Proceedings of the Sawtooth Software Conference

1991

Foreword

It is our pleasure to present these Proceedings of the fourth Sawtooth Software Conference, held in Sun Valley, Idaho, in January, 1991.

These papers address PC interviewing and analysis, and include such diverse topics as: non-response bias in telephone interviewing; the availability of PC technology in field interviewing locations; measuring brand equity with conjoint analysis; and data input devices. Several papers focus on the validity and accuracy of our adaptive conjoint analysis method and some compare that technique with the full profile method.

The 1991 conference was our first that included prepared discussants. The discussants presented complementary and contrasting views, which encouraged audience participation in the "question and answer" periods that followed each presentation. We thank the many discussants who chose to publish their comments in this Proceedings.

We wish it had been possible to include the lively discussions that were launched by the speakers and discussants. These discussions were full of useful comments, challenges, and suggestions. Practical advice was shared, and ideas for future research were proposed. We hope the readers who did not attend — and those who did — will join us and their colleagues at our next conference, in July, 1992.

The papers presented are in the words of the authors; only light copy editing was performed. We thank all of the authors for their efforts and for the contribution they have made to both theory and practice.

Margo Metegrano
Editor
April, 1991

Table of Contents

INTERVIEWER ERROR: A COMPARISON OF METHODS

Thomas L. Van Valey
Sue R. Crull
Kercher Center for Social Research
Western Michigan University

INTRODUCTION

Computer-assisted telephone interviewing (CATI) procedures have become very sophisticated in the past decade. A number of studies comparing CATI and paper methods have been conducted (Groves and Mathiowetz, 1984; House, 1984; Coulter, 1985; Harlow, Rosenthal and Seigler, 1985; Catlin and Ingram, 1988; and Catlin, Ingram, and Hunter; 1988). In addition, Groves and Nicholls have provided several reviews of the effects of CATI (Groves and Nicholls, 1986; Nicholls and Groves, 1986a; Nicholls and Groves, 1986b). Common elements among the studies are comparisons of response rates, time and cost factors, and data quality. Our study will replicate comparison elements of most of the previous studies, but will look at data quality with a new emphasis on interviewer error, using a detailed assessment procedure based on interviewer monitoring.

Data quality is directly influenced by interviewer performance. In this study, three types of interviewer error are evaluated: skip-pattern error, interpretation error and response selection error. Skip-pattern error involves moving from one question or section of the survey instrument to another incorrectly, given the reply to a screening question or set of questions. In computer interviewing, the pattern is incorporated into the computer program, and should eliminate skip errors entirely (barring errors in program logic that are not corrected). This is well documented in previous research. In paper methods, such movements are entirely the responsibility of the interviewer (although word and graphic aids can also be incorporated into the paper instruments).

Interpretation error involves the incorrect selection of a response due to the misunderstanding of the respondent's verbal answer to a given question. Previous research does not indicate the effect of the computer method on interpretation error. Listening skills are especially important for items that require judgments about degrees of response, such as the ubiquitous Likert and similar type questions calling for satisfaction or agreement on the part of the respondent.

Response selection error, which has not been assessed in any CATI comparisons, involves pressing the wrong key for the computer method or indicating the wrong number with the pencil for the paper questionnaire. For the paper method, it also includes errors that are introduced during data entry (which, of course, are verifiable and can be eliminated). Although computer programs can reduce keystroke errors with consistency checks, there is no way to eliminate some within-range keystroke errors. The importance of keying errors was presented in the discussion of Catlin, Ingram, and Hunter (1988) by Tinari (1988:301):

> ...many sources of error are common to both modes, for example, misunderstanding questions and misrecording answers. Just because both sources of error exist in both modes, this does not mean they are equal. I believe that keying errors are more likely on CATI than writing errors using pencil and paper. Granted, keying errors may be caught by the built-in checks on CATI, but this is not always the case and not all types of errors can be detected if they are within range.

The lack of assessment of keystroke error is also mentioned by Groves and Nicholls (1986:127) in their discussion of measurement error:

Since studies comparing recording errors in CATI with those on paper forms have not been undertaken, it remains unknown whether on balance CATI increases or decreases their frequency.

PROCEDURES

We have been aware of keystroke error at the Kercher Center since we computerized our telephone interviewing more than five years ago. We decided to study the magnitude of the error and developed a small, controlled study to measure keystroke error. The project for which the data were collected was the fourth annual random-digit-dial telephone survey for a local municipality, involving a sample of over 400 residents. The focus of the survey was the residents' attitudes and opinions about city services and community issues.

A total of 20 interviewers were used on the project, 10 experienced with paper methods, 10 experienced with computer methods. Interviewers did not cross over in method, therefore removing the possibility of any bias for one method or another. All interviewers received a one-hour training session specific to the survey instrument. Training with respect to method was not necessary as all interviewers were experienced in their respective methods. Interviewers were asked to complete ten interviews, and were paid on a piece rate basis. The balance of the interviews on the project were carried out using computer methods, which is the more common approach at the Kercher Center.

The instrument contained a total of 65 items, with a potential of 11 open-ends (although in actuality no one selected more than 6). There were 5 skip patterns: 2 skipping a single item, 1 skipping three items, 1 skipping five items (and crossing a page), and 1 skipping six items (and crossing a page). The two forms of the instrument were identical in structure and wording, although appropriate directions (both word and graphic) were included for the interview schedules using traditional paper methods. The computer method, of course, required no special directions regarding skip patterns for the interviewers.

To determine whether or not the various types of error are taking place, it is of course necessary to know what the respondent is saying in response to the questions posed by the interviewer. Since it would have been necessary to notify both parties, and since that could have affected the responses, we decided to use monitors instead of tape recording the calls. Each interviewer's work was

independently monitored by two experienced graduate students through the use of special telephone equipment. In addition, the monitors recorded the number and nature of all calls placed, the start and finish of every interview carried out, the interviewer starting and ending times, and any break times.

Data collection began on October 13 and was completed on the 29th (an unusually long period of time was required to schedule all the monitoring). Calling was done from 5PM to 9:30PM on weekdays, and 10AM to 9:30PM on weekends. No screening for eligible households was carried out prior to the initiation of the project, and no call-backs were made. The interviewers were simply given randomly selected call sheets at the beginning of each session. For those interviewers using paper methods, the monitors used paper methods; for the computer methods, both the interviewers and the monitors used stand-alone computers with Ci2 software (Ci2 System for Computer Interviewing). In neither instance were automated calling procedures included. Interviewers were informed that some of their work would probably be monitored. However, in an effort to distract attention from the fact that unusual procedures were being used in the project, the interviewers were told that we were investigating: 1) the utility of piece rate payment and, 2) the quality of interviewer training (each interviewer was given pre- and post-interviewing questionnaires dealing with these subjects).

To measure each of the three types of interviewer error, we compared the individual responses of the interviewers with those of the two monitors who listened to the interviewing. In an effort to be conservative, we decided that a "known" error took place only when the two monitors agreed and also disagreed with the interviewer. Then, the type of error was determined, by the nature of the responses selected. For example, if the interviewer recorded a "2" (which might refer to the category "somewhat satisfied"), and the monitors both recorded a "3" (which might refer to the category "neutral"), this would be classified as a response selection error on the part of the interviewer. In the same situation, if the two monitors both recorded a "7" or an "8" (referring to monitor categories regarding answers that were not specific enough), this would be an interpretation error. Skip errors could only occur when there was no disagreement on the screening question of a skip pattern, yet disagreement on the items within the skip. We also found it necessary to designate a residual category of error, "unknown error," for those situations where the interviewer and the two monitors all recorded different responses. For open-ends, the texts were checked for spelling errors and total word counts.

RESULTS

Although the emphasis of this project is upon interviewer error, information regarding results and cost are first presented to establish a context for the analysis of error. Table 1 presents information for each interviewer regarding the total number of calls made, the total work time required to complete the requested ten interviews, the average interview time (from "Hello" to hang up), and the proportion of work time spent actually interviewing.

Table 1: Comparison of Time Spent, Paper and Computer

Interviewer	Completed Interviews	Total Time	Average Interview	Percent of Time Spent Interviewing	Total Calls
Paper					
1	10	2.1 hr	9.1 min	71.1%	29
2	10	2.7	11.7	73.6	38
3	10	3.2	9.8	50.8	60
4	10	3.2	12.8	66.7	55
5	10	3.5	11.0	52.4	32
6	8	2.7	12.0	60.0	61
7	10	2.9	10.2	58.6	75
8	10	2.8	10.7	62.9	49
9	10	3.0	10.8	59.3	84
10	10	3.1	12.7	68.3	35
	98	29.2 hr	11.1 min	62.4%	518
Data entry		4.5			
		33.7 hr			
Computer					
1	10	2.1 hr	9.0 min	71.4%	34
2	10	3.0	10.8	59.3	63
3	10	2.5	11.0	73.3	33
4	10	2.3	10.1	72.1	36
5	10	3.4	11.4	55.3	34
6	10	2.4	10.0	70.9	36
7	10	2.7	11.0	67.5	37
8	10	1.9	8.6	74.8	39
9	10	2.5	10.8	72.0	46
10	10	2.9	10.2	58.3	67
	10	25.7 hr	10.4 min	67.5%	425

Although there is considerable variation among both the paper and computer interviewers, it is clear from the totals reported in Table 1 that in all four aspects the computer method was more efficient than the paper method. The paper interviewers made 93 more phone calls, required 1.5 hours more overall time to achieve their goal of ten completed interviews, and took longer carrying out the interviews themselves (11.1 minutes, as compared to 10.4 minutes by computer). Moreover, they spent a somewhat smaller proportion of their time actually interviewing (62.4%, as compared with 67.5%). In contrast, the open-ended responses for the paper method were somewhat longer than those for the computer method (69.6 words as compared with 58.0 words). It should be noted, also, that this does not take into account either the additional 4.5 hours that must be added to the paper method for data entry, nor the fact that one of the paper interviewers did not complete the requested ten interviews.

At the very least, these data indicate that the computer method appears to be measurably faster than the identical paper method (6.3% in average length of interview, 12.0% in overall time). In addition, the computer method displays a higher potential productivity, given the 8.2 percent more time spent on actual interviewing.

The important question to ask at this juncture is whether the comparison is fair. It is certainly possible that the paper interviewers fell victim to an unfortunate sampling distribution, or to the vagaries of inefficient scheduling. Table 2, which presents the various rates of response, sheds some light on the issue.

Table 2: Comparison of Results and Cost, Paper and Computer

Interviewer	n	Percent Answered	Percent Eligible of Answered	Percent Completed of Eligible	Cost @$8/hour
Paper					
1	10	69.0%	75.0%	66.7%	$16.80
2	10	52.6	65.0	76.9	21.60
3	10	51.7	41.9	76.9	25.60
4	10	30.9	100.0	58.8	25.60
5	10	46.9	73.3	90.9	28.00
6	8	34.4	71.4	53.3	21.60
7	10	46.7	62.9	45.4	23.20
8	10	36.7	77.8	71.4	22.40
9	10	32.1	66.7	55.6	24.00
10	10	57.1	75.0	66.7	24.80
	98	45.8%	70.9%	66.3%	$233.60
Data entry					36.00
					$269.60
Computer					
1	10	55.9%	73.7%	71.4%	$16.80
2	10	44.4	60.7	58.8	24.00
3	10	57.6	63.2	83.3	20.00
4	10	55.6	65.0	76.9	18.40
5	10	41.2	71.4	100.0	27.20
6	10	55.6	60.0	83.3	19.20
7	10	46.0	64.7	90.9	21.60
8	10	56.4	77.3	58.8	15.20
9	10	39.1	83.3	66.7	20.00
10	10	32.8	68.2	66.7	23.20
	100	48.5%	68.8%	75.7%	$205.60

Once again, it is apparent that there is considerable variation among the interviewers. Nevertheless, the totals are informative. With respect to the percent answered (which excluded calls to answering machines, no answers, non-working numbers, and so on) and percent eligible (which excluded businesses, children and non-residents), the difference between the two methods is negligible. The paper interviewers achieved a slightly smaller proportion of calls that were answered by a person (2.7%), and the computer interviewers reached slightly fewer respondents who were eligible to be interviewed (2.1%).

The more telling difference between the two methods is the percent completed. The computer interviewers managed to complete 9.4 percent more of the eligible phone calls than the paper interviewers. Because of the method of calculating refusal rates (100 - Completion Rate), those are not shown. In conjunction with shorter interviews, less overall time spent working, and more time spent at interviewing, it is little surprise that the per-case cost figures are quite different ($2.75/case for paper methods, in contrast to $2.06/case for computer — a reduction of more than 25%). When calculated over this small project, the result is a modest differential in cost of slightly more than $275. However, in a survey research establishment that does 15,000 interviews a year, the differential would be in the range of $10,000.

The final issue at hand is that of interviewer error. While the data have demonstrated that computer interviewing is clearly faster and more efficient than traditional paper methods, is it accomplished only by paying a price in reduced data quality? Table 3 presents the results for each of the three major types of interviewer error.

Looking first at response selection error, it is clear that the paper method produced fewer errors than the computer method (35 as compared with 68). Even when the errors produced through data entry are added in, the total (53) is still smaller than those produced by the computer interviewers. The errors in the open-ended items were virtually the same (20 for paper, 19 for computer). The amount of interpretation error is slightly larger for the paper method (105) than for the computer (94). For skip-pattern errors, the paper method does show two errors in contrast to the zero errors that would be expected for the computer method.

Although not shown in the table, the unknown errors displayed a different pattern in that the paper method produced more than the computer method (79 as compared to 64). Because the resolution of these errors could substantially increase either the response selection and/or the interpretation errors, further analysis will be carried out as to their impact.

Table 3: Comparison of Errors, Paper and Computer

Interviewer	n	Selection Errors		Skip Errors		Interpretation Errors	
Paper							
1	10	0		0		11	
2	10	0		0		4	
3	10	4		0		50	
4	10	6		0		9	
5	10	4		0		1	
6	8	6		2		8	
7	10	1		0		4	
8	10	1		0		1	
9	10	6		0		3	
10	10	7		0		14	
	98	35	(.55%)*	2	(.03%)*	105	(1.65%)*
Data entry errors		18					
		53	(.83%)*				
Computer							
1	10	21		0		19	
2	10	7		0		0	
3	10	6		0		9	
4	10	5		0		2	
5	10	7		0		2	
6	10	4		0		25	
7	10	2		0		6	
8	10	11		0		28	
9	10	2		0		1	
10	10	3		0		2	
	100	68	(1.05%)*	0	(.00%)*	94	(1.45%)*

* The numbers in parentheses are error rates, based on the total number of interviewer-controlled keystrokes in each case.

When interviewer errors are divided by the total number of keystrokes over which the interviewers had control (65 times the number of completed interviews), it is clear that the overall rates of error are all quite small. For response selection error, it is no larger than .83 percent for the paper method (and that is if the data entry errors are left uncorrected), as compared to 1.05 percent for the computer method. The rate of skip-pattern error is negligible for the paper method and non-existent

for the computer. Even the rates of interpretation error, which were the largest detected, were in the range of 1.5 percent, for both methods.

DISCUSSION

In this study of 198 monitored telephone interviews (98 paper and pencil, 100 computer-assisted) involving 20 experienced interviewers, comparisons were made of response rates, time and cost factors, and data quality. The paper method displayed more total calls, more total work time, and a greater average interview time than the computer method. In addition, the percent of time spent interviewing was less for the paper method than the computer method. In contrast, the open-end responses for the paper method were somewhat longer than those for the computer method.

Compared to other studies, our findings were anomalous with regard to average length of interview. House (1984) found no difference in interview length between the two methods. Four other studies found the computer interviews to be longer than those using the paper method (Nicholls and Grove, 1986b). Catlin, Ingram and Hunter (1988) also found longer interview times for the computer method, but felt that it might be due to computer downtime and differences in the structure of the instruments. With respect to productivity, our results are similar to those reported in the literature (Nicholls and Grove, 1986b). As to open-ends, our findings support Nicholls and Grove (1988) who speculated that the computer method may result in less complete entries than the paper method. Catlin and Ingram (1988) found no significant differences in the word count of open-ends for the two methods.

Regarding response rates, we found little difference between paper and computer methods with respect to the percent of calls answered or eligible for interviewing. However, the computer method resulted in a larger proportion of completed interviews. The bottom line is also clear; computer interviews proved to be substantially less costly than paper interviews.

Comparing our results with respect to response rates to those reported in the literature is somewhat problematic because there tends to be considerable variability in the calculation of completion rates. Generally, previous studies found no differences in refusal rates produced by the two methods. However, we found that the paper method had higher refusal rates (lower completion rates). As to costs, the two studies that reported cost data calculated them differently and also reported contradictory results. Catlin and Ingram (1986) found the cost per interview to be slightly higher for the computer method. However, Ferrari (1986), as cited in Nicholls and Grove (1986b), found — as we did — that the cost per completed interview was greater for the paper method than for the computer.

Turning to the issue of data quality, response selection error was greater for the computer method than the paper method. As was expected, there was no skip-pattern error for the computer method, and a negligible amount for the paper method. Interpretation errors were higher than expected, but were similar for both the paper and computer methods. Since there is so little in the literature regarding this issue, it is pointless to compare our results. Nevertheless, it does appear that typing

errors (which are analogous to our response selection errors for the computer method) are likely to be in the range of 2-5 percent. We found much lower rates. To what extent these results generalize to data entry during an interview is not at all clear.

A more important issue has to do with strategies for reducing the amount of error that invariably does creep into data. Perhaps, as is often heard, careful training and close supervision are the key. Certainly these are excellent ideas and cannot hurt. However, from our data it does appear that all interviewers make some errors (although some are markedly better than others). It may simply be that some level of error will have to be tolerated.

Another strategy that would appear to have promise is a program of continued monitoring of interviews where both sides of the interview are noted (as distinguished from supervision where only the interviewer's words are noted). In our data for example, by eliminating the three interviewers with double-digit error rates in each of the methods, the number of response selection errors could have been reduced by 31.4 percent (paper) and 52.9 percent (computer). For the number of interpretation errors, the results are even more dramatic — 82.4 percent lower for the paper method and 76.6 percent lower for the computer method! While this might introduce certain tradeoffs (such as speed of interviewing — thus alterations in productivity and cost — or morale problems among interviewers), it could also result in significant reductions in the amount of error that exists in a telephone survey.

REFERENCES

Catlin, Gary, and Ingram, Susan, "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing," in *Telephone Survey Methodology*, Robert G. Groves, *et al.*, (editors), New York: John Wiley and Sons, 1988, pp. 437-450.

Catlin, Gary, Ingram, Susan, and Hunter, Lecily, "The Effect of CATI on Data Quality: A Comparison of CATI and Paper Methods," *Proceedings of the Fourth Annual Research Conference of the U.S. Bureau of the Census*, U.S. Bureau of the Census, 1988, pp. 291-299.

Coulter, R., "A Comparison of CATI and non-CATI on a Nebraska Hog Survey," Staff Report No. 85, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC, April 1985.

Groves, R.M., *Survey Costs and Survey Errors*, John Wiley & Sons, New York, 1989.

