Conjoint Data Collection at the Chicago Auto Show
Linda S. Middleton, Chicago Tribune

Computer Interviewing Applications in the Navy
Emanuel P. Somer and Dianne J. Murphy, Navy Personnel Research & Development Center

Computer Interviewing in Europe: What U.S. Researchers Need to Know
Martin Stefflre, PA, London, England

A Taste of Japan
Ray Poynter, Sandpiper International Limited

PERCEPTUAL MAPPING

Using Perceptual Mapping for Market-Entry Decisions
Richard H. Siemer, Dow Chemical Company

Evaluating Distribution Channels with Perceptual Mapping
Bob Block, John Morton Company

Repositioning A Service: Advanced Marketing Research and Associated Management Issues
David L. Masterson, Masterson & Associates

Perceptual Mapping and Cluster Analysis: Some Problems and Solutions
Charles I. Stannard, D'Arcy Masius Benton & Bowles

1989, continued

A Correspondence Analysis Approach to Perceptual Maps and Ideal Points
Vincent Shahim, Markinor House
Michael Greenacre, University of South Africa

Discriminant Versus Factor-Based Perceptual Maps: Practical Considerations
Thomas L. Pilon, IntelliQuest, Inc.

CONJOINT ANALYSIS APPLICATIONS

The Manager Versus the Customer: A Comparison of Values
Richard B. Ross and Larry G. Gulledge, Elrick and Lavidge, Inc.

Conjoint Analysis Across the Business System
Neil Allison, McKinsey & Company, Inc.

Using Conjoint Strategically to Enhance Business Engineering
Roger Moore, The Boston Consulting Group

Gaining A Competitive Advantage By Combining Perceptual Mapping and Conjoint Analysis
Harla L. Hutchinson, John Morton Company

Bias in the First Choice Rule for Predicting Share
Terry Elrod and S. Krishna Kumar, Vanderbilt University

Assessing the Validity of Conjoint Analysis
Richard M. Johnson, Sawtooth Software

CLUSTER ANALYSIS

New Findings With Old Data Using Cluster Analysis
Tara Thomas, Blue Cross and Blue Shield of Iowa

A Comparison of Clustering Methods
William D. Neal, SDR, Inc.

Reliability, Discrimination, and Common Sense in Cluster Analysis
Natalie M. Guerlain, POPULUS, Inc.

1988

CONJOINT ANALYSIS

Conjoint Analysis: Its Reliability, Validity and Usefulness
Dick R. Wittink, Cornell University

Conjoint Predictions 15 Years Later
John A. Fiedler, POPULUS, Inc.

Reliability Issues in Attribute Selection
Douglas L. MacLachlan, University of Washington;
Michael Mulhern, Mulhern & Associates
Allan Shocker, University of Washington

Comparison of Conjoint Method
Manoj Agarwal, State University of New York

A Comparison of Rating and Choice Responses in Conjoint Tasks
Jordon J. Louviere, University of Alberta

Comparison of Conjoint Choice Simulators
Carl Finkbeiner, National Analysts, Division of Booz•Allen & Hamilton

Statistical Software for Conjoint Analysis
Scott M. Smith, Brigham Young University

Software for Full-Profile Conjoint Analysis
Steve Herman, Bretton-Clark

Solving Practical Problems

Conjoint Analysis By Telephone
Brent Stahl, Minnesota Opinion Research

Conjoint Analysis by Mail
Dan Cerro, Bain & Co.

Complex Computer Interviews
Robert Zimmermann, Maritz Marketing Research

PERCEPTUAL MAPPING

Overview of Perceptual Mapping
William D. Neal, SDR, Inc.

Comparing Mapping and Conjoint Analysis: The Political Landscape
Joel Huber, Duke University
John A. Fiedler, POPULUS, Inc.

The Effects of Familiarity: Who Should Rate What?
William G. McLauchlan, McLauchlan & Associates

Major Research Companies and Perceptual Mapping

A Comparison of Techniques for Perceptual Mapping
Robert W. Ceurvorst, Market Facts

Preparing Data for Mapping
Roger Buldain, Burke Marketing Research

Decision Criteria for Selecting Mapping Techniques
Gordon A. Wyner, M/A/R/C

Perceptual Mapping: A Comparison of APM with Paper and Pencil Data
Herb Hupfer, Elrick & Lavidge

SPECIAL SECTION

Emerging Technology and Its Impact on Data Collection
Vincent Vaccarelli, Xerox Corporation

Increasing the Use of Market Research and the Status of Market Researchers
Marc Prensky, MicroMentor, Inc.