Groves, R.M., and Mathiowetz, N.A., "Computer Assisted Telephone Interviewing: Effects on Interviewers and Respondents," *Public Opinion Quarterly*, Vol. 48, No. 1B, 1984, pp. 356-369.

Groves, R.M., and Nicholls, W.L., II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," *Journal of Official Statistics*, Vol. 2, No. 2, 1986, pp. 117-134.

Harlow, B.L., Rosenthal, J.F., and Ziegler, R.G., "A Comparison of Computer-Assisted and Hard Copy Telephone Interviewing," *American Journal of Epidemiology*, Vol. 122, 1985, pp. 335-340.

House, C.C., "Computer-Assisted Telephone Interviewing on Cattle Multiple Frame Survey," Staff Report No. 82, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC, 1984.

Nicholls, W.L., II, and Groves, R.M., "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," *Journal of Official Statistics*, Vol. 2, No. 2, 1986, pp. 93-115.

----------, "The Status of Computer-Assisted Telephone Interviewing," *Proceedings of the 18th Symposium on the Interface, Computer Science and Statistics*, 1986, pp. 130-137.

Tinari, Robert, "Discussion of 'The Effect of CATI on Data Quality: A Comparison of CATI and Paper Methods'", (By Catlin, Ingram, and Hunter), *Proceedings of the Fourth Annual Research Conference of the U.S. Bureau of the Census*, U.S. Bureau of the Census, 1988, pp. 300-304.

Comment on Van Valey and Crull

Marcus Schmidt
Southern Denmark Business School

The Van Valey and Crull paper discusses some important sources of error in telephone interviewing:

1. Skip-pattern error (moving from one question or section of the survey instrument to another incorrectly),

2. Interpretation error (incorrect selection of a response due to the misunderstanding of the respondent's verbal answer), and

3. Response selection error (pressing the wrong within-range key for the computer method or indicating the wrong number with the pencil for the paper questionnaire).

The authors report some main findings resulting from an experimental design which compares the impact of the above error types.

Ten interviewers (approximately 100 interviews) were conducting the telephone interviews using traditional paper-and-pencil methods, while another 10 interviewers (100 interviews) were using computer interviewing software (Ci2).

One of the above error types — skip-pattern error — is eliminated when using computer interviewing because the pattern is incorporated into the program. Errors in program logic might still exist, but these are not interviewer errors.

DISCUSSION OF THE RESULTS

Table 1 summarizes some of the findings of Van Valey and Crull.

Table 1: Comparison of Time spent, Paper and Computer

Method	Total Time
Paper & Pencil (n=98)	29.2 hr + 4.5 hr (data entry) 33.7 hr
Computer (n=100)	25.7 hr

Of course, when using computer interviewing, it is not necessary to spend time on data entry.

However, I do not think that the above comparison is quite fair with regard to paper-and-pencil methods since it only covers field interviewing. While the computer approach indeed seems to outperform the paper-and-pencil technique during the field interviewing phase, the opposite might well be true during the questionnaire construction phase.

It seems to be much easier to construct a paper-and-pencil questionnaire than a computerized questionnaire.

This is due to the fact that one does not need to care about program logic when using paper and pencil. Often, it can be a time-consuming task to develop the logic for a computer interview. And, one has to use many files simultaneously: frames, logic, numstart, and so on.

Moreover, some time is spent aggregating interviews from field diskettes and converting the files for data processing.

On the other hand, some other non-field activities such as

- copying the questionnaire
- data checking, and
- recoding

seem to contribute to the advantages of computer interviewing.

So, if one wants to perform a comprehensive comparison of the two interviewing methods, the whole research process will have to be taken into account.

This somewhat broader approach is also supposed to affect and change cost considerations (Table 2 in Van Valey and Crull).

TELEPHONE VS. FACE-TO-FACE VS. MAIL

The paper of Van Valey and Crull deals only with telephone interviewing. However, it might be interesting to evaluate the time- and cost-effectiveness of paper-and-pencil methods vs. computer methods using face-to-face interviews and mail interviews, respectively. See Table 2.

Table 2: Computer interviewing approaches

	Who fills in the questionnaire ?	How does the interview take place ?
Telephone	Interviewer	CATI at Agency
Face-to-Face	Interviewer or respondent	Interviewer with PC/Laptop
Mail	Respondent - Self-administered Questionnaire	Agency sends a Diskette with Q. to respondents

Whereas Van Valey and Crull have investigated the first of the three above approaches, the latter two have not been investigated in systematic ways (at least as far as this researcher is aware).

In each of the three above approaches the respondent is reacting; he or she is being contacted by an interviewer. Nowhere is the respondent acting (for example, by contacting the interviewer). Perhaps the respondent's answers would be different if he or she were acting.

In fact, it is possible to use a fourth way of gathering data: Voice response!

A few comments on how voice response might work in a survey research context might be appropriate:

WILL VOICE RESPONSE (VR) MAKE COMPUTERIZED TELEPHONE INTERVIEWING TECHNIQUES LIKE CATI OBSOLETE ?

This is how a VR Interview Application might work:

- A panel consisting of approx. 1500 respondents is established, using traditional recruiting techniques. The panel, of course, has to be representative with regard to the purpose of the recruiting criteria ("Voter," "Head of Household," "Export Manager," or the like).

- Each respondent must have easy access to a push-button telephone (either at home or at work). In Denmark 90% of all households own at least one push-button telephone (according to KTAS — the Public Copenhagen Telephone Company).

- The respondents are "trained" to phone an 800-number once a day. When hearing a special tone, they press their individual passwords.

- Next, the respondent is asked several questions by a pre-recorded voice, such as:

Do you favor or oppose question xyz ?

Press "1" for favor
Press "2" for oppose
Press "3" for don't know ... and so on.

The respondent just presses the button that represents his or her opinion.

It is, of course, possible to "go backwards" in the questionnaire — as long as the line-connection exists — by using other buttons such as the "#" or "*" buttons, when one wants to correct errors. Skip pattern errors do not exist, **neither do response selection errors** — because it is the respondent, not the interviewer who presses the buttons.

While interpretation errors might still exist, their scope will be different and their shape will be reversed: The respondent might misunderstand the interviewer (the pre-recorded voice) but not vice versa! It might be difficult, if not impossible, to detect this kind of error, because trying to do so would involve phoning the respondents individually and asking them questions with regard to their understanding of the questions.

It should be stressed that most of the advantages of computerized interviews (such as randomization of questions and response alternatives) still exist when using VR. Moreover, it is possible to use voices (such as friendly young female voices) which optimize response-willingness (and result in a minimal number of interviews interrupted during the interviewing-process). It is also possible to let respondents choose among different interviewing voices.

Immediately after the interview is finished and the line is disconnected, data entry has taken place and the VR ASCII-data file contains a new observation. So, when the final interview has been conducted, an aggregated data file is present and the statistical computation can begin.

Even if the call is disconnected during the interview the data gathered so far will be present, which means that each numeric answer (keystroke) is transmitted separately in ASCII-character format to the VR-computer together with a lot of "noise" — line number and individual password. Indeed, this might seem to be somewhat clumsy, but so far it has proved technically difficult to overcome this clumsiness.

The capacity of a VR system is limited only by the available hardware and software budget.

VOICE RESPONSE promises to be a quick, cost effective, and reliable way of gathering data.

VR might even outperform CATI with regard to what might be called its "homegrounds," namely speed and cost — areas where up till now CATI has been believed to be a superior way of gathering market research information.

A VR panel is, of course, a panel and as such it inherits the advantages and disadvantages one has to consider when using traditional paper-and-pencil panels (mainly panel-effects). However, traditional panels are a well established phenomena of theory and practice in modern marketing research. Why should not the same apply to VR panels?

Besides, it is at least doubtful if it will ever be possible to develop an efficient VR system without using the panel recruitment technique. This is due to the fact that the panel approach is based on the respondent's willingness to participate.

Other possible VR systems might use one of two approaches:

1. Make the VR system call respondents according to a randomized technique. When using this approach one faces severe integrity problems. All respondents will be taken by surprise and be unprepared, and most of them will be low-involved when listening to a pre-recorded voice. So, refusal rates and the number of disconnected interviews could be high. Moreover, this approach offends the privacy of the respondents. In Denmark this approach might indeed be deemed illegal by the authorities (as might the traditional CATI approach) because — according to several systems — presently it does not differentiate between published and non-published phone numbers. So respondents with non-published numbers are also phoned by the system.

2. Another way to make a VR system work would be to make the respondents call the 800 VR system number in response to an announcement in the printed press or in the electronic media. But this approach will definitely not be cost effective — due to heavy advertising costs for announcements in a sample of media (one would have to cover and reach a broad audience).

The approach would also raise new problems with regard to the sample being representative. The sample might be biased, so one would have to perform weighting procedures or post-stratification procedures with regard to the sampling criteria.

In the global village, times are changing and so are the ways of gathering data!

PERSPECTIVES ON DATA QUALITY
THE CLIENT AND THE MARKETING RESEARCHER

Rebecca Elmore-Yalch
MACS, Inc.
Jane Glascock
Metro Transit, Seattle

It is frequently said that the design and execution of a marketing research study is as much an art as a science. To be sure, the "art" is there, as every study is different. On the other hand, the "science" is frequently lacking due in part to the fact that survey and marketing research is a methodology without a unifying theory. Robert Groves (Groves, 1987) argues that since survey research is not itself an academic discipline with a common language, a common set of principles for evaluating new ideas, and a well-organized professional reference group, there are no clear-cut guidelines as to what constitutes data quality.

This argument was borne out in a recent qualitative survey by MACS of some research industry professionals and individuals representing the "client" side of the equation. When asked to define "quality data," few individuals were able to come up with a single cohesive statement which would suggest the existence of an industry-wide definition of quality data. Moreover, few respondents were able to identify methods to insure that the end products were based on quality data.

This paper, therefore, first sets forth two alternative perspectives on data quality that of the client and that of the researcher. With these two perspectives in mind, the paper then identifies ways in which the client and the researcher attempt to manage the process by which data are collected so as to increase the overall quality. Then the paper discusses the potential contribution of computerization of data collection. Lastly, the realization of these improvements is evaluated using a longitudinal analysis of an annual, large sample survey that has evolved from a traditional pen and paper procedure with manual sample control, to a stand-alone computer-assisted interview with manual call management procedures, to a fully-integrated system where both questionnaire administration and sample/call management are automated.

STUDY BACKGROUND

To provide some background, the Transit Department of the Municipality of Metropolitan Seattle (Metro) has conducted a telephone survey of King County residents for the past several years. This annual Rider/Non-Rider survey has taken a variety of formats over the years, but serves the basic purpose of measuring customers' and non-customers' awareness of and attitudes toward Metro and its services. Further, the survey has been used as a forum for probing public sentiment about issues such as personal security and the building of the downtown Seattle bus tunnel. The annual survey has provided an important source of longitudinal data and has been used in developing short- and

long-range transit plans and in evaluating transit agency performance. For the past several years, major sections of the survey have remained the same and as such serve as an important tracking tool of public sentiment.

The Metro surveys offer a unique opportunity for examining the effects of new technologies on quality of data. In 1986, the survey was conducted using traditional pen-and-paper methods of questionnaire administration. Sampling control was non-automated and handled by the labor-intensive method of continually monitoring and updating call-record cards or sheets. In 1987, a stand-alone computer system was used for the first time to automate the administration of the questionnaire. From 1987 to 1988, the sample continued to be handled manually. In 1989 integrated computer-assisted telephone interviewing with sample and call management capabilities was introduced. Both questionnaire administration and sample / call management were automated and handled through the use of a local area network (LAN).

The Metro study allows a comparison of procedures for implementing as the survey have been the same since 1986. The questionnaire has been similar and has averaged approximately 20 minutes in length. Similar sampling procedures have been used by both research firms who conducted the surveys, with one exception. In 1986 and 1987, a maximum of four call-back attempts were made to each household; this number was increased in 1990 to five.

DATA QUALITY ISSUES

In any survey or marketing research project, clients are most often concerned with obtaining data that can be generalized legitimately to the population under study that is, the data are replicable and provide unbiased estimates of true population parameters. Clients are likely to talk about obtaining measurements which are ***reliable*** and ***valid***. Moreover, clients talk about using sample sizes large enough to obtain a level of ***precision*** and ***confidence*** upon which to base important business and marketing decisions. The processes by which the measurements become reliable and valid, as well as the management of the sample, are typically left to the research firm. It is the management of these processes which typically insures that the quality of data is indeed intact.

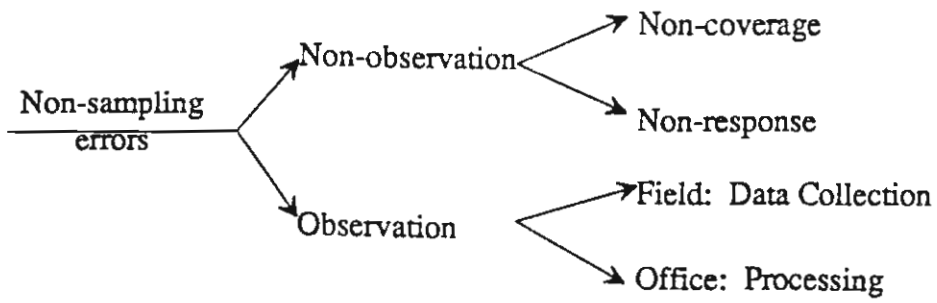
The biases which are of concern to clients such as Metro occur largely through the introduction of error. Two basic types of error occur in any marketing research study sampling and non-sampling error. Sampling error is random error which occurs in the selection of respondents. It is part of every study and cannot be minimized by improvements in the data collection procedures. Reducing sampling error typically means increasing the sample size, and therefore, is largely a budget issue. Researchers are mainly concerned with reducing non-sampling error, particularly non-sampling errors that are not random in nature. These systematic errors bias the value of sample statistics away from population parameters. Non-sampling errors occur because of errors in the research design, misinterpretation of respondent replies, errors in data analysis, errors in tabulation or coding, or errors in reporting the results. Frederick Mosteller (1968) states of non-sampling error:

The roster of possible troubles seems to have grown with increasing knowledge. By participating in the work of a specific field, one can, in a few years, work up considerable methodological expertise, much of which has not been and is not likely to be written down. To attempt to discuss every way a study can go wrong would be a hopeless venture.

It would be impossible to address every way to reduce non-sampling error in a single paper. Instead this paper focuses on the non-sampling error that occurs during data collection and the impact of new data collection technologies in reducing non-sampling error. Before addressing non-sampling error reduction, it is appropriate to review the sources of this error.

In data collection we are concerned with the two basic types of non-sampling error non-obser-vation error and observation or measurement error.

Overview of Non-sampling Errors



Source: Leslie Kish, Survey Sampling (New York: John Wiley & Sons, Inc., 1965) p. 519

Non-observation errors occur in sampling because part of the population of interest was not included or because some elements designated for inclusion in the sample did not respond. In addition to the sampling error inherent in any random sample of the population, the primary types of non-observation errors are (1) non-coverage error and (2) non-response error. *Non-coverage* errors denotes a failure to include some units, or entire sections, of the defined survey population in the actual operational sampling frame. For example, an entire zip code could be missed because the post office continually adds zip codes as the population in an area grows. Since telephone prefixes are closely tied to zip codes in King County, failure to include a new zip code could represent a serious problem. Another non-coverage error that organizations like Metro face is the inability to reach households without telephones. The fact that these households are frequently lower income households which include individuals who are more likely to ride the bus makes this a particular concern to Metro. *Over-coverage* error can also be a source of bias because of duplication in the list of sampling elements.

With today's advances in sampling techniques and random number generation, this is less of a problem than in previous years. However, over-coverage continues to occur to a limited extent due to the increased number of households with two phone lines. Although less of a problem for data quality, the problem of over-coverage is a public relations problem for the industry as repeated calls to the same household is an irritant most research firms would prefer to avoid.

Non-response error represents a failure to obtain information from some elements of the population that were selected as part of the sample. Over time, response rates have declined for household surveys (Steeh, 1981); a major factor being the increase in refusal rates. A recent *New York Times* article reported that refusal rates of all adults contacted had increased from 15% in 1982 to 34% in 1988. Metro suffers less from this problem with refusal rates between 16-23% each year since 1986.

Also of concern in terms of non-response error is the failure to reach a household despite repeated callbacks over the course of a study. The amount of non-response error resulting from this factor on the Metro surveys has varied significantly over the course of time during which Metro has conducted this survey. Reasons for this variance and its impact on data quality will be examined later in this paper.

Observation or measurement errors occur because inaccurate information is secured from the sample elements or because errors are introduced in the processing of the data or in reporting the findings. There are four potential sources of observation or measurement error of concern to Metro: (1) the questionnaire, (2) the mode of the interview, (3) the interviewer, and (4) the respondent. Observation errors therefore occur because of the influence of the interviewer, the weakness of the survey questions, failures of the respondent to give appropriate answers to the survey questions, and the effect of the mode of data collection on survey answers.

Measurement errors arising from the *questionnaire* itself come from a variety of sources: (1) the client may be confused about the purposes of the study and as a result may offer vague or even misleading direction to the researcher; (2) the researcher may be inept at resolving the client's confusion and helping the client detail the purposes of the study; and (3) the researcher may misunderstand the study purposes and pose the wrong questions. In addition, questionnaire errors arise in the development of the final instrument and its administration and result primarily from errors in question order, structure, and wording.

The use of telephone surveys as the *mode of the interview* introduces some unique types of error. Some of the errors that occur in telephone surveys result from the inability to reach special locations or respondents; the limits on the types of information that can be gathered (for example, only a limited amount of probing or open-ended questions is possible); difficulties in reaching the correct respondent in a household; difficulties in getting cooperation over the telephone; and an increased potential for interviewer bias.

At the *interviewer* level, Metro is concerned primarily with two sources of error: (1) interviewers failing to read the entire question, rewording or leaving out key information in a question, failing to ask follow-up questions where indicated, or failing to communicate effectively; and (2) interviewers inaccurately recording the respondent's information.

Respondent error occurs primarily as a result of social and psychological factors which influence a respondent's answer and the cognitive demands of the interview. Physical factors such as hearing difficulties can also affect respondent error. At the respondent level, Metro is concerned primarily with obtaining information which accurately represents the attitudes and behavior of the survey respondent.

While all error contributes to concerns about the quality of data, observation errors are frequently more troublesome to both the client and the researcher because it is difficult to assess the extent to which they are present. We can measure the extent of non-coverage and non-response errors, but with observation errors, we may not even be aware that a problem exists.

IMPROVING DATA QUALITY — THE CLIENT'S RESPONSE

While data quality is an important concept in any marketing research study, it is a particular issue to Metro because the comparability of data over time is a major study objective. There is as much interest in the changes from year to year as in the actual levels of response, because these changes indicate shifts in attitudes and behavior that are either the result of changes in Metro's operations or environmental factors that may require adjustments in Metro's operations. Biases are a concern especially if they vary from year to year. If so, changes in responses from year to year might be attributable to biases as well as actual changes in attitudes and behavior. Thus, Metro evaluates changes in the sampling procedures, the questionnaire, or the administration of the interview in terms of any biases that may be introduced.

Over the years during which the Rider/Non-Rider Survey has been conducted, Metro has attempted to help minimize the error inherent in any research study. Metro staff reviews, evaluates, and approves the research design and sampling plan to help assure that coverage is appropriate. Explicit study purposes and a draft questionnaire are also developed in-house, minimizing potential misunderstandings with the researcher about what should be included as the scope of the research. Metro staff also briefs the interviewing and research staff on questionnaire rationales, explains the meaning of possible responses, and defines transit-related terms to help educate interviewers about the study and ensure a fully knowledgeable interviewer pool. Staff monitors interviews as the study progresses, as well, and provides feedback to help ensure that questions are being asked consistently and responses are being understood.

Moreover, Metro's effort to minimize error has included selecting research firms to conduct its survey with either established manual procedures or a computer technology able to identify and thus

eliminate error at its source. Where error cannot be completely eliminated (for example, non-response error), Metro requires a measure of this error so that the overall quality of its data can be evaluated.

Finally, in addition to ongoing communications with the research firm, Metro's staff participates in the design of analyses, conducts reliability and validity assessments of the resulting data, and reviews and helps revise the final report. Metro strongly advocates this participatory client role as a way of improving overall data quality.

IMPROVING QUALITY — THE RESEARCHER'S RESPONSE

The responsibility for the quality of data ultimately rests with the research firm completing the study. When clients do not provide adequate information on research design and study specifications, it is the responsibility of the research firm to ask for that additional information. When clients make requests that will ultimately affect the overall quality of the data, it is the research firm's responsibility to inform the client of the ramifications of that request and to make appropriate recommendations. And because many clients do not know the intricacies of data collection and analysis, it is the research firm's responsibility to educate clients about the reasons research is done in a certain way.

There are two important ways for the research firm to minimize the error inherent in the data collection process: (1) thoroughly training interviewers, and (2) adopting a set of strict procedures and/or computer technologies that help control error. Much has been written about strategies for minimizing error through appropriate training procedures and instituting strict manual procedures for questionnaire administration and sample control. Much less has been written on the impact of new technologies on quality of data. Therefore, the remainder of this paper focuses on the effects of computer-assisted telephone interviewing on data quality.

It is frequently assumed that the use of computer-assisted telephone interviewing has a positive effect on the overall quality of data, but this assumption is based largely on qualitative factors. Little empirical evidence exists to substantiate claims that computer-assisted data collection has a positive effect on data quality. Moreover, it is not clear exactly how much improvement in data quality can be achieved by the use of stand-alone computer interviewing and what further improvements result when using a system with call management capabilities.

Much of the research published in applied marketing research and academic publications has evaluated the effectiveness of computer-assisted interviewing only in qualitative terms. For example, it has been suggested that the advent of computer-assisted interviewing has broadened the opportunities for quick and accurate survey results, reduced costs, and simplified the interviewing task for respondents and interviewers (Carpenter, 1989); has improved the efficiency and effectiveness of the method by which data are collected (Gates & Jarboo, 1987); and has positive effects on productivity and interviewing quality (Burg, 1986).

A more extensive review on the use of microcomputers in marketing research (Malhotra, Tashchian & Mahmoud, 1987) identifies three important areas in which errors are potentially reduced: (1) verification of data entry at the time of the interview and elimination of second-stage data entry; (2) automatic recording of output; and (3) reduction of time and error in data collection. It is suggested that electronic questionnaires allowing for automatic handling of the question order, the response format, and the skipping and branching instructions, leave the interviewer free to concentrate on asking questions, entering responses, and establishing a rapport with the respondent. Further, it is suggested that the automatic coding and data verification made possible by microcomputers significantly reduces the time, cost, and errors associated with manual data entry. This in turn can result in improved productivity, faster turn around, and higher quality data.

Some efforts have been made to quantify the effects of stand-alone computer-assisted interviewing. The effects of using a stand-alone computer assisted telephone interviewing system compared with traditional pen-and-paper methods on response distributions, interviewer behavior, and other non-sampling errors were studied (Groves and Mathiowetz, 1984). It was found that there were few differences in response distributions, some evidence of reduced interviewer variance among computer-assisted interview cases, but little difference in response rates between the two methods. Moreover, it was found that interviewers tended to have no clear preference between the two different methods, but favored the computer-assisted system for ease in following the questionnaire. Another study looked at nine surveys conducted over a six year period at the Survey Research Center at Michigan (Groves and Magilavy, 1986) and further examined interviewer effects on survey statistics. Small reductions in these effects were obtained when a stand-alone computer-assisted system was used. No research was found which directly compared the effects on data quality of stand-alone computer-assisted interviewing and computer-assisted interviewing with automated call management capabilities.

EFFECTS OF COMPUTERIZATION ON DATA QUALITY

Over the past two years during which time MACS has been involved in the Metro research, we have documented the effects of stand-alone computer-assisted interviewing and the introduction of a computer-assisted call management system on the overall quality of data notably, on non-observation error resulting from non-response.

As a first step in tracking this effect, response rates were calculated for each year the study was conducted. As shown in the following table, response rates increased steadily from 1986 to 1989. However, a substantial decrease occurred in 1990. An analysis of sources of non-response error revealed the areas in which the greatest improvements were achieved as well as possible explanations for the decrease in 1990.

EFFECT OF STAND-ALONE COMPUTER-ASSISTED INTERVIEWING AND COMPUTER-ASSISTED INTERVIEWING WITH CALL MANAGEMENT CAPABILITIES ON RESPONSE RATES
Metro RIDER / NON-RIDER SURVEY
1986 - 1990

YEAR	RESPONSE RATE
1986	69.7%
1987	77.0
1988	79.2
1989	81.5
1990	71.8

Response Rate = (Refusals + Rejects + Ineligibles + Terminations + Completed Interviews) / (Total Sample -Business / Non-Working Numbers)

Non-response error is primarily a function of three factors: (1) loss of interviews due to immediate refusals; (2) loss of interviews due to mid-terminates with failure to complete them at later, more convenient times; and (3) loss of interviews due to a failure to reach a household because of busy numbers, no answers, or answering machines. The following table tracks the level of error resulting from non-response from 1986 to 1990. During this period, stand-alone computer-assisted interviewing and ultimately computer-assisted call management capabilities were introduced.

EFFECT OF STAND-ALONE COMPUTER-ASSISTED INTERVIEWING AND COMPUTER-ASSISTED INTERVIEWING WITH CALL MANAGEMENT CAPABILITIES ON NON-RESPONSE ERROR — 1986-1990

	% OF SAMPLE DISPOSITIONED *					
YEAR	TOTAL NON-RESPONSE	REFUSAL	MID-TERMINATES	TOTAL NOT REACHED	NO ANSWER/ BUSY	CALL-BACKS
1986	51.7%	20.0%	0.9%	30.8%	26.1%	4.7%
1987	46.9	23.2	0.7	23.0	19.5	3.6
1988	38.8	17.0	1.0	20.8	19.1	1.7
1989	41.3	22.2	0.6	18.5	16.8	1.7
1990	44.5	15.9	0.6	28.0	25.6	2.4

** Based on usable sample. Does not include sample dispositioned as business / non-working numbers or as out of area.*

As shown in the table, total non-response error was highest in 1986 when the questionnaire was administered using traditional pen-and-paper methods and the sample was manually controlled. In 1986 total non-response error was 51.7%. The largest portion of this non-response error was a result of not reaching a large segment (30.8%) because they were not at home or their phones were busy. This number remained high despite attempts to schedule call-backs. Fifteen percent of the numbers dispositioned as households not reached had scheduled call-backs which were not achieved. This high degree of non-response could be attributed to the need to spend so much time monitoring the administration of the questionnaire due to the many complicated skip patterns, thereby limiting the amount of attention which could be devoted to maintaining sample control.

In 1987, stand-alone computer-assisted interviewing was introduced for the first time; sample continued to be handled manually. Non-response error dropped to 46.9% despite a increase in the number of immediate refusals. The most notable decrease was in the percent of sample that was not reached because of no answers or busy numbers (from 26.1% to 19.5% of sample). Although a smaller percent of the total sample was finally dispositioned as not reached despite a scheduled callback (3.6% of the total sample), the percent of scheduled call-backs reached in 1987 did not increase (15.4% of the numbers not reached).

In 1988, the error resulting from non-response continued to decrease to 38.8%. A large portion of the decrease is attributable to the decrease in the refusal rate, as opposed to better sample management. However, improvements were made in reaching households with scheduled call backs.

The data provide some support for the hypothesis that the use of stand-alone computer-assisted interviewing has a positive effect on data quality. Field supervisors may spend less time monitoring questionnaire administration and focus more on sample and call management with a resulting improvement in data quality.

In 1989, the use of computer-assisted telephone interviewing with computerized sample and call management capabilities was introduced. Total non-response error increased in 1989 to 41.3% due largely to an increase in refusal rates (22.2%). However, the increase in refusal rates was more than offset by a decrease in all other sources of non-response error.

Of particular interest is the decrease in percent of sample households not interviewed because the interview was terminated midway through the survey. Before the introduction of an automated call management system, the percent of households not included as completed interviews because interviews were terminated midway through the process ranged from 0.7 - 1.0%. In 1989, when an automated call management system was used, a 43% improvement was achieved in completing interviews which were terminated partway through. This is because it is relatively easy to retrieve respondent data and call records from the network to complete these interviews at another time. There were also decreases in the percent of sample not reached because no one was at home or the

number was busy. The automated sample management system did not improve the percent of scheduled call-backs reached.

In 1990, the level of non-response error appears to have stabilized, and in some instances it has increased. There was little change in the number of mid-terminated interviews that were later completed. Refusal rates dropped substantially.

On the other hand, the percent of sample that was not reached increased substantially. Additional analysis revealed that a large portion of that increase was due to the fact that respondents are becoming increasingly hard to reach over the phone, a challenge being faced by the entire research industry. Between 1989 and 1990 we began to track separately the impact of answering machines as well as not-at-homes and busy numbers on non-response error. An increase was noted in all areas except for reaching busy numbers.

IMPACT OF ANSWERING MACHINES ON % OF SAMPLE NOT REACHED BECAUSE OF NO ANSWER [1989 - 1990]

	% OF SAMPLE DISPOSITIONED *				
YEAR	TOTAL NOT REACHED	NOT-AT-HOME	BUSY	ANSWERING MACHINE	CALL-BACKS
1989	18.5%	10.8%	1.5%	4.6%	1.7%
1990	28.0	16.2	1.6	7.8	2.4

** Based on usable sample. Does not include sample dispositioned as business / non-working numbers or as out of area.*

Despite repeated call-backs (up to five per household), problems were noted both in reaching households who were not-at-home or where an answering machine was reached. The greatest impact occurred in households with answering machines. Seventy one percent more calls were dispositioned as not reached because of answering machines in 1990 than in 1989. On the other hand, there was only a 50% increase in the number of households where there was no answer. Despite the increase in the number of scheduled callbacks not reached as a percent of the total sample (2.4%), there was no change in the percent of households not reached because of scheduled callbacks between 1989 and 1990.

CONCLUSIONS

This paper has discussed several issues underlying the quality of data generated in a survey and identified sources of error that can affect data quality. Moreover, this paper represents a first attempt to track the effects of computer-assisted interviewing and call management techniques in reducing non-sampling error, most notably the non-sampling error which results from non-response. As a first attempt, this paper appears to raise more questions than it actually answers.

There are, of course, several limitations to the methodology employed to obtain these results. Ideally, an experiment is needed where interviews are randomly assigned to an interviewing and sampling methodology. It is a rare client, however, who will agree to that kind of experimentation and it is a rarer marketing research firm that can afford the luxury of testing at its own expense. Therefore, we must rely on tracking data to provide us with these indications of error reduction. The Metro survey represented an excellent opportunity to examine this issue, since the basic sampling plan, the questionnaire, and the questionnaire length all remained fairly constant for a number of years.

Further, it would be ideal if additional measures were available by which to judge the effectiveness of computer-assisted interviewing with call management capabilities. In particular, data on interviewer hours and interviewer productivity would provide additional insight into the issue. Also, data on the scheduling and management of call-backs would be useful to identify whether computerized call management introduced more varied calling hours as well as better follow-through on the actual number of call-backs made to any given number. Because MACS was not involved with the work early on, this information is not available.

Despite the limited number of data points, it appears that the introduction of stand-alone computer-assisted interviewing and ultimately a computer-assisted sample/call management system has had a positive effect on some areas of non-response error for the Rider/Non-Rider survey. Most notably there were improvements in the area of completing interviews with respondents who had terminated midway through. This represents an extremely cost-effective way for research suppliers to reduce non-response error, while eliminating a potential source of bias from the client's view. Moreover, some improvements were achieved in reaching scheduled call-backs. Although similar improvements can be attained using a manual call management procedure, there is no doubt that considerable cost savings can result by managing this process with a computer.

It is less clear that computer-assisted call management can achieve further improvements in reaching individuals who are rarely home or whose phones are often busy. Additional research would be needed to determine whether the increase in non-response in 1990 was due to an overall industry trend or simply due to the fact that it is unrealistic to expect any further improvements in this area.

In light of other industry trends affecting non-response error, the decreases achieved in the Metro study bode well for the future use of computer technology. Although questions could be raised

raised regarding the statistical significance of this decrease, it is the authors' belief that any improvements which have an impact on decreasing the error inherent in marketing research studies should be carefully evaluated. Continued improvements in this technology are likely to continue to benefit the research industry and its clients.

Before closing, some additional comments concerning the use of computer-assisted interviewing should be noted. While computer technology has had some effect on non-response, there are other issues which must be addressed in study design to achieve significant reductions in non-response. In particular, attention must be paid to reducing the percent of households who are not reached because they are not-at-home or utilize an answering machine to screen calls. Possible solutions include leaving messages on answering machines and offering incentives for a return call. Computer-assisted interviewing with call management capabilities offers some special benefits for this procedure. As with interviews terminated partway through, the ease by which call records can be retrieved from the data base simplifies the process by which research firms can take inbound calls and match these calls with the appropriate piece of sample.

Computer-assisted interviewing also appears to have positive effects on other sources of error – in particular interviewer and measurement effects. Computer-assisted call management capabilities have taken the burden of sample management from the field supervisors thereby allowing these individuals to focus more of their time and expertise on better training, monitoring, and retraining of interviewers over the course of a study.

Moreover, the benefits of computer-assisted interviewing in allowing for the accurate administration of more complex questionnaires, as well as more accurate and immediate entry of data have long been extolled and are further supported by this joint experience. These advances also indirectly impact the overall quality of data.

Finally, the considerable cost savings achieved through the use of these technologies cannot be ignored. With ever-tightening budgets as well as demands for increased precision in data resulting in larger sample sizes, the need for cost reductions remains a major issue in marketing research. The increased automation of traditionally labor-intensive tasks is likely to have a significant impact on the overall cost of data collection. This cost savings may allow research firms to increase pay rates at the interviewer level, thereby insuring higher quality interviewers and hence better data.

As our knowledge of the capabilities of marketing research continues to expand, it is likely that our demands in terms of the complexity of questionnaires that are administered, the sampling plans developed, and our expectations of the interviewers' skills in probing for the correct responses are likely to increase. Moreover, as microcomputers continue to improve in terms of speed and computing capabilities as well as decrease in cost, the barriers to their use in marketing research are likely to decrease. Computer-assisted telephone interviewing, with both the questionnaire and the

sample controlled by the computer, is likely to become the only acceptable methodology as both clients and marketing researchers demand the additional capabilities and benefits offered by this integrated system.

BIBLIOGRAPHY

Burg, Ellen (1986), "Computers Measure Interviewers' Job Performance," *Marketing News*, March 14, 1986, 36.

Burg, Ellen (1986), "Quality Research Requires Step-by-Step Control," *Marketing News*, January 3, 1986, 19.

Carpenter, Edwin H. (1989), "Software Tools for Data Collection: Microcomputer-Assisted Interviewing," *Applied Marketing Research*, 29, 23-32.

Curry, Joseph (1990), "Interviewing by PC: What We Could Not Do Before," *Applied Marketing Research*, 30, 30-37.

Gates, Roger H. and Jarboo, Glen R. (1987), "Changing Trends in Data Acquisition for Marketing Research," *Journal of Data Collection*, 27, 25-29.

Gershowitz, Howard (1990), "Entering the 1990s — The State of Data Collection — Telephone Data Collection," *Applied Marketing Research*, 30, 16-19.

Groves, Robert M. (1987), "Research on Survey Data Quality," *Public Opinion Quarterly*, 51, S156-172.

Groves, Robert M. and Magilavy, Lou J. (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," *Public Opinion Quarterly*, 50, 251-266.

Groves, Robert M. and Mathiowetz, Nancy A. (1984), "Computer Assisted Telephone Interviewing: Effects on Interviewers and Respondents," *Public Opinion Quarterly*, 48, 356-369.

Malhotra, Naresh K.; Taschian, Armen; and Mahmoud, Essam (1987), "The Integration of Microcomputers in Marketing Research and Decision Making," *Journal of the Academy of Marketing Science*, 15, 69-82.

Rothenberg, Randall (1990), "Surveys Proliferate, but Answers Dwindle," *New York Times*, October 5, 1990.

Comment on Elmore-Yalch and Glascock

Robert V. Miller
MarketVision Research, Inc.

This study is important because it is longitudinal and such studies are rarely done or reported upon. Longitudinal studies are valuable because they are able to track attitudinal information and permit subtle changes in data gathering structures, while holding other variables constant. Such an approach allows researchers to measure, over time, the effects of independent variables which relate to data quality.

RESEARCH DESIGN AND THEORY BUILDING

On a theoretical basis, this research still leaves the reader confused regarding the writers' arguments as to which variables in data collection are most important and why. Also, it is important to know the writers' perspectives on how issues related to data quality fit into a grounded theoretical framework of marketing research and how these issues relate to client expectations.

FINDINGS

The central finding of this study is that CATI systems and sample/call management appear to improve data quality. Other research in published studies has put forth much the same argument and probably most researchers believe that CATI and sample/call management improve data quality, particularly with complicated interviews, although not with every study. However, clear published evidence supporting the belief that CATI improves data quality is not abundant. Also, little published evidence supports the view that sample/call management is a major causal variable which leads to consistent improvement in data quality.

Other variables have also been shown to affect data quality, such as voice tonality/voice characteristics; perceived interviewer experience; interviewer training; interviewer supervision; time of day; interview length; and interviewer race, sex, and age. The difficulty with this study and others which attempt to confirm data quality is that they do not account or control for these other variables. As a result, our confidence in the conclusions put forth in this research — that CATI and sample/call management are the primary independent variables — is lessened.

PATH FORWARD

1. Modify design of research. A complicating factor in this study is that there were substantial changes in design from wave to wave, which were necessary and required by the client. It would have been easier to measure data quality improvements had the research been more consistent.

Moving from a purely field survey to a quasi-experimental design would allow for control of other data quality variables which could be held constant. An example of this approach is a "mode comparison study" in which two or more surveys with the same population would be concluded simultaneously, differing only in the mode of interviewing (for example, CATI versus paper and pencil).