COMPUTER INTERVIEWING

Computer Interviewing: Current Practices and Cautions
Richard Miller, Consumer Pulse, Inc.

Door-to-Door Interviewing with Laptop Computers A Year Later
Joel Gottfried, National Analysts, Division of Booz•Allen & Hamilton

PC-Based Research: Europe Versus the U.S.
Dirk Huisman, SKIM Market & Policy Research

Special Opportunities & Problems

Use of Computer Interactive Interviewing at Trade Shows
Jacqueline Labatt-Simon, Cahners Exposition Group

Computer Interviewing With the Mobile Van
Carlos Barroso, Procter and Gamble Co.

Developing Complex Computerized Questionnaires
Ann Weaver, American Medical Association

Unattended Kiosk Interviewing
Glenn Okimoto, Hawaii Dept. of Transportation

Disks-by-Mail: A New Survey Modality
Marshall G. Greenberg, National Analysts, Division of Booz•Allen & Hamilton
Lesley Bahner, POPULUS, Inc.
Richena Morrison, Morrison & Morrison
Brent Dahle, CSR Institute
Thomas Pilon, IntelliQuest, Inc.
Harris Goldstein, Trade-Off Research

Statistical Analysis and the Market Researcher
Tony Babinec, SPSS, Inc.

Survey Research Software: From Expert System Sampling through Computer Interviewing, Data Analysis and Presentation to Publication
Ed Carpenter, University of Arizona

1987

PERCEPTUAL MAPPING

Perceptual Mapping: Its Origins, Methods, and Prospects
Allan D. Shocker, U. of Washington

Adaptive Perceptual Mapping
Richard M. Johnson, Sawtooth Software

Analysis and Interpretation of Results
Paul N, Ries, Procter and Gamble
Paul Hase, Hase/Schannen Research

How to Sell Perceptual Mapping
Michael Baumgardner, Burke Marketing Services
Edward (Ted) Evans, Ortho Consumer Products Division, Chevron Chemical Company

How to Design a Study
Betty A. Sproule, Hewlett-Packard
William G. McLauchlan, Burke Marketing Research

Presenting Results
Bruce J. Morrison, General Electric
John A. Fiedler, POPULUS, Inc.

CONJOINT ANALYSIS

Conjoint Analysis: How We Got Here and Where We Are
Joel Huber, Duke University

Adaptive Conjoint Analysis
Richard M. Johnson, Sawtooth Software

How to Sell Conjoint Analysis
Verne B. Churchill, Market Facts, Inc.
Vincent P. Vaccarelli, Xerox Corporation

How to Design a Study
James J. Tumbusch, Procter and Gamble
Richard D. Smallwood, Applied Decision Analysis

Analysis and Interpretation of Results
Marshall G. Greenberg, National Analysts, Division of Booz•Allen & Hamilton
Donald Marshall, Smith Kline & French Laboratories

Presenting Results
Peter B. Bogda, M/A/R/C Inc.
Marc R. Prensky, The Boston Consulting Group

COMPUTER INTERVIEWING

Historical Perspectives and the Future of Computer Interviewing
Lawrence Dandurand, University of Nevada

Doing Traditional Research By Computer: What Has Been Learned So Far

Long Self-Administered Interviews
Lesley A. Bahner, POPULUS, Inc.

Political Polling
Brent Stahl, MORI, Inc.

Consumer Taste Tests
David Griscavage, PepsiCo, Inc.

Telephone Interviewing
Richard Miller, Consumer Pulse, Inc.

Children's Research
Ira Goodman, J.M.R. Marketing Services

New Horizons — Taking Computers Where They Haven't been Before

Computer Surveys By Mail
Harris Goldstein, Trade-Off Marketing Services,

Complex Interviews with Laptops
Joel Gottfried, National Analysts, Division of Booz•Allen & Hamilton

Computers and Hard-To-Interview Respondents
William Tooley, IMR Systems, Ltd.

Who Should Do What? — Field vs. Supplier vs. Client

Audrey Bowen, General Foods
David Griscavage, PepsiCo, Inc.
David Santee, Hallmark Cards, Inc.
Peter Honig, Peter Honig Associates
Elizabeth Bradley, National Analysts, Division of Booz•Allen & Hamilton
Richard Miller, Consumer Pulse, Inc.
Bernadette Schleis, Interactive Network, Div. of Bernadette Schleis & Associates

Ci2 in the University

Arthur Saltzman, Cal State University
Lawrence Dandurand, University of Nevada