2. Theory Building. The authors should move toward a more unified theory to support their research. It would be helpful to have more explanatory information and more insights as to how variables affecting data quality can be accounted for on a theoretical level, as a backdrop for this research. This would help to clarify the objectives of the research and provide for some discussion of other data quality variables which were not included.

QUALITY IN COMPUTER INTERVIEWING — TRICKS OF THE TRADE

Emily Wyatt
Burke Marketing Research

Quality has many definitions. Everyone of us would probably define quality in a slightly different way. Webster's Dictionary defines quality as " the degree of excellence which a thing possesses." For the purposes of this article, I would like to define Quality in Computer Interviewing as the collection of accurate and reliable data.

Computer Interviewing can add quality when used in either a telephone methodology or in a mall intercept. Many of the problems encountered when using Computer Interviewing are common to both methodologies, but let's look at the methods separately.

TELEPHONE

Comparison of Cost/Efficiency - Computer Interviewing vs. Paper

About a year ago, Burke was involved in a test which compared the efficiency of using a computer interview or a paper-and-pencil interview for a consumer telephone study. This was a study of average complexity with a ten minute interview length, and a variety of question types. These included unaided and aided awareness, ratings, and verbatim response questions. We used a national sample, and the qualifications were that the respondent be 18 or older and that there be at least one working television in the home. We had a quota of 100 completed interviews for each method.

Time Comparison for Computer and Paper Methods

	Hours		
Steps	Computer	Paper	% Difference
Building	17.25	NA	
Questionnaire Testing	4.25	NA	+207%
Data Entry	NA	7.00	
Data Preparation	17.00	18.00	-6%
Data Collection	76.50	101.00	-24%
Cleaning	2.00	13.00	-85%
Tabulation	17.00	17.00	0%
TOTAL	134.00	156.00	-14%

figure 1

The first three steps listed in figure 1 are applicable in only one of the two methods. Building refers to the creation of the computer interview. Building and testing of the questionnaire are specific to computer interviewing. The data entry step is used only in the paper-and-pencil method. At this point, if you compare just these three steps, the computer interview is taking more time. However, this time is more than compensated for in other savings, particularly the data collection phase.

Data prep refers to the process of editing and coding the verbatim responses. This process is a little easier when using a computer interviewing package. We saw a 6% time savings. The data collection phase used 24 fewer hours (a 24% savings), when the data were collected using a computer interviewing package. The other area of significant time savings was in the cleaning process. The cleaning took 11 fewer hours for the data collected with a computer interviewing package. This was an 85% savings of time for this phase. The actual tabulation of the data used the same amount of time, being independent of the method used to collect the data. Overall, using a computer interviewing package generated a 14% savings in time in this test.

The two phases with the highest percentage of time saved were the data collection and the cleaning. We can break these phases down further to see where the savings actually occurred.

Data Collection

	Hours		
Steps	Computer	Paper	% Difference
Briefing of Supv	1.00	1.00	0%
Briefing of Intv	11.25	13.50	-17%
Sample Preparation	0.50	4.00	-88%
Interviewer Hours	37.00	47.00	-21%
Telephone Connect Hrs	16.00	19.50	-18%
Supervision	5.50	7.00	-21%
Study Control	5.00	5.00	0%
Sample Disposition	0.25	4.00	-94%
TOTAL	76.50	101.00	-24%

figure 2

The time spent briefing the supervisor was the same for both methods. This time is used by the supervisor to become familiar with the study and to prepare to brief the interviewers. We do see a reduction in time spent training the interviewers, resulting in a 17% savings for this step. It takes less time to train an interviewer on a computer interview, because we don't have to spend a lot of time explaining the skip patterns. The computer interviewing package will handle this for us.

Sample preparation includes the set up of the sample in randomized replicates, dividing the sample by time zones, and creating a filing system for the interviewers and supervisors to use during the

dialing of the study. We used the computer sample management system in the computer interviewing half of the test. This system handles most of the sample preparation automatically, resulting in an 88% reduction in time for sample preparation. The sample disposition reporting was reduced by 94%, again because the computer package handles this automatically.

There was a decrease in the total interviewing hours with computer interviewing. This was due to a reduction in the interview length when using the computer rather than paper. The interviewer did not have to decide what question to ask next when using the computer, so the hesitation between questions was eliminated. The interview length when using the computer was 9 minutes. The paper half of the test had an interview length of 10.5 minutes. Not only did this reduce the interviewer hours, but it also reduced the "phone connect" time, which produced a further savings in phone charges.

There was an overall reduction in supervisory time used in the computer half of the test. This reduction was due to the absence of paper interviews to check. One interesting note, the supervisory time spent monitoring and developing the interviewers actually increased on the computer half of the study. This was not a necessity, but a benefit derived from the supervisor having more time available to spend in this way.

Cleaning

Hours

Steps	Computer	Paper	% Difference
Specs (program)	0.5	7.0	-93%
Specs (tested)	0.5	1.0	-50%
Edit Data	1.0	1.5	-33%
Pull problems	NA	1.0	——
Resolve problems	NA	2.5	——
TOTAL	2.0	13.0	-85%

figure 3

Data generated by a computer interviewing package are essentially clean when the interview is completed. In this test, very minor editing needed to be done. For the paper interviews, just writing and testing the program to clean the data took a day. This impacts not only the cost, but the timing available for the report on these data. A process that took 2 hours for the computer interviews took over a day and a half for the paper.

This test was for a single "average" study, but the results can be extrapolated for other studies. In general, a very simple study with easy skip patterns and a short interview length will see some increase in productivity, but not as much as we saw in this test. There will still be a benefit in terms

of timing for the report, with less elapsed time from the end of data collection to the final report. On the other hand, more time is needed before the start of the data collection phase to set up the computer interview. A more complex study will see an even greater increase in productivity than was achieved in this test. Complex skip patterns, a very long interview, or an interview where the interviewers frequently need to reference a respondent's previous answers will achieve the most benefit from using a computer interviewing package.

Common Problem and Solutions - Quality Issues

Some of the common problems that we see in data collection can be solved by using a computer interviewing package. As I mentioned above, an interview with a very complex skip pattern can be more easily completed by computer. The time needed to brief the interviewers will be significantly reduced. The supervisor's time will be freed to spend more time monitoring and developing the interviewer rather than checking very difficult paper interviews. The interviewing hours and phone connect time will be reduced, resulting in a savings in cost and timing.

Other common problems involve the logic of the answers given by the respondent. At times, the respondent will give an answer that is impossible. This may be an answer about a form of product that does not exist or it may be an answer that is not logical based on the previous responses given by the respondent. In either case, the interview can be set up to not allow the inappropriate response. The interviewers do not have to be familiar with all of the types and forms of products that are in the interview, nor remember all of the previous responses; the computer will do this for them. The problem can be immediately resolved by clearing the response with the respondent and getting an acceptable response. If it turns out that a previous answer needs to be changed, the interviewer can do that and then continue with the interview. The fact that the computer interview handles this type of problem helps to reduce the interviewer confusion. This is another reason that we saw a reduced interview length in the computer half of our comparison test.

When asking a question in which the answers must add to a particular number, the computer interview can be set up to handle the math. This could be a question about the percentages of time spent in certain activities, which must add to 100%, or a question where 10 points need to be allocated among a variety of products but must add up to 10. We do not have to rely on the math skills of an interviewer, who is also trying to complete an interview as quickly as possible without irritating the respondent. If the answers do not add to the correct constant, the interviewer can be alerted and the problem resolved immediately with the respondent.

In a study where the same respondents will be interviewed multiple times over the course of the data collection phase, such as a recruit and recall or a panel of respondents, you may want to use information from a previous interview in the current interview. This is much easier when using a computer interviewing package. The information can be included from a file into the new interview and then used for skip patterns. If the interview is paper, the interviewer needs to have the previous interview in front of him and use not only the information he is gathering in this phone call but

information from the previous interview as well. This is a very confusing process and mistakes are easily made.

In all of these cases, using a computer interviewing package makes the data collection phase easier and helps to create cleaner data.

MALL INTERCEPT

The reasons that you would want to use a computer interviewing package in a mall intercept method are very similar to those for a telephone method.

Timing can be reduced just as we saw in the telephone test. The time spent to clean the data will be significantly reduced, since data generated through a computer interviewing package are essentially clean when the interview is completed. The interview length will be shortened by reducing the hesitation between questions. Again, this is particularly true for interviews with complex skip patterns.

Most of the quality issues mentioned above are also true in a mall intercept method. The interview can be set up so that "impossible" answers will not be accepted. Any questions which require math can be handled quickly and easily with a computer interviewing package.

When using a computer interviewing package for interactive interviewing, the respondents are entering the data themselves. This significantly reduces the amount of time spent by interviewers and supervisors. The interviewer is needed for the mall intercept, but the main interview is administered by the respondent. There is a need for an interviewer to be available to assist respondents, but this interviewer should be able to oversee more than one respondent. We no longer need a one-to-one ratio of interviewers to respondents, when conducting the interview.

Common Problems and Solutions

There are problems that are unique to mall intercept interviewing. One of the problems that we experienced early in our use of computer interviewing for mall intercepts was a compatibility issue. The individuals in our system who are responsible for the scheduling of work are not very knowledgeable about computer interviewing. The people responsible for scheduling at the agency may or may not be knowledgeable about their equipment.

To resolve this issue, Burke sent a questionnaire to each of the agencies that we use most frequently. This was done just over a year ago. The questionnaire was basically an inventory of the equipment owned by each agency. Questions included the brand of PC, the type of monitor (monochrome or color), the version of DOS being used, size of RAM, and the size of the disk drives.

Of the 24 agencies in this survey, 21 agencies owned PCs. These agencies comprise a total of 112 mall locations, 88 of which were equipped with PCs. The total number of PCs reported were 356. This was an average of 17 PCs per agency that owned PCs and 4 PCs per location.

Summary of Field Agency Information

Equipment			Number	%
Monitors				
Color			261	75%
Monochrome			89	25%
Random Access Memory				
256K			44	13%
512K			15	4%
640K			288	83%
Type of Computer				
IBM XT			111	31.2%
IBM PS2 Model 30			66	18.5%
Standard Turbo			24	6.7%
IBM XT Clone			20	5.6%
Compaq 286			18	5.1%
Zenith Super Sport Portable			16	4.5%
Standard 88			11	3.1%
Zenith Z181 Portable			10	2.8%
ITT			9	2.5%
IBM PS2 Model 25			8	2.2%
Tandy			8	2.2%
Leading Edge/Hyundai			8	2.2%
Other			47	13.2%
Disk Drive Configurations				
Hard Drive	5.25	3.5		
1	1	1	2	0.6%
1	1		21	6.5%
1		1	29	9.0%
	1	1	7	2.2%
1	2		47	14.5%
	1		55	17.0%
		1	8	2.5%
	2		109	33.6%
		2	46	14.2%
Version of DOS				
2.1			54	15.6%
3.0			13	3.7%
3.1			29	8.4%
3.2			154	44.4%
3.3			72	20.7%
4.0			25	7.2%

figure 4

As you can see from figure 4, a variety of configurations are represented. Once you know what an agency owns, the compatibility issues are significantly reduced and the supplies can be shipped on either 5.25 or 3.5 inch disks, whichever is appropriate.

The other significant problem that we encounter is missing or bad data. This can be caused by the compatibility issue mentioned earlier, disks may be damaged in the mail, disks may be mishandled, or disks may be inadvertently reformatted or erased.

To avoid some of these problems, it is helpful to have a "dry run." Ship a "practice package" to each location at least one day prior to the start of the project. This package should include: a practice disk for each computer being used on the study, a paper copy of the interview, supervisors' instructions specific to the study, and a pre-addressed cardboard mailer.

The agency completes practice interviews on each of the machines to be used for the study. This also gives the interviewers an opportunity to become familiar with the interview. These practice interviews are returned before the study starts. The data on the practice disks are checked immediately. This assures that any hardware or software incompatibilities will be detected early.

As computer interviewing has become more common, problems with mishandling of disks and the errors which lead to erasing the data or reformatting the disks have become less frequent. If the practice disks do come in blank, the problem can be resolved before any "live" interviews have been lost.

Conclusion

The problems that a computer interviewing package can resolve and the benefits that it provides far outweigh the new problems inherent in any system of hardware and software.

CONTROLLING NON-RESPONSE BIAS AND ITEM NON-RESPONSE BIAS USING COMPUTER-ASSISTED TELEPHONE INTERVIEWING (CATI) TECHNIQUES

Michael Sullivan
Freeman, Sullivan & Company

INTRODUCTION

The title of this paper is perhaps a little ambitious. It suggests that computer-assisted telephone interviewing (CATI) techniques can be used to *control* two very troublesome "threats" to the validity of survey measurements. As will be apparent momentarily, non-response bias and item non-response bias cannot be completely controlled using any currently available techniques. Nevertheless, CATI techniques offer some very powerful and cost effective capabilities for reducing these sources of bias in surveys. In addition to facilitating tight control of non-response bias in telephone surveying, CATI techniques are extremely useful in mixed mode surveying — an approach to controlling survey non-response bias which combines the strengths of two or more survey modes. They also provide technology necessary to efficiently conduct interviews using a measurement technique known as bounded recall, a procedure which greatly reduces item non-response bias. This paper will first discuss non-response bias in general. Then, I will present some examples of surveys using mixed mode techniques and bounded recall, focusing on what we think works and what doesn't.

Let's begin by discussing non-response bias in a little more detail.

NON-RESPONSE BIAS DEFINED

What exactly are non-response bias and item non-response bias? There is a substantial academic literature discussing non-response and item non-response bias. A good overview of the problem is presented in Telephone Survey Methodology by Robert Groves, *et al.* (You should read Chapters 12 and 13 of this 1988 book if you want an overview of the subject.) In a nutshell non-response bias and item non-response bias are exactly what they sound like — bias in survey measurements due either to the fact that a respondent could not be contacted at all, or to the fact that the respondent refused or failed to provide some subset of the information sought by the surveyor.

Non-response is a necessary but not sufficient condition for non-response bias. Non-response bias (item or otherwise) actually has two components. *It is made up of the non-response rate and the difference between respondents and non-respondents.* For simple sample statistics such as means and proportions, non-response bias can be viewed as a simple linear function as follows:

$$y_t = y_r + (nr/(r+nr)\ (y_r - y_{nr})).$$

Where:

y_t = the "true value" of the sample statistic
y_r = the value of the statistic for the r respondents
y_{nr} = the value of the statistic for the nr non-respondents

This simple mathematical construct illustrates some interesting properties of non-response bias. First, it is clear that non-response bias is not simply the result of the non-response rate. A survey with a non-response rate of 99 percent may have little or no non-response bias if the difference between the observed and unobserved respondents is little or nothing. The converse is also possible. That is, a survey with a relatively low non-response rate (say 10 to 20 percent) may suffer from significant non-response bias if the difference between observed and unobserved respondents is sufficiently great.

SOURCES OF NON-RESPONSE BIAS

All major modes of survey contact (in-person, telephone and mail) are susceptible to non-response bias of different degrees and kinds. Until fairly recently the three modes of surveying were presented as competing alternative measurement techniques, with the primary determinant of choice among the alternatives being cost. The conventional wisdom has been that in-person interviewing generally produces superior response rates and data quality, followed by telephone interviewing, followed by mail surveying. However, in recent years, the apparent superiority of in-person interviewing over the other survey modes has been questioned. As systematic studies of non-response bias associated with the different modes accumulate, it has become increasingly clear that the different survey modes experience non-response for different reasons and therefore experience different (and potentially offsetting) non-response biases.

Looking at the possible outcomes of a survey contact, it is apparent that non-response bias can arise in a number of ways. *In general, any systematic failure in attempting to survey respondents can result in non-response bias.* Such failures can occur for the following important reasons:

1 initial contact cannot be established with the sampled respondent — because the respondent has moved, is not home, lives in a dangerous neighborhood or is somehow screening contacts with the outside world (for example, using security guards, secretaries or telephone answering machines);

2 respondents (or their "representatives") refuse to participate in the survey; and

3 respondents are physically incapacitated or unable to understand and speak any of the languages being used in surveying.

Non-response for some of these reasons is *prima facie* evidence of the existence of non-response bias.

Respondents who cannot write or speak the languages in which surveying is being conducted are very likely to be systematically different from those who can on a number of dimensions. They are likely to be less wealthy, possess less formal education and be less acculturated than those respondents who write or speak the languages in which the survey is being conducted. If these respondents constitute a significant fraction of the population, failing to include them in the survey will significantly bias survey results. In some regions of the United States, great care must be taken to control this source of non-response bias. In California, for example, about 8 percent of the general population does not speak English well enough to be interviewed using that language. The fraction of the California population that does not write English well enough to understand and respond to a mail survey is probably substantially higher. To control for non-response bias due to differences in acculturation, interviewing must be routinely conducted in Spanish in statewide surveys; and in some counties it must be conducted in Mandarin, Cantonese or Vietnamese to obtain representative samples. It is more difficult to control for non-response bias due to differences in literacy. Usually, interviewing by telephone or in-person is required in these populations.

Another fairly automatic source of non-response bias is non-response due to respondents living in a dangerous place or to their screening contact with the outside world. In my experience, this sort of bias is most often encountered in urban areas where significant segments of the population live or work in dangerous locations or in high security areas. The problem is particularly acute for in-person survey techniques. In fact, it has been suggested that in many urban areas in the United States, non-response bias due to these factors may favor telephone interviewing over in-person interviewing.

Of course, telephone interviewing is susceptible to other kinds of screening. In particular, screening resulting from use of telephone answering machines and from individuals who may refuse "by proxy" for the respondent (for example, a person answering the telephone refuses for the entire household.) Recent research at FSC suggests that answering machines are not a very significant source of non-response bias. In a recent statewide telephone survey in California, only about 4 percent of sampled observations could not be reached after 10 contact attempts because of the constant presence of an answering machine on the line. The proxy refusal and direct respondent refusal are far more serious problems, if only because most non-responses in telephone surveys arise in these categories.

While refusals have great potential to induce non-response bias, the mere fact that respondents refuse to participate in surveys is not necessarily evidence of the existence of non-response bias — except perhaps in extreme cases. To be sure, respondents do refuse to participate in surveys because of some aspect of survey content, which is likely to lead to non-response bias. However, they also refuse to participate for a large number of other reasons, most of which are probably unsystematic and unrelated to the content of the survey. Most respondents appear to be reacting

to the mode of the survey or to numerous other factors which are potentially unrelated to the content of the survey and thus are unlikely to produce non-response bias.

For example, in telephone surveys conducted at FSC about 60 percent of refusals occur before the appropriate respondent can be identified. That is, they occur during or immediately after the introduction to the survey. Moreover, in excess of 90 percent of refusals occur before the interview proceeds beyond the point of identifying the appropriate respondent. At this stage of the interview, the respondents have not really been exposed to the actual content of the survey so it can hardly be said that they are reacting to the content of the study — though it is clear they often are reacting to the mode of administration (that is, the cold telephone call).

Our interviewers, if possible, normally ask the respondent why they are refusing. (Interviewers seldom have time to ask this question, since most refusals at this point are rather definite and respondents usually hang up before the interviewer can speak.) Most respondents who answer this question indicate that they are too busy, that they are too tired, that they consider surveying to be an invasion of privacy or that that they consider the survey to be an unwanted inconvenience. Few mention anything in relation to the content of the survey as their reason for refusing to participate. If we take these respondents at their word, it appears likely that the majority of refusals in telephone interviewing are probably unrelated to survey content and thus are unlikely to produce non-response bias. (Numerous surveyors have reported similar findings. See for example, "Nonresponse: The U.K. Experience," by Collins *et al.*, in Telephone Survey Methodology, by Groves *et al.*, eds. John Wiley and Sons 1988.)

Of course, to the extent that the above reasons for refusing to participate in surveys tend to be geographically clustered, there may indeed be non-response bias induced by these differences. Urban populations, for example are much more likely to refuse to participate in surveys citing the reasons outlined above. There is reason then in surveys which target urban and non-urban populations (for example, statewide surveys) to pay careful attention to the effects that differential response rates from these areas may have on survey results.

CONTROLLING NON-RESPONSE BIAS — AN OVERVIEW

Except in relatively obvious cases such as those indicated above, researchers seldom know the extent of difference that may exist between respondents and non-respondents. Non-respondents by definition escape observation on most significant dimensions. Consequently, most efforts to control non-response bias focus on minimizing the survey's non-response rate and adjusting for non-response bias after the fact using analytical techniques when possible. (Efforts to analytically adjust survey estimates after the fact to take account of differences between respondents and non-respondents are not commonly used today, though more sophisticated survey designs sometimes anticipate the need for such adjustments and attempt to collect information that may be useful.

In practice, survey designs involving such adjustments are difficult to explain and defend because so little is known about non-respondents.)

Another reason that control of non-response bias tends to focus on non-response rates is that clients tend to have an obsessive concern with these rates. Clients usually have strong opinions about whether a response rate is "good" or at least good enough, based either on their training or their prior experience with the population of interest or the subject under study. They tend to use survey response rate as a sort of catchall indicator of the quality of the survey effort; and surveyors who want to keep their clients are well advised to manage their client's perception of non-response rates carefully. It is often the only indicator that will be used to judge the quality of the survey work that has been undertaken.

CONTROLLING NON-RESPONSE BIAS USING CATI TECHNIQUES

Computer-assisted telephone interviewing offers a number of facilities that can be used to manage (though not eliminate) non-response bias at the survey and item levels. These include use of CATI systems to:

1 cost-effectively enhance overall response rates for all kinds of surveys (mail, telephone and in-person) ;

2 collect information to augment survey measurements (taken using mail or in-person survey techniques) for purposes of making statistical adjustments to population level estimates;

3 measure the effects of non-response bias from mail and in-person survey techniques;

4 collect survey information recursively — collecting information for survey items that were previously not completed by the respondent (in mail or in-person surveys) — allowing researchers to eliminate existing item non-response bias; and,

5 provide respondents with information that facilitates bounded recall — preventing the occurrence of item non-response bias.

Techniques one through four require either that the survey be conducted over the telephone or that telephone interviewing be integrated with other survey techniques such as mail and in-person interviewing. Use of a CATI system in telephone interviewing can greatly enhance the economic efficiency of interviewing, sample management and respondent data management — making possible the execution of mixed mode survey designs that would be otherwise prohibitively expensive to accomplish. However, a CATI system is not technically required in using the first four techniques. The last technique is virtually impossible to execute without a CATI system.

CATI AS AN INTEGRAL TOOL IN SURVEYING

Unlike a watched pot, survey data have a tendency to fall painfully short of expectations if they are not continuously and closely inspected as they are being collected. To ensure data quality, professionals who have a substantive understanding of the data being collected (consultants and project managers) should be able to routinely and easily inspect the operational results of surveying and analyze incoming data to identify problems that may be occurring. To facilitate this process, I believe the CATI system should be fully integrated with the other research facilities that may be part of the survey shop.

At FSC we conduct in-person, mail and telephone surveys, and various mixed mode versions of these surveys. To facilitate overall survey operations we have integrated our CATI facility with the other computer systems used by our consultants and project managers. We use a CATI/Ci2 (Ci2 System for Computer Interviewing) system in telephone surveying. The system runs on a dedicated 20-station computer network using Novell Netware v. 2.15. The CATI facility is connected to the 17-station front office network (also Novell Netware v. 2.15) using an internal bridge.

Because these systems are completely integrated, professionals working in the front office can easily attach to the CATI facility even when it is in operation. This allows them to:

- inspect results of operations (such as completions and refusals)
- observe interviews in progress (not very useful but it impresses clients),
- analyze incoming telephone survey data quickly and efficiently, and
- transfer sample management data to and from CATI/Ci2 (useful in mixed mode surveys and surveys using bounded recall).

MANAGEMENT OF TELEPHONE SURVEY RESPONSE RATES USING CATI TECHNIQUES

With the above operating design, it is possible for consultants or project managers to quickly analyze the operational or substantive results of surveying without leaving their desks. Depending on the professional's familiarity with CATI/Ci2 and other available software systems, he or she can analyze survey results using the "canned" facilities available in CATI/Ci2, or can load the data into one of the data base or statistical packages available on the front office computer network.

Consultants and project managers who are skilled in dBase have become adept at moving files back and forth between the call management data base (DB.CON) and dBase. This facility makes possible fairly in depth analyses of the results of survey operations. For example, it is possible to analyze response patterns by area code, telephone prefix, zip code, city and other geographic location information (if known). In practice, the professionals do not routinely use this facility to monitor non-response rates in the laboratory. Instead they tend to rely on the summary reports provided by the laboratory staff on a weekly or nightly basis. They also rely primarily on the laboratory supervisors to monitor the performance of interviewers and take corrective action as

required. Analysts tend to analyze the call management data base to answer unusual questions such as:

- how successful have been efforts to date to convert initial refusals;
- how long are the interviews taking and how much time on the average is being spent in the various stages of interviewing;
- proportionately how many answering machines are being encountered and how many calls are typically required to "get around" them;
- how are response rates varying by geographic location; and
- in mixed mode surveying where we are telephone interviewing to follow up on unreturned mail surveys, how many of the respondents who are claiming to have sent in their surveys actually have done so.

In addition to analyzing the call management data base, professionals also have the ability to rapidly load the results of interviewing into data base and statistical packages such as SAS and SPSS. To facilitate this process, we have written specialized programs which parse the results of the Ci2 CONV2 program (the program that translates Ci2 results into ASCII format) and load them into dBase files. From dBase, these data can be analyzed directly or easily transferred to SAS or SPSS for subsequent analysis.

This facility is used in three ways. First, it is used to inspect early results from the survey. On more occasions than I like to admit, we have identified potential problems with question wording and logic by analyzing early survey returns. This facility provides us with the ability to do so. Clients also like to have preliminary results from telephone surveying. Often they are working under serious time pressure for preparing reports, and they like to use preliminary data to prepare analysis programs and begin "getting a feel" for the data that will eventually arrive. Finally, the facility is used in preparation of final survey deliverables.

MIXED MODE SURVEYING USING CATI TECHNIQUES

Mixed mode surveying offers a powerful approach to controlling and measuring non-response bias in surveying. By combining the various survey modes, it is possible to significantly reduce uncertainty about results due to the possible presence of non-response bias.

For example, mail surveys to businesses often produce relatively low response rates because the surveys are not addressed to a person in the organization who has the authority and responsibility for maintaining the information being sought. Response rates between 10 and 25 percent are quite common in this circumstance. It is difficult to have much confidence in survey results which contain such a large potential for non-response bias.

It is possible to significantly improve response rates to mail surveys of businesses by initially identifying the appropriate respondent in each business, through telephone interviewing. In this way, if the target of the survey is the purchasing manager, the survey gets delivered to the purchasing

manager, who expects its arrival and has agreed to participate in the study. Using this two-stage survey technique, response rates on critical variables ranging from 65 to 80 percent are likely. By collecting basic information during the initial telephone interview that is critical for judging the eventual existence of non-response bias in the mail stage of the survey, it is possible to systematically study the presence of non-response bias and adjust for it if any is found. There is some additional cost involved, but few would argue that the improvement isn't worth it — especially if the validity of the data might eventually be challenged.

There are other mixed mode survey combinations that help to control non-response bias. These include:

- *telephone-mail-telephone designs* — surveys which initially identify the respondent by telephone, send a self-administered instrument in the mail, and call back to the respondent to collect the required information;

- *mail with telephone follow up to non-respondents* — the objective of the telephone follow up is to observe the differences between non-respondents to the mail survey and others who were willing to provide the information over the telephone. It doesn't completely eliminate non-response bias, but it can provide greater confidence in mail survey data and a means to adjust survey results to take account of non-response bias.

There are other mixed mode survey techniques which greatly reduce the cost of interviewing but have unknown impacts on non-response bias. For example, it is possible to combine telephone interviewing with in-person interviewing — using the former to identify and recruit respondents to the latter. The most commonly applied sample design used in in-person interviewing is the area probability sample usually with clustering. Surveys based on such designs are very difficult and expensive to carry out. If any selection criteria are applied within the sample (for example, sampling only for households with adolescent children), the costs of surveying using this technique may be prohibitive. Moreover, because of screening and other problems outlined above, these designs tend to be susceptible to serious non-response bias in urban populations. Telephone surveying to "recruit" respondents to in-person interviewing can greatly reduce the cost of in-person interviewing, and it is less susceptible to non-response bias due to screening and other biases. However, refusal rates on the telephone using this technique are quite high. For this reason, the jury is still out on this approach to surveying.

CATI techniques offer great improvements in efficiency over manual survey management approaches, in accomplishing mixed mode surveys. Mixed-mode surveying typically requires that data be transferred either into or out of the telephone survey mode (for example, respondent name, address and other particulars that may be needed).

For example, in a telephone-mail survey, respondent contact information is typically collected during the telephone survey mode and used to address the mail survey mode. Other information obtained in the telephone mode may also be used to identify the appropriate mail survey version that will

be sent to the respondent, if necessary. In this circumstance, it is possible to directly transfer respondent contact information from the CATI system into the mail processing system being used. This is particularly helpful in continuous surveying where results of one day's telephone interviewing "drive" the next day's mail survey batch. This approach also can be used to improve the quality of the mail survey materials. By loading respondent address information into word processing systems such as Microsoft Word or Word Perfect, it is possible to prepare highly personalized correspondence introducing the survey — an approach which has been shown to dramatically affect response rates.

Moreover, if the CATI system is tightly integrated with the other research facilities being used, it is possible to effectively track the status of sample observations as the survey proceeds. In mixed-mode surveying this can be a particularly difficult problem, because it is easy for respondents to get lost in the relatively automatic data collection process. Some of the clients we work for are extremely concerned about "bothering" the respondents (their customers). When respondents complain to the client that they don't want to be bothered anymore by the survey, it is necessary to exclude these respondents from further contact attempts. This can be easily be done if the current status (mode and stage) of the respondent is known. If the CATI system is tightly integrated with the other tracking systems being used to track the status of cases, this is an easy problem which can be solved in a few minutes.

USING CATI TECHNIQUES IN STUDIES USING BOUNDED RECALL

Finally, CATI systems are particularly useful in survey instrument designs which employ bounded recall as an approach to controlling both response and non-response bias. A major source of error in survey measurements is the respondent's memory. In some cases, the respondent's ability or inability to remember important details may significantly affect survey results. If respondents cannot remember details that are required to answer survey questions, they are less able to answer them. Moreover, respondents are more likely to refuse to continue participation when this occurs.

For example, in a recent study we performed for the Pacific Gas and Electric Company (PG&E) we were asked to determine whether or not a sample of their customers who had received advice concerning the costs and consequences of certain energy conservation options had implemented them. The recommendations were made to these customers from 1983 through 1989. The study was conducted in 1990.

Two separate "programs" were evaluated using similar surveys. However, in the case of one of the programs, detailed digital records had been maintained on the recommendations that had been made to each target respondent. With minor editing, it was possible to directly load the details of the recommendations and the consulting contact into the CATI/Ci2 system. Target respondents for this program were given advice between 1983 and 1986.

For the other program, which took place between 1987 and 1989, detailed information was not available concerning the recommendations that were made. All that was available were the name and address of the target respondent.

The differences between the response rates for the two studies were significant. For the earlier program, about 45 percent of respondents could be located and could remember receiving the information provided by PG&E. For the later program, about 31 percent of respondents could be located and remember receiving the information. Differences in recall rates were more dramatic. For the earlier program, about 80 percent of respondents could remember having received the specific recommendations (read to them by the interviewer) and could report what their organization had done about them. For the later program, only about 55 percent of respondents could remember specific recommendations well enough to be able to report how their organization had acted.

This example demonstrates that providing respondents with cues and other bounding information can significantly improve response rates, even for events and information which occurred quite some time ago.

Because surveys using bounded recall essentially involve a "unique" survey for each respondent, a CATI system is almost required to complete them. Features of the CATI which are critical to this kind of study are:

- the ability to load digital information in the survey system from outside,
- the ability to display it at the screen,
- the ability to act on imported data logically for purposes of controlling the flow of the interview; and
- the ability to randomly vary the order of presentation of questions, to control bias that might be induced by the ordering of questions presented at the screen.

CONCLUDING REMARKS

CATI systems used in conjunction with survey techniques designed to control non-response bias can reduce problems associated with non-response bias. It seems to me that using a CATI system as I have described is one of the best available approaches to controlling such bias. This paper has shown you some of the approaches that can be used.

However, non-response bias remains a sticky problem, which is not likely to go away any time soon. It is a problem which is probably getting worse in all modes of surveying. Since many readers are involved in telephone surveying, I would like to end with a few comments about what I think the major challenges are for controlling non-response bias in telephone surveying.

Telephone surveying using RDD sampling techniques offers an attractive sample frame and relatively inexpensive interviewing costs. However, there is one big problem with this approach to surveying — the rate at which respondents or their representatives refuse to participate.

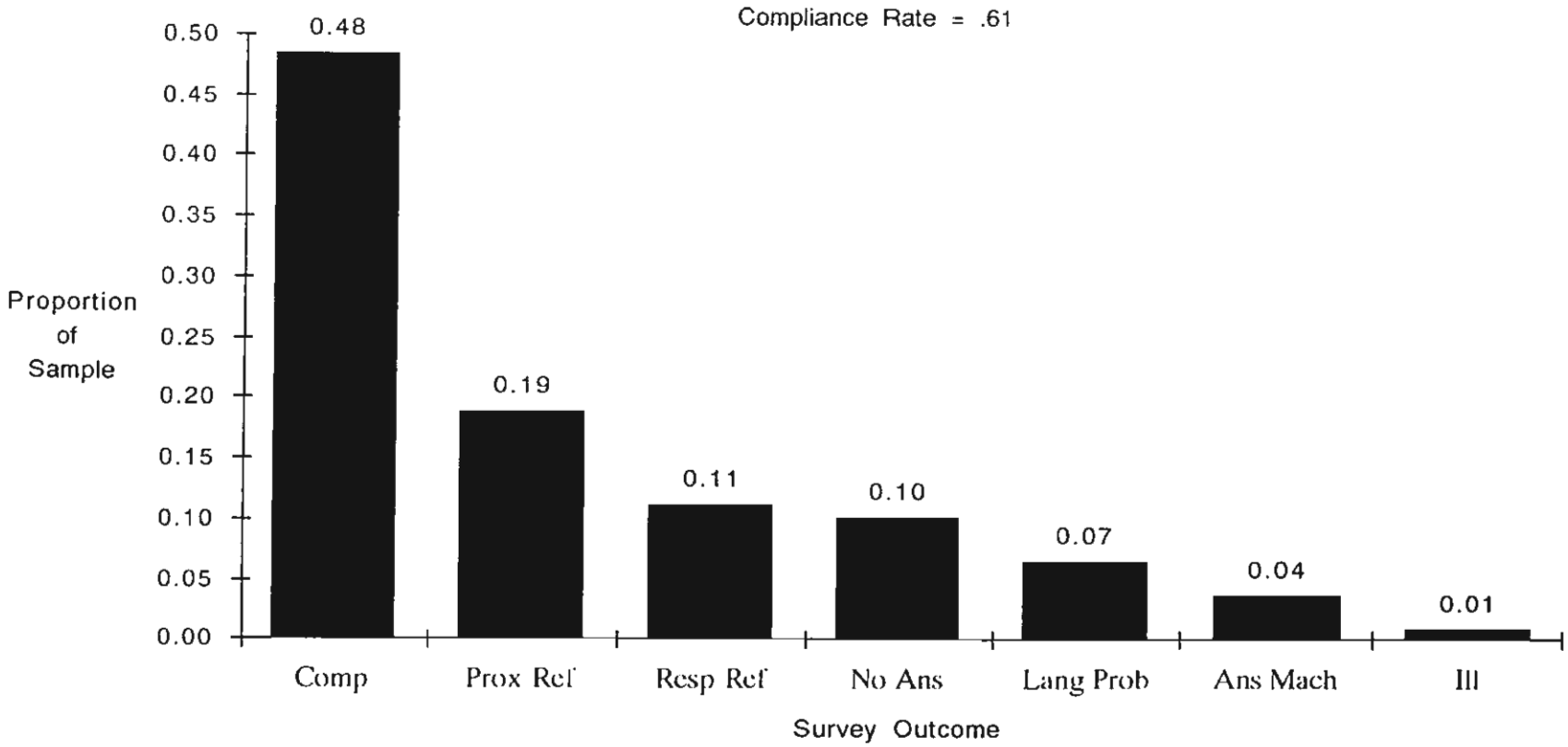
To date, the most common approach to controlling refusal rates is to attempt to "convert" refusals. This approach is costly and typically doesn't yield much improvement in the refusal rate. In our experience, in the California market only about 15 percent of initial refusals can subsequently be persuaded to participate in the study. Considering that the average compliance rate for surveys in California is about 50 percent, this sort of improvement falls somewhat short of being impressive.

It seems to me that a significant methodological breakthrough will be required to really control the problem of refusal rates in telephone interviewing. We as surveyors probably need to shift our attention away from trying to reduce the size of refusal rates to trying to measure or understand how those who refuse might be different from those who do not. Remember, it doesn't matter whether people refuse to participate in a survey if the probability of their having done so is unrelated to any of the measurements we are taking.

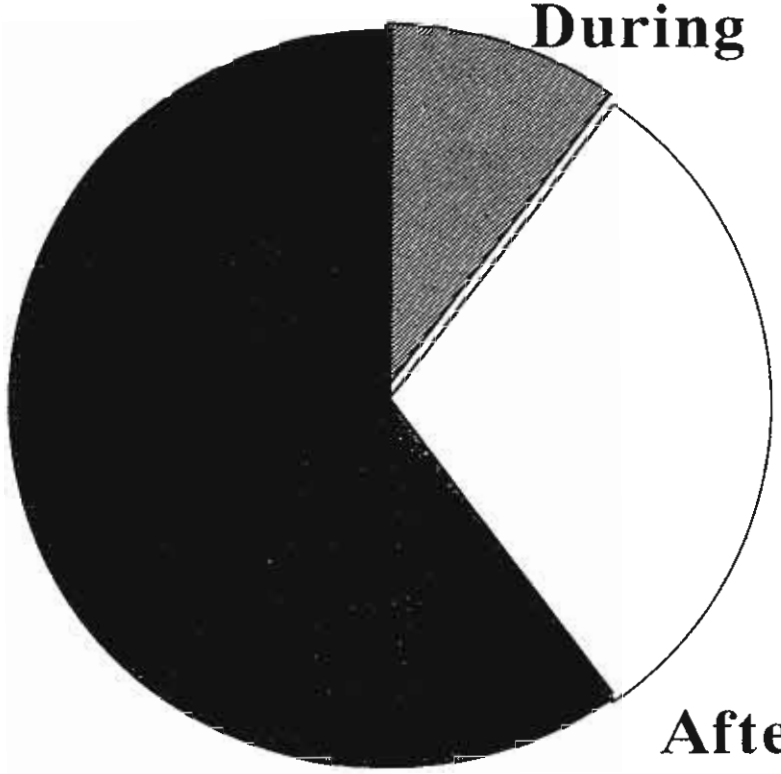
I think there are some ways of doing this which involve systematically studying refusals in an experimental fashion — but that is another topic.

California

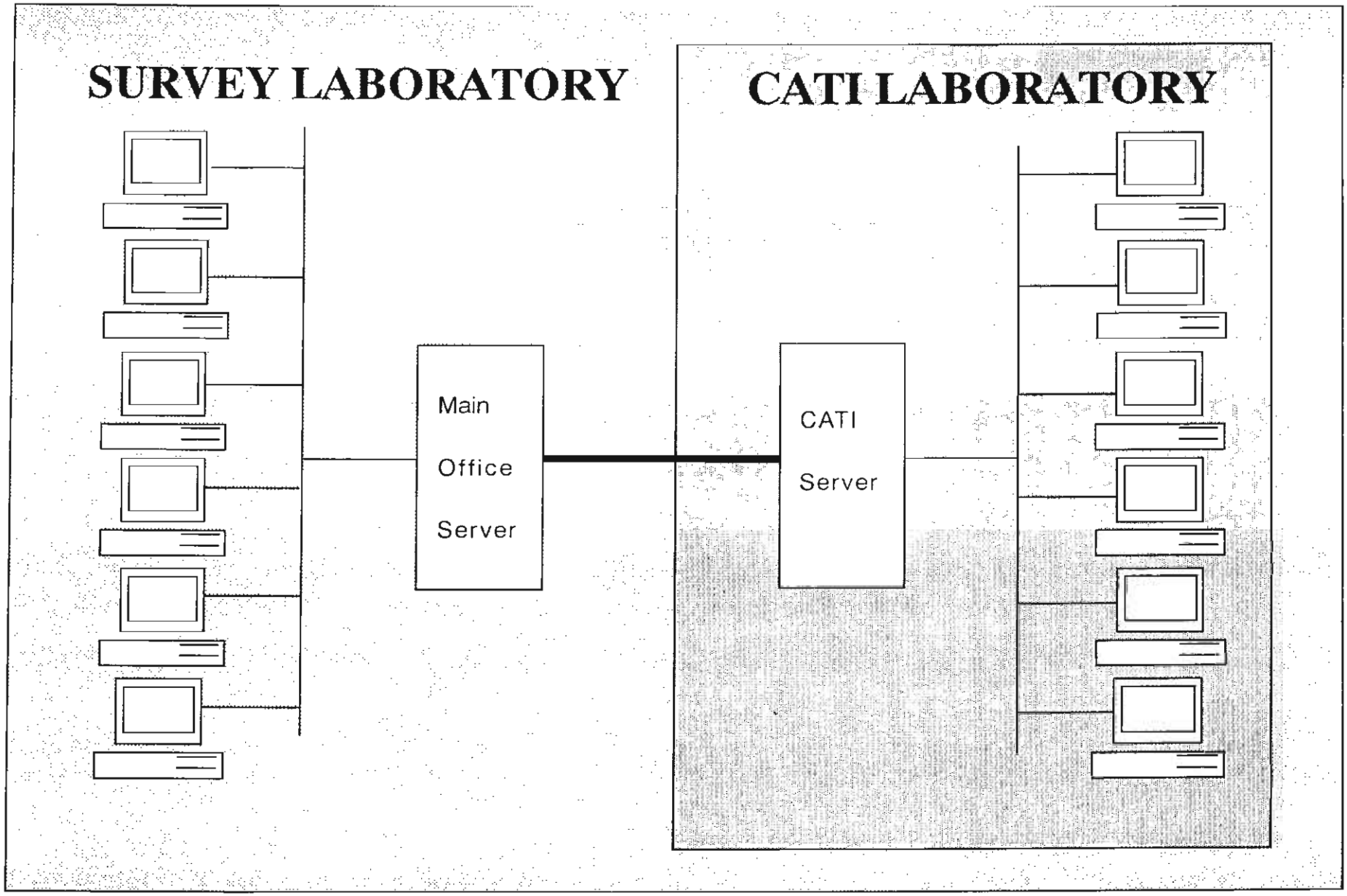
Results of Survey Operations
1990 Statewide Survey of California Households
Regarding Alcohol Consumption and Attitudes About Prices
Sample Size 1635



Typical Distribution Of Refusals
By Stage of Progress Through Interviewing
Random Digit Dialing -- California



SCHEMATIC DIAGRAM: INGETRATED CATI FACILITY, -- FREEMAN, SULLIVAN & COMPANY



Comment on Sullivan

Robert L. Zimmermann
Maritz Marketing Research

There was very little presented in the Sullivan paper with which I would disagree. We have employed most of these methods in coping with non-response bias, with perhaps some differences in emphasis. I am simply going to add a few addendums to his excellent review.

Documenting non-response bias is relatively simple, and it is surprising that there are not more hard data available. Whenever multiple callbacks are included in the design, it is only necessary to break down demographics by the callback count on which a resolution is finally reached. Perhaps the problem is that those studies with six to eight callbacks have little need to analyze non-response bias, and such clients may not be interested in funding analyses that benefit less thorough researchers.

We have received reports from clients that "intent to buy" for state-of-the-art consumer electronics and communications devices increases at least to the sixth callback. There is also considerable anecdotal evidence that demographics of respondents differ by time of day and day of the week, so that callbacks should be distributed across the call time frame.

Multiple callbacks are expensive, and there is an inevitable tradeoff between sample size and number of callbacks. There is always a point at which the non-response bias is trivial compared with standard error of estimate for our statistics. The trouble is, we can provide precise power functions to delineate the impact of changing sample size, but our estimates of non-response bias are often very fuzzy. We find clients frequently put their money behind increased sample sizes.

We use mixed mode telephone and mail interviewing routinely. In addition to costs, other factors suggesting use of mail surveys include the need to utilize visual materials, product placements, very long questionnaires, and questionnaires that are more easily and accurately completed if the respondents can reference their files. To decrease non-response bias, we use telephone pre-recruit, call-back of non-respondents, reminder mailings, telephone recording of mailed questionnaires, monetary incentives, and attention-getting devices such as sending the questionnaires via Federal Express.

Non-response bias may be especially critical in some segments of business or professional (especially medical) respondents. Target populations may be very small, and particular potential respondents may be critical because of their high volume within the defined market. The problem here is less that of being unable to contact the respondent than that of inducing cooperation. The incremental cost per interview may run into hundreds of dollars, and in such cases we will often use large incentives.

While there seems to be more concern over non-response bias in telephone and mail interviews than with in-person interviewing, one form of in-person interviewing is at risk of very large distortions due to selective interviewing: Mall intercept samples are distorted by the demographics of the mall, the wide variations in time-at-mall of different individuals, daily cycles of mall usage, and whims of both the interviewee and interviewer. The probability of being selected for a mall intercept interview varies across individuals by a factor of 20 to 40 or more.

When the sample basis is individuals rather than households, and more than one individual within the household qualifies, a common practice is to randomly select an individual within each household among those qualifying. This may necessitate a second callback and a second possibility of refusal. The probability of a second refusal is directly related to the number of qualified respondents within the household, and thus subject to all the biases associated with household size.

I agree completely with the stress placed on collecting telephone interview data via an integrated CATI system, with the ability to track the factors impacting non-response through the course of a study. Without on-line recording of detailed screener data, it is virtually impossible to analyze non-respondent factors at the individual respondent level. In monitoring a large-scale study, these data should be accessible on a daily basis, at least during the early phases of the interviewing. In addition, it is difficult to randomly assign multiple callbacks by time-slot, without computer control.

OF MICE AND MEN, AND WOMEN

Ray Poynter
Sandpiper International Limited, UK.

INTRODUCTION

The purpose of this paper is to share our findings with respect to using mice (and to a lesser extent other analogue devices) for data entry in computer-aided personal interviews (CAPI). The paper starts with an application where Sandpiper have acquired a broad skill in using mice to collect data that would have been difficult or even impossible without analogue input, and goes on to review more ambitious uses of the mouse using techniques we are still developing. The final two sections deal with some of the details of getting the technology running and describe our experience with other data entry media.

In keeping with the spirit of this conference the paper will attempt to remain in contact with the theme "Doing things we couldn't do before," and an attempt will be made to define why these things couldn't be done previously.

The bulk of Sandpiper's studies feature and centre upon the collection of matrices of item by attribute ratings. It is in this context, in particular, that we have gravitated towards the mouse as our primary medium for data collection. It should be stressed that techniques which offer significant advantages in this niche may be less applicable for other types of studies. Indeed on most occasions we still use a keyboard to select such things as demographic breaks.

The findings and examples in this paper are drawn from Sandpiper's extensive programme of mouse assisted computer aided personal interviewing. We conduct more than 20,000 such interviews per year, across eleven countries and four continents and have used mice not just in conjunction with Roman characters but also with Japanese and Thai scripts.

This paper sets out to review the advantages of using mice as a tool for data entry and concludes that for some types of studies there are savings in time and improvements in measurement. During the course of this paper I refer frequently to IBM personal computers; the same comments would be true of any high-quality IBM compatible PC, and in fact the majority of the machines we use are laptop IBM compatibles.

ONE DIMENSION

All researchers will be familiar with the five-point scale, the seven-point scale and other versions of the linear scale. The well established problems with this scale are: the effort required to complete them, the effects of the labelling of the intervals, the limited variation that such scales can produce (particularly at the individual respondent level), and the time required to collect the data.

CAPI systems such as Sawtooth's Ci2 (Ci2 System for Computer Interviewing) provide the opportunity to use analogue scales. These scales allow a cursor to be manoeuvred with cursor keys and record the comparative positions of the items. Lesley Bahner in her paper, "Long Self-Administered Questionnaires," included in the 1988 Sawtooth Software Conference Proceedings, provided useful coverage of the advantages of the analogue scale. In particular the paper illustrated the increased levels of meaningful discrimination that analogue scales can provide.

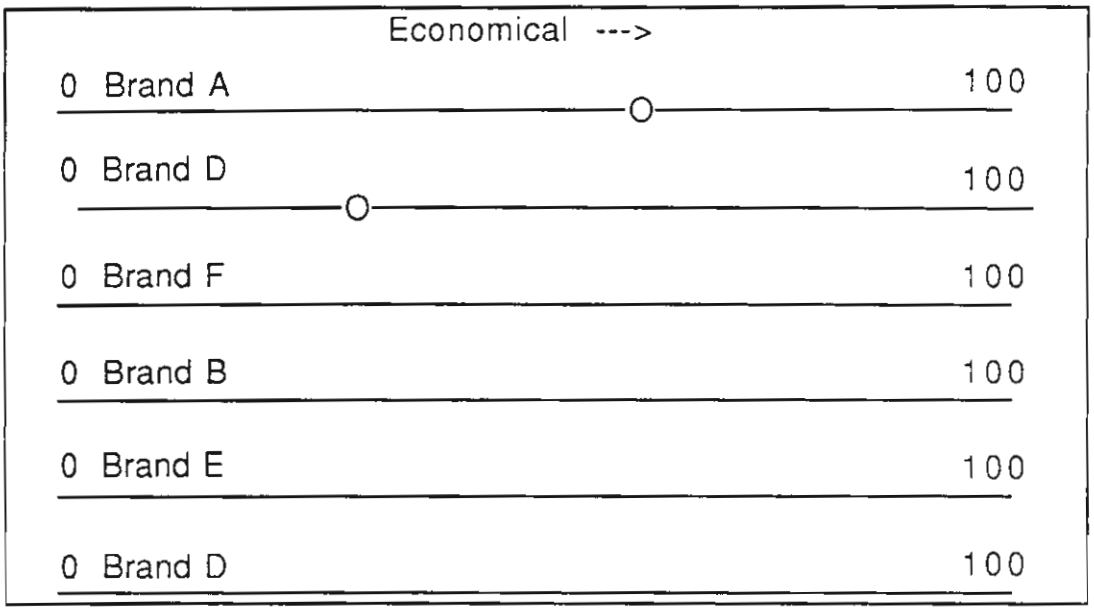
One outcome of the increased discrimination that analogue scales provide is that more options become viable down at the individual respondent level. Where a respondent's attribute ratings are constrained to a seven-point scale, the researcher is restricted by the inter-product variation that is possible and the analysis is similarly constrained. As the number of intervals increases so do the options with respect to analysing data at the individual respondent level.

Perhaps the most interesting outcome of this sort of data is the ability to construct disaggregated consumer simulation models based on analogue scales, including emotional as well as rational scales. Such models allow simulations where the options presented to the "electronic respondents" are consistent with the paradigm of the models' construction; we are not forcing continuous input into an highly discrete structure.

As is indicated in Dr. Ian Greig's paper ("Interviewing by computer for explanation and prediction in sensory analysis: a case example for fragrance" *Food Quality and Preference* 1,4/5,187-90,1989), the use of mice to move and control the cursor during the rating on analogue scales can produce additional benefits over and above those already mentioned. As detailed below, the mouse reduces respondent fatigue and increases data sensitivity and speed of collection. In addition the mouse enhances the positive elements referred to in Lesley Bahner's paper. In particular, the fact that the respondents perceive even an interview of 60 to 90 minutes as fun.

The data Sandpiper tends to collect consist of one or more matrices of item-by-attribute ratings, plus a few demographic classifications and assorted pick n items from a list of m-type questions. A standard screen is shown below. Each time the respondent presses the left mouse button the position of the cursor is recorded and the cursor moves to the line below. As the respondent moves the mouse left and right the cursor moves left and right. All up and down movements by the mouse are ignored. The software uses the right button to allow the respondent to move back up the scales to correct earlier entries.

Normally the software randomises the order of the attributes between interviews and the order of the items for each attribute, to balance out any order effects.



When the cursor reaches the bottom line displayed the respondent enters his or her response in the same way as on the other scales. The software then scrolls the list of brands up and displays the next scale to be rated.

This system ensures that during the rating process the respondent is using nothing more than the mouse and its two buttons. This provides a degree of security from the respondent pressing unwanted keys. Most of the people we interview have never used a mouse before the interview and some familiarisation is necessary. On the first screen the interviewer "trains" the respondents by asking them to place the marker on the mid-point, the end point, under the first letter of the item name, etcetera. This process is repeated until the respondent is clearly competent — this typically takes about 60 seconds. It IS that easy!

The only other elements of training relate to what happens when the respondent reaches the last item on the screen, the last item for the first attribute, or wants to correct earlier entries. This training process has not proved to be a net speed handicap. However, survey designs which include a smaller amount of scaling may find the net speed reduced.

It is worth pointing out that at present we use the mouse for data entry only for analogue scales and for the two dimensional uses outlined in the next section. For selecting an item from a list (and of course for typed-in responses) we still rely on the keyboard (for now).

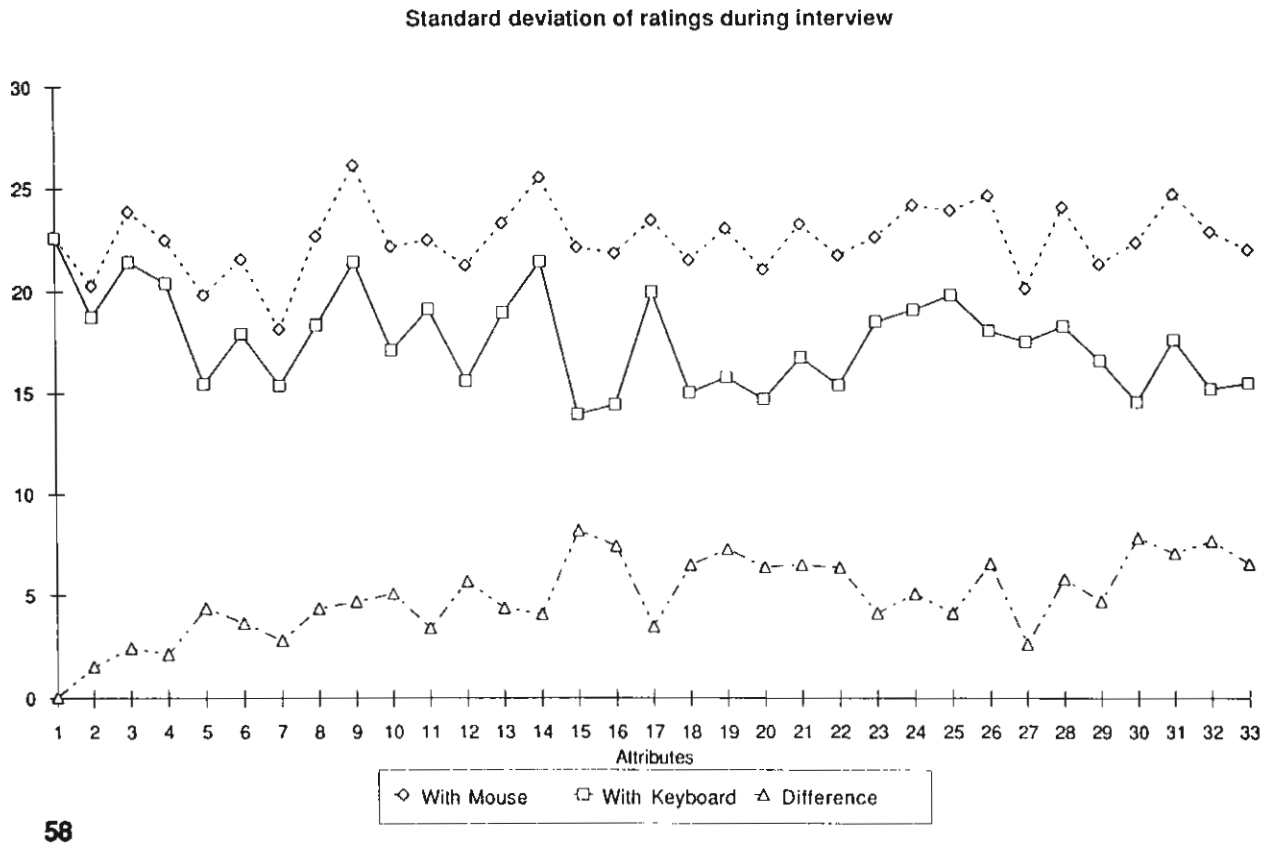
To cater to the logistical problems that can and do happen, the software can be operated without the mouse, using the four cursor keys. This feature has allowed us to compare cursor-key entry with mouse entry. The following example is drawn from one of our regular tests to monitor performance.

The data were collected as part of an ongoing shampoo study in the UK. The data represent two matched samples of respondents (100 in total) answering for 30 shampoo brands on 33 attributes. For the purpose of this experiment the order of the attributes were held constant, whereas normally the software would randomise the order of attributes between each interview.

Not surprisingly, the basic inferences that can be drawn from the data do not differ between the keyboard data set and the mouse-captured data. There are, however, differences in the data that are useful to report.

The first difference is in the length of time taken to complete the interview. The keyboard data recorded a mean time of 65 minutes, including a few selection and demographic classification sections. The mean time for the mouse-based interviews, (which required as part of that time mouse familiarisation) was 50 minutes. The time taken for the scaling part of the interview was about 60 minutes for the keyboard medium and 45 minutes for the mouse medium. The mouse showed a significant 25% reduction in the time taken to complete the scaling matrix. This finding is consistent with other monitoring exercises we have conducted.

The second difference relates to the quality of the discrimination achieved on the scales. The graph below plots the standard deviations achieved by both the keyboard and mouse entry data sets, and the differences between the two. The numbers were arrived at by first calculating for each respondent the standard deviation for each attribute across the shampoo brands. The members of the two data sets were then averaged to produce a figure for each attribute for both groups.



The data clearly, and significantly, show differences in the discriminative abilities of the data entry media. The mouse exhibits a basically rectangular distribution, indicating no fall off with time. The keyboard data consistently score lower than mouse data and exhibit what may be modest fatigue effects; this inference is drawn not so much from the shape of the curve relating to the keyboard data but from the differences between it and the mouse data. The data indicate that performance of keyboard data entry can be improved upon by using a mouse. The data also show that, even in a keyboard situation, the discriminative ability of an attribute is determined more by what the attribute is measuring than by the position in the interview it occupies; a finding that echoes Lesley Bahner's findings.

In conclusion to this section let me restate our findings. Mice can be a real asset in producing better data, faster. The interview does not exhibit the pattern of fatigue that can occur with long keyboard interviews. From in-depth discussions with respondents we have concluded that the physical and mental effort in using the full scale is no greater than that when using just part of it.

TWO DIMENSIONS

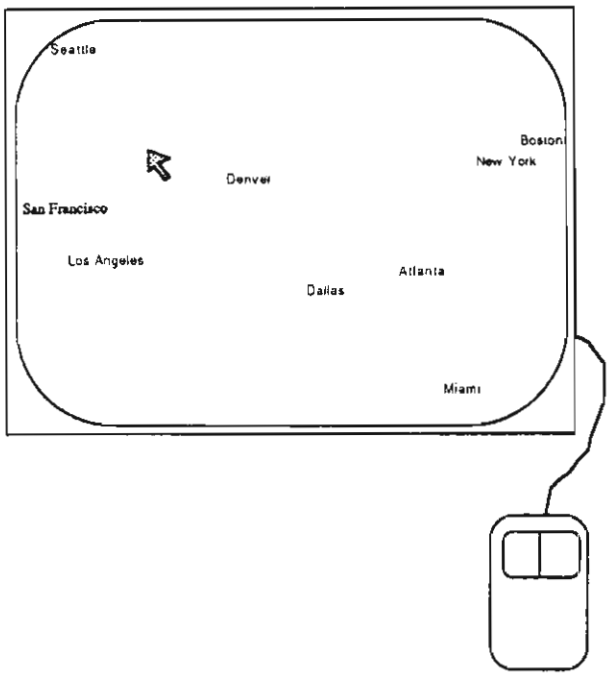
We have used the mouse in its fuller sense of a two dimensional device in a number of different ways.

We would like to share one of the uses that the system has been put to, namely that of positioning new data on existing perceptual maps. This is achieved by presenting the map on the screen and recording where the respondent thinks the additional point should be positioned. Our experience to date has been that respondents do not readily take to positioning a pointer in two dimensions. Consequently when we ask respondents to position items on maps we ask the respondents to point to the position on the screen and we use the interviewer to position the pointer.

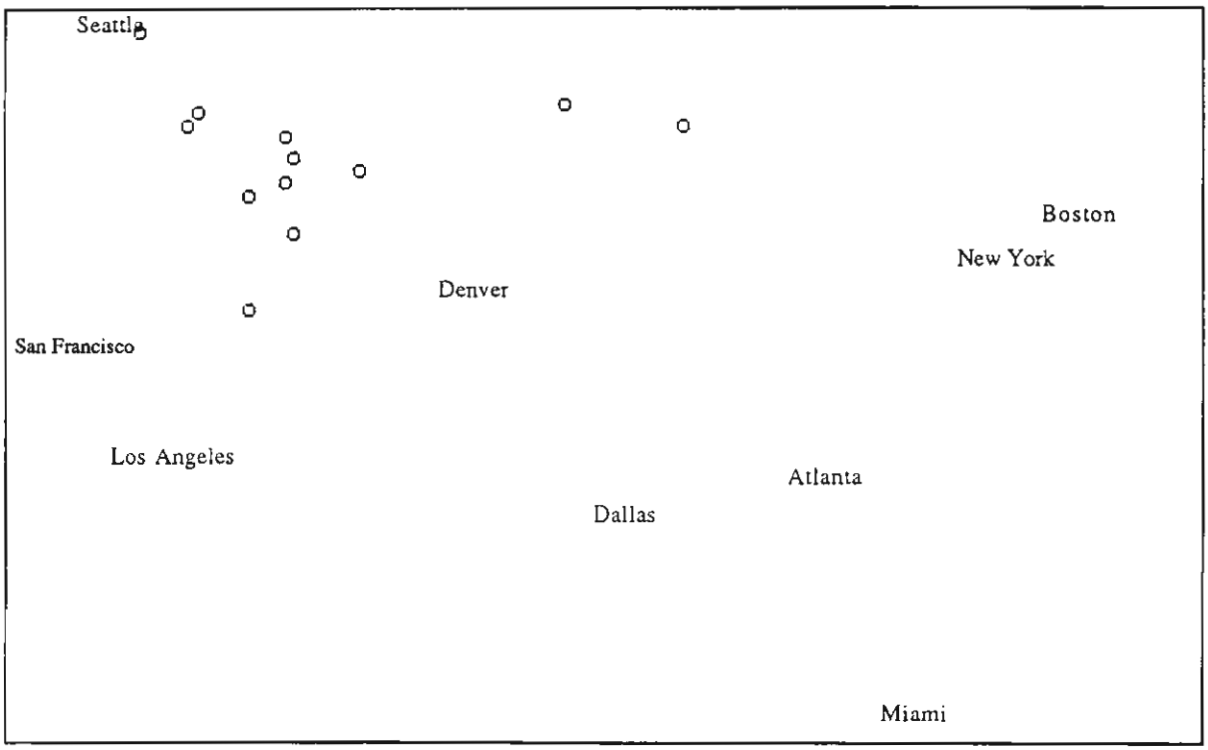
The advantage of this technique is that no new calculations have to be made. Most researchers will have suffered from the habit of perceptual map software to re-orient maps when additional data are added. The points that are added to the map are actual representations of the positions the respondents see the brand in; there are no intervening calculations or assumptions. A presentation can readily be constructed showing the distribution of the brand positions.

One needs to be aware, of course, that the technique is only valid if it is used to present new data in the context of an accepted original map. If the original map is dated or contentious then the additional data will be of little practical use.

The following is an example where, for the purposes of this paper, we interviewed a sample of twelve Nottingham residents (located in the immediate vicinity of Sandpiper's office) and asked them to place Salt Lake City on a map of USA cities. A representation of the interview and the resultant data are displayed graphically below.



ESTIMATES OF SALT LAKE CITY'S LOCATION



We have found this approach very useful in getting positions for new or re-launched products. With maps based on research rather than cartographical research we would of course use a number of views (x versus y, x versus z, y versus z, for example).

In addition to the above, rather straightforward application, we have used similar techniques in more advanced ways. For example we have constructed interviews where the respondent is offered a variety of maps in order to estimate which attributes differentiate the respondent's perception of a brand from either his or her own earlier perception or a group perception. We have even used the power of the computer to generate two dimensional screens from the respondent's own data, from earlier analogue scales, enabling interesting image-shift type analysis.

One view of the mouse is that it is simply a better method of collecting analogue scale data. However, this view would miss the new opportunities that using the mouse in two dimensions can offer. New research possibilities are opened up with this approach. We at Sandpiper have only begun to scratch the surface of the new possibilities and we feel convinced that there are significant iconic and pictorial designs that will take research a quantum leap forward.

WHO IS DRIVING MICKEY?

Although the mouse is becoming increasingly common on PCs, it is still not supported directly by the MS-DOS operating system. Before the mouse can operate, it is necessary to load an external device driver. The Microsoft mouse is supplied with a device driver in the form of a program called MOUSE.COM which remains resident once run. Other manufacturers usually supply similar drivers with their products.

The mouse driver installs itself as DOS interrupt 51 and all calls to the mouse must use this interrupt. There are some 35 different function calls that can be made. It is possible to set up the parameters for each call in the machine registers and issue the interrupt directly. However, this is unnecessarily tedious. Microsoft sell a Mouse Programmer's Toolkit and Reference Guide at a modest price. The Reference Guide is essential reading for anyone contemplating mouse programming as it contains the definitions of all the mouse functions and several example programs in high level languages.

The Toolkit contains a library called MOUSE.LIB which can be linked with any Microsoft high level language, such as QuickBASIC, Pascal and C. It can also be linked with Microsoft compatible object code produced by other supplier's compilers, such as Borland's Turbo-C. The library essentially contains only one function which acts as the interface between the program and the driver interrupt. The function exists in several different versions for the different memory models available in some compilers.

The definition of the function in Turbo C is:

```
extern void cmouses(int *m1, int *m2, int *m3, int *m4);
```

This is the function for the small model. There are equivalent functions called cmousem, cmousec and cmousel for the medium, compact and large models. There are always four parameters to the function. Their usage depends on which mouse function is being requested, but it is often as follows:

m1 function number (0 to 36, not 17 or 18)
m2 button number
m3 horizontal cursor coordinate
m4 vertical cursor coordinate

To make programming the mouse more straightforward, it is worth spending a small amount of time developing a library of specific mouse functions, each of which calls cmouses with the correct parameters. Examples of some of the common functions are as follows:

```
/* Mouse Reset and Status - function 0 */
void MouseReset(int *status, int *buttons)
    {
    int    m1, m2, m3, m4;

    m1 = 0;
    mouse(&m1, &m2, &m3, &m4);
    *status = m1;
    *buttons = m2;
    }

/* Show Mouse Cursor - function 1 */
void ShowMouseCursor()
    {
    int    m1, m2, m3, m4;

    m1 = 1;
    mouse(&m1, &m2, &m3, &m4);
    }

/* Hide Mouse Cursor - function 2 */
void HideMouseCursor()
    {
    int    m1, m2, m3, m4;
```

```
    m1 = 2;
    mouse(&m1, &m2, &m3, &m4);
    }

/* Get Mouse Button Status and Position - function 3 */
void GetMousePosition(int *status, int *cursorH, int *cursorV)
    {
    int    m1, m2, m3, m4;

    m1 = 3;
    mouse(&m1, &m2, &m3, &m4);
    *status = m2;
    *cursorH = m3;
    *cursorV = m4;
    }
```

The unit of movement that is returned by the mouse is the Mickey. A Mickey is equivalent to a standard mouse moving over a distance of one two-hundreths of an inch. Perhaps we could call this *Mickey Mouse research.*

More recently Microsoft have introduced Mice that record units of one four-hundredth of an inch, still calling it a Mickey. Clearly if your software is configured to respond sensitively to one two-hundredth of an inch then the new resolution can cause startling effects. If your cursor starts behaving like Mickey Mouse on amphetamines then it is probably because you have a modern mouse sending back modern Mickies.

OTHER MEDIA

At Sandpiper we have experimented with a number of data entry media . The following is a brief note on our experiences and conclusions:

Paddle:

The first non-keyboard device we used was a paddle. A paddle is a simple device that was widely used on games machines and came as standard with the Apple II microcomputers. The paddle is a small plastic box, with a button and a dial, connected to the computer by a cable. The button registers simply on and off while the dial acts upon a potentiometer, the computer reading this as an integer in the range 0 to 255.

For one-dimensional work the paddle is a fine device for data entry. The respondent is able to sit well back from the work surface, the training is quicker than that with a mouse and the respondent can choose whether to rest the device on a surface or to hold it in one hand and turn it in the other.

As a two-dimensional data device the paddle is a disaster. It requires two paddles to be used simultaneously, with the subtlety of touch of a brain surgeon.

When we moved to the IBM PC, Sandpiper produced some in-house paddles for this machine, created out of mice that we first broke and then reassembled with a dial and new case.
As devices these were very popular with our field people but impractical. It would not have been viable to support and insist on third party companies using our own proprietary device.

Light Pen:

We first used a light pen back in 1982, attached to an Apple II computer. The CAPI software which was written for a paddle was easily converted — in AppleBasic nothing was too complicated. The system had great style but was far too tiring to use. Trials and real usage both confirmed that it is simply too much hard work to keep your hand up near the screen for the 30 to 90 minutes that our typical interview takes.

Joystick:

We have not made extensive use of joysticks, using them on occasions as back-ups if paddles or mice are not available. They have proved to be almost as friendly as paddles and mice with respect to one-dimensional data entry (most respondents seem to find it confusing for such an obviously two-dimensional device to be used in a one-dimensional way). As a two-dimensional data entry medium the joystick has proved to be very friendly and intuitive. The real disadvantage of the joystick is, like that of the paddle, the lack of availability. Also the variability between one joystick and another can lead to problems.

Touch Screens:

To date we have not used touch screens. For the sort of data we collect the problems we found with the light pens are likely to be repeated.

Stylus:

Our belief is that the market research of the future, indeed the computer of the future, is going to have a lot in common with the Linus Write-Top, displayed in a market research context at the 1989 Sawtooth Software conference. The flat screen looks like a note pad and the stylus is like a pencil on its note pad. The stylus can be used to check boxes, write numbers or characters, indicate positions on analogue scales or two-dimensional maps, and has a host of other possibilities. This machine of the future will depend: upon reliability (can it read 99% of people's handwriting?), cheapness, availability and support.

Plethysmograph:

One novel, but disappointing route we followed was the plethysmograph. One of our key project analysts took a plethysmograph and hooked it up to an IBM computer and investigated the sort of data we could collect over a period of six months. The device was a small clip that fitted on the ear lobe and recorded the flow of blood. Despite producing quite a lot of interesting data, we were unable to produce repeatable usable data. However, the experience gained has proved useful in linking various devices to computers at data collection time.

Comment on Poynter

Steven M. Cooley
Blue Cross and Blue Shield Association

The author has done a good job of discussing types of peripheral devices available for interactive data entry. My comments are about how these technologies affect data collection.

BENEFITS OF THE TECHNOLOGY

Using mice or other auxiliary computer devices seems to yield two distinct benefits. One is in the realm of measurement, or more specifically, the process of data collection. The second is in the data analysis. The paper identified a number of data collection economies including lower time requirements, less respondent fatigue, and the potential for respondent enjoyment of a relatively novel experience. These aspects are clearly of concern to applied research interests as the industry must balance respondent rights, client information needs, and budgetary constraints.

On the academic or basic research side of the aisle, data sensitivity appears to be the greatest benefit of collecting analog data in the manner presented. As mentioned in the paper, analog scales tend to produce more variability and therefore increase the options for data analyses at the respondent level. While our own research objectives may differ from the author's, I believe we all might like to increase the sensitivity of our measures, regardless of the unit of analysis.

DATA QUALITY

The quality of data collected in computer interviewing has been a major conference topic. This paper describes a brief but effective training process for respondents to become familiar with the mouse device. From my experience in using Apple Macintosh products, I can attest to the naturalness of the device and ease in adapting to using it in both one- and two-dimensional tasks. I believe that respondents would be better able to handle the two-dimensional problems posed to them if given the opportunity to use the mouse in a graphically-based application during the training. I also feel that the mouse could easily be extended for use in the collection of categorical data. The author may already offer this option to survey participants.

MEASUREMENT THEORY

I would be remiss if I shunned my academic training and failed to consider the role of measurement theory in this discussion. A quick review of any basic measurement or psychometric text reveals a vast literature on measurement scale issues. After decades of attempting to refine scales and interval levels, I feel the papers presented by Lesley Bahner (1987) and Ray Poynter signal a

renewed interest in natural scales. I, too, believe analog scales may better approximate naturally occurring metrics by allowing respondents to self-explicate intervals. However, the role of anchors is still an important issue. The semantic value of words used as endpoints could greatly affect respondents use of the scale.

AGENDA

Many market researchers reside in the grey area between basic and applied research disciplines. "Pure" academics must deal with pragmatic issues when engaging in consulting work. Practitioners also have a vested interest in expanding the knowledge base while pursuing client or company objectives. While taking these issues into account, I propose that this paper presents an agenda for each.

Applied research should continue to press for new ways of presenting stimuli and collecting data from respondents who are becoming increasingly difficult to locate and expensive to measure. Increasing advances in microcomputer capabilities should prove a fertile platform for new and interesting approaches such as multimedia presentations on one machine. Basic research will be necessary to validate new methods and can help to resolve the scale anchor issue which remains with us, after all these years.

Finally, I think we should all note that this is neither frivolous nor trivial research. The paper's title belies its relevance to the industry in demonstrating how one firm deals with "Doing Things We Couldn't Do Before." As such, it is an important contribution which the author has chosen to share with the research community.

REFERENCE

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

A COMPARISON OF FACTOR ANALYSIS AND CORRESPONDENCE ANALYSIS AS DATA REDUCTION TECHNIQUES WITH CLUSTERING IN MARKET SEGMENTATION

Marsha A. Wilcox
Decision Research Corp.

There are many research methods and statistical techniques to conduct market segmentation. The purpose of this paper is to compare factor analysis and multiple correspondence analysis as data reduction techniques in the applied setting of market segmentation. It should be noted that the comparison here is of an applied nature. While it is possible to use a tetrachoric or polychoric correlation matrix (or polyserial for both continuous and discrete data) as input for the factor analysis, allowing a more direct comparison with the correspondence analysis axes, common practice is to use a Pearson product moment correlation, as was done in the present analysis.

The purpose of this comparison is to illustrate the difference between the results of two approaches, one being a relatively common practice (factor analysis) and the other relatively new (correspondence analysis). Each technique will be discussed briefly followed by an analysis using data concerning leisure time and activities.

Ultimately, the evaluation of the worth or appropriateness of a segmentation scheme is its utility to those initiating the study. There are no statistical tests that prove or disprove the utility of a particular cluster or segment solution. Therefore, a definitive measure of the superiority of one technique over the other is not available. The degree to which the segmentation scheme meets the purposes and objectives stated prior to the research then becomes a standard for evaluation.

Generally, the purposes of market segmentation are: a) to identify and locate the group or groups most likely to respond to marketing and/or advertising efforts, and b) to identify and profile product/ service user groups. An effective market segmentation scheme is characterized by homogeneity within segments and meaningful differences between segments. Additionally, the segmentation scheme should provide adequate face validity with segment profiles that can be easily identified as discrete types of customers. The differences between clusters or segments on continuous variables may be examined with techniques such as analysis of variance (ANOVA), and with chi-square (c^2) for categorical data, to name only two. These two will be the criteria considered with the examination of each of the methods.

PURPOSE AND METHOD OF THE STUDY

The purpose of the study was to examine attitudes and behaviors concerning leisure time and activities and define useful segments for marketing and advertising purposes. Six hundred telephone interviews were completed in January, 1987. The sample was nationally representative of households with annual incomes of at least $25,000. Respondents were 21 years or older. Men and women were equally represented in the sample.

AREAS OF INVESTIGATION

Items assessing attitudes toward work, leisure time, money, response to advertising, shopping, and risk taking were included in the survey instrument. Several classes of behavior were inventoried: use of leisure time, media consumption, and spending related to leisure time. Demographic information was also recorded for each respondent.

Factor Analysis

Factor analysis assumes that the observed (measured) variables are linear combinations of a smaller number of underlying dimensions or factors. It is based on the correlations among the measured variables. The correlation normally used is the Pearson product moment correlation, with the assumptions that the data are normally distributed, and are at least interval level in nature. Factors are derived from the patterns of relationships among the variables. Principal components analysis generates factors (components) that explain the variation in the observed, or measured, variables.

Determining the appropriate number of factors or components involves several considerations. First, there is the tension inherent in explaining as much of the variation in the data as is possible while minimizing the number of factors retained. Each successive factor retained should explain at least as much of the remaining variation as an average variable (hence the eigenvalue >= 1 "rule of thumb"). Finally, from an applied perspective, the factors must be interpretable and useful in the situation at hand. Usually, an examination of the eigenvalues for the principal components solution will provide a guide for the number of factors to be retained.

The factors, or components, are usually rotated to optimize discrimination between factors and consistency within factors. Most often the rotation is orthogonal in nature. Other types of rotation (for example, oblique or procrustes) may be employed when there are theoretical reasons for doing so. Such reasons may include an hypothesized market structure, or a confirmatory factor analysis in which the purpose of the study is to replicate, if possible, prior findings.

Once the factor structure has been determined, factor or component scores are computed for each individual in the sample. The scores are then used as input for a clustering procedure. Scores based on principal components are particularly amenable to a variety of clustering algorithms because they are standardized (with a mean of zero and a standard deviation of one). Often, the clusters are profiled, using the appropriate bivariate or multivariate statistics, on the variables used in the creation

of the factor structure as well as other variables in the study (using crosstabulations of the categorical variables). Descriptions of the clusters are then used in marketing and/or advertising efforts. For general advertising purposes, attitudinally distinct groups are valuable because the primary form of communication with consumers and potential customers is usually through broadcast media, where that information is used to create ads to appeal to those groups. However, in the arena of direct or targeted marketing, clusters with clear demographic as well as attitudinal differences are most useful.

Segments based on factor analysis are usually constructed using attitudinal data. Fuller profiles, including relevant demographics, reported behavior, and other available data, are usually generated once the segments have been defined.

Example Data

Principal components analysis was conducted on the battery of twenty two attitudinal items included in the survey. Seven principal components with an orthogonal rotation were retained, based on interpretability. There were eight components with eigenvalues greater than one. The scree plot showed a distinction at six and eight following the principal components analysis. However, after examining the interpretability of the components, it was decided that the seven factor solution was optimal. This solution accounted for 50% of the variation in the data. The principal components and variable loadings appear in Appendix A.

The seven factors were subjectively named, Risk, Thrift, Indulgence, Conservatism, Advertising, Control and Busy.

Each respondent was then assigned a score on each of the components. The cluster analysis was done using these principal component scores. The SAS Fastclus procedure was used. Fastclus performs disjoint clustering based on Euclidean distances computed from quantitative variables. Three through ten cluster solutions were examined. The four cluster solution was chosen, in part, based on the cubic clustering criterion (SAS technical report A-108). A profile of each cluster was developed using crosstabulations of the categorical behavioral and demographic data. Additionally, the c2 for each table was calculated. Other continuous data, including the component scores, were compared using t-tests and analysis of variance. As would be expected, there are differences between cluster means on almost all of the components. The fourth factor failed to assist in differentiating the clusters. A t-test of the means within clusters (testing the hypothesis that the mean of the population from which the cluster is drawn is different from zero) was also helpful in defining the clusters. Figure 1 shows the mean scores for each factor within each cluster.

A brief description of each of the four clusters follows. The descriptions include those variables for which there were significant differences in the cluster means or the chi-square for the crosstabulation table was significant at the .10 level.

Cluster 1 — Risk takers

The members of this group see themselves as risk takers, although much less so than the Fast Track segment, described below. In fact, their only positive component mean is on the "Risk" factor. This factor is characterized by three statements: a) "I enjoy activities that are fun, even if they're a little bit dangerous," b) "I like to take risks in my leisure time," and c) "My favorite leisure activities require physical effort." The members of this group are decidedly not indulgent nor are they in great control of their spending habits. They don't see themselves as thrifty. While their media habits in general are not significantly different from the other consumer segments, they are slightly more likely to enjoy TV commercials.

This group's members are more likely to be male (63%), and married. Forty-six percent are between the ages of 35 and 54. As is consistent with their attitudes, they prefer outdoor sports in which they can participate with their partners. However, jogging does not fall into this category for them. This group tends to spend leisure time at home or at a racquet club.

Their shopping behavior is not markedly different from the other clusters, except that they are somewhat less likely to shop in discount stores and do not tend to shop by mail. They are less likely to own an instant camera than the other groups and do not regularly use a bank ATM.

Cluster 2 — Indulgent Homebodies

This group's members are decidedly not risk takers. They see themselves as indulgent, but in control of their spending. Though they are in control of their spending, they are not thrifty nor are they bargain hunters.

These members tend to be female (71%) high school graduates working part-time. Fifty-nine percent of this group are in the 35-54 age bracket. They prefer to spend leisure time indoors and prefer watching to participating in sports. However, many in this group own exercise bicycles. They are slightly more disposed to spend their time alone than with partners or friends. This is consistent with their preferred activities: reading books and magazines and watching TV.

While their media consumption is relatively high, the members tend not to buy based on advertising and they are less likely to report enjoying print or broadcast advertising than the other segments.

Their spending habits include shopping by phone and mail order as well as specialty stores.

Cluster 3 — Thrifty Ad Watchers

The members of this group see themselves as busy people who are thrifty and in control of their spending.

The Thrifty Ad Watchers are truly "tuned-in" to advertising. Eight-three percent agree that certain television commercials are enjoyable and 79% have similar feelings about print ads. They do not perceive themselves as risk takers, nor are they indulgent.

This group's members tend to be slightly older than those in Cluster 1 or 2. They tend to be from the south (35%) with annual incomes below $35,000. They prefer to spend leisure time at home, and prefer indoor activities such as watching TV and reading magazines. In fact, this group is the most likely to report watching TV daily. As would be expected, the members of this group prefer watching to participating in sports. They are by far more likely to seek leisure activities that are free (89%).

Consistent with their income, the members of this segment tend to shop at discount stores and through the mail.

Cluster 4 — Fast Track

Risk-taking and indulgence top this segment's component means. The Fast Track are by far the most adventuresome. A full 93% of this group enjoy activities that have an element of danger, and 75% report that they like to take risks in their leisure time. Seventy-one percent agree with the statement that they are "the first to try something new."

The members see themselves as indulgent, but in control of their spending. Seventy-four percent agree with the statement that they will buy something they want, even if it costs more than they planned to spend. They tend not to be influenced by or enjoy advertising.

The members of this group are more likely to be younger (44% are in the 21-34 age bracket), college educated (49% have at least completed college), single (32%) men (61%). They tend to be employed full time and report more hours at work than any other segment. They prefer outdoor activities away from home in the company of others, more often friends than partners. Participatory sports are more compelling than watching sports for them. Evidence of this can be seen in their inventory of sporting equipment: golf clubs, skis, tennis rackets, and running shoes. This group's members are most likely to jog regularly (11%). Further evidence of their interest in health and social interaction can be seen in their clubs: health clubs, racquet clubs and golf clubs.

As would be expected, this segment is less likely to spend time reading books and magazines or watching TV.

While this group's members are active, they also pursue other interests and are more likely to own a 35mm camera than those in other segments.

Their shopping takes them to both discount and specialty stores. In fact, this group's members are more likely than any other to shop in specialty stores. They are also the most likely to use bank ATMs.

Correspondence Analysis

Until recently, correspondence analysis has been relatively obscure in the U.S. In fact, it was not included in standard statistical software until 1990 (SAS release 6.06). The technique has had different names in different fields; among them: dual scaling (both rows and columns of the data matrix), optimal scaling, reciprocal averages, principal component analysis of qualitative data, and simultaneous linear regression for contingency tables. The reader is referred to the paper by Hoffman & Franke (1986) and Michael Greenacre's 1984 text for further historical descriptions.

Correspondence analysis is a multivariate descriptive statistical technique that graphically represents the rows and columns of a categorical data matrix (Hoffman & Franke, 1986). Unlike most multivariate methods, the only data requirement, in terms of data type or distribution, is that the input data matrix be rectangular and contain non-negative numbers. This is a great advantage in that demographic data, categorical behavioral data and other categorical data can be included in the analysis, as can rating scale data.

Like factor analysis, correspondence analysis is a data reduction technique. For segmentation purposes, the input data matrix has the individual respondents as rows and categorical variables as columns.

Continuous, or pseudo-continuous data such as Likert type scales, are entered as categories. It is important to note that Likert type scales are often thought of as truly continuous variables. There are several reasons why they are not. First, respondents do not all use the scale (any scale with a small number of points) in the same manner. An examination of the response patterns will usually show that there are respondents who tend to use the extreme ends and others who do not. In other words, the scale has different meanings for individual respondents. Also, the distributions on such items may be (but probably are not) normally, or near normally, distributed. Often there is a significant skew, or bimodality, in responses. These factors make this data type less than ideal for traditional multivariate techniques assuming truly continuous measures.

In correspondence analysis, the data are reduced to principal axes, each respondent is given a coordinate position in the multidimensional space, and then coordinates are used as input for the clustering procedures. The outcome is segments that are driven by all of the relevant variables, not just those continuous, normally distributed variables that are amenable to traditional factor analytic methods.

Example Data

Categories for the attitudinal (four point Likert-type) variables were computed based on their distributions. Frequencies were computed for each of the variables of interest; in this case, the attitudinal battery plus relevant media and purchase behavior. Principal axes were computed and the six-factor solution was chosen based on a histogram of the eigenvalues. Coordinates on each

of the principal axes were computed for each respondent. These coordinates were then the basis for a k-means type, followed by hierarchical clustering algorithms. Three through seven-segment solutions were examined.

The four-segment solution was optimal given this method. The clusters were then profiled with respect to the variables driving the segmentation and other relevant variables from the questionnaire.

The segments have been named the Penny Pinchers, Catalogue Shoppers, Trendy Spenders, and High Rollers.

As with the discussion of factor analysis, a brief description of each of the segments follows.

Cluster 1 — Penny Pinchers

The Penny Pinchers tend to have lower incomes and to be somewhat older than the other segments. This group's members are characterized by conservative attitudes toward spending and risk-taking. They do not engage in leisure activities that require any physical effort such as jogging or tennis, nor do they engage in other leisure activities that require spending such as going to the movies. They are likely to spend time at home reading or watching TV. They are the least likely of the consumer segments to own a VCR or subscribe to cable or pay TV.

This group's members are accessible to advertisers through television, newspapers, and magazines. However, they shop much less than members of other segments. When they do shop, their preferred channel is the department store and not mail or telephone. In fact, they are suspicious of mail/phone shopping with credit cards.

Cluster 2 — Catalogue Shoppers

This segment is predominantly female, well educated, and affluent. The members are the most active group of consumers in the sample. They are busy shoppers, taking advantage of mail, phone, specialty and discount stores. They see themselves as thrifty bargain hunters.

This group has a relatively low interest in activities involving physical effort. The members are, however, as likely as more athletically-oriented segments to own sports clothing and equipment.

The Catalogue Shoppers report heavy use of broadcast and print media and are generally open to advertising.

Cluster 3 — Trendy Spenders

Most members of this group are male (60%). The Trendy Spenders are well educated and affluent. They are predominantly middle-aged (35-54). Members of this group are very interested in physical activities, including sports for exercise, participatory sports (including hunting and fishing), as well as

watching sports for entertainment. While members of this segment are somewhat open to taking risks and spending, they also consider themselves thrifty and in control of their spending.

This segment's members are moderate consumers of print media, but heavy consumers of broadcast TV and pre-recorded video materials. They are the heaviest users of cable TV among the four segments, and half subscribe to special pay TV services. This group also has the highest incidence of watching videotaped material at home.

This group also reports a high degree of openness to advertising. Seventy-two percent report enjoying certain TV ads, and sixty percent report enjoy certain magazine ads. Nearly two thirds, (sixty-four percent), disagreed with the statement that they "never buy products based on advertising."

The members of this group are second to the Catalog Shoppers in their use of discount and specialty stores; fifty-nine percent made one or more mail order purchases in the past year. This group's members tend to agree with the statement that "it is dangerous to use credit cards over the telephone." However, more than one third had ordered one or more products using an 800 number in the past year.

Ownership of items associated with leisure activities characterizes the Trendy Spenders. These items include: 35mm cameras, tennis rackets, and running shoes. They also report high levels of out-of-home entertainment including movies, sports events, and vacations.

Cluster 4 — High Rollers

This segment is predominantly male (60%), and tends to be younger (43% are 21-34). These members tend to have the lowest education level of the four segments, and tend to work more than sixty hours per week.

The High Rollers report little reading and few indoor activities in their leisure time, and generally pursue a variety of participatory sports.

This segment is distinguished by members' strong willingness to take risks in leisure activities and to spend. They also express strong openness to advertising. The High Rollers strongly agree that if they want something they'd buy it even if it cost more than they liked (73%). Seventy-nine percent said that it is fun to buy things, while 64% said they will pay more for convenience. They are less likely than other groups to seek activities that are free.

The High Rollers media habits include daily newspapers and TV. This group has the highest incidence of subscription to special pay TV, and 60% subscribe to cable TV. While this group enjoys certain TV and print advertising, 66% report that they never buy based on advertising.

A comparison of the segmentation techniques

The two segmentation schemes will be compared on the basis of demographic distinctions, attitudinal differentiation between the clusters, and media and purchasing/shopping behavior. These criteria are chosen in light of the assumption that to be useful, a segmentation scheme must clearly identify the members of each segment, and clearly identify the means by which they are most likely to be reached.

Interestingly, there are similarities in the two schemes. Both resulted in four consumer groups. The Risk Takers of the factor analysis (FA) scheme are somewhat like the High Rollers of the correspondence analysis (CA) scheme. Both are characterized by risk-taking behavior and participation in sports during their leisure time. Neither are characterized by thriftiness or control of spending. There are demographic and media differences which are more distinct for the CA segmentation.

There are also similarities between the Indulgent Homebodies in the FA scheme and the Catalogue Shoppers of the CA scheme. Both groups are predominantly female and have low interest in activities involving physical effort, and both groups are heavy consumers of broadcast media. However, the Catalogue Shoppers are more distinct in their shopping behavior and in their interest in bargain hunting.

The Thrifty Ad Watchers and the Penny Pinchers share lower incomes, sedentary activities and susceptibility to advertising. Both groups tend toward conservative attitudes in spending and risk-taking. The demographic characteristics of this segment are more dramatically portrayed in the correspondence analysis segmentation.

The Fast Track segment (FA) and the Trendy Spenders (CA) share their love of active leisure time. Both groups report a high degree of involvement in physical exercise. There are some differences in their media habits. The Fast Track have a lower overall involvement with both broadcast and print media and are less distinct in their spending patterns than the Trendy Spenders. Generally, the attitudinal distinction of the Fast Track is clearer than that of the Trendy Spenders. However, the demographic, media and shopping profiles of the Trendy Spenders are more distinctive than those of the Fast Track.

Chi-square measures for each of the techniques

The correspondence analysis method generally yielded a greater distinction between groups on behavioral and demographic variables than did the factor analysis method. Seven of the contingency tables with cluster membership and leisure activity participation had chi-squares significant at the .10 level for the correspondence analysis-based segmentation, while six of the same tables contained

significant differences in the factor analytic approach. There were six tables showing significant differences for the categories describing where and with whom respondents spent leisure time in the correspondence analysis-based solution, and five in the factor analytic solution.

Larger differences between techniques were shown in the areas of frequency of participation in specific leisure activities and types of equipment owned. Four of the contingency tables for frequency of participation showed significant differences for the CA segmentation, while only one of the FA contingency tables showed a significant difference among the groups. In the area of equipment owned, ten of the CA tables were significant while only six of the FA based tables showed significant differences.

In the areas of club memberships and purchase behavior, three and five tables respectively showed significant differences among the CA based groups, compared with two and three for the FA based segments.

Age, income and gender were significantly different among the segments for both of the segmentation schemes. However, the observed chi-squares for age and income were much higher for the correspondence analysis based solution. Additionally, education, marital status and employment status were areas in which the chi-squares for the contingency tables for the groups based on correspondence analysis were statistically significant.

Both segmentation schemes produced interesting and useful consumer groups. However, the scheme produced using correspondence analysis was clearer with respect to several critical areas. The demographic distinctions and the media and spending habits of each of the four segments were clearer than those of the scheme based only on principal component scores. This difference is valuable in light of the usual purposes for market segmentation. These groups were more clearly identified and distinguished from one another in settings in which the attitudinal variables are not available, such as targeted marketing and advertising efforts. They were also more accessible because the differences in their media and spending habits were more distinct than in the factor analysis scheme.

Demographic differences between clusters were generally more distinct in the correspondence analysis segmentation scheme. Figures 2 and 2b show that age, income and education were more differentiated among the CA segments than they are in the factor analysis segmentation scheme. Figures 3 and 3b show the difference between schemes in the distinctions between clusters for openness to consumer spending. The distinctions between schemes for risk taking are shown in figures 4 and 4b. Shopping channels and openness to advertising for each of the segments and both segmentation schemes are shown in figures 5 and 5b, and 6 and 6b respectively. The final set of figures, 7 and 7b, shows distinctions in media use by segment for each of the segmentation methods.

Figures 8, 9, and 10 are simple correspondence analysis maps showing the eight clusters and their relative positions with respect to the attitudinal, behavioral and demographic variables. The map of the attitudinal variables shows that both approaches produce attitudinally distinct groups. With the exception of the High Rollers, the segments defined using correspondence analysis are much more distinct on the map of the behavioral variables. A similar pattern is true of the demographic variables.

In conclusion, given the flexible data requirements and the clarity of interpretation, market segmentation based on correspondence analysis may produce more useful segmentation schemes than traditional segmentation based on attitudinal data and factor analysis prior to clustering.

REFERENCES

Greenacre, M. (1984). *Theory and applications of correspondence analysis.* Academic Press, New York.

Greenacre, M. (1978). "Some objective methods of graphical display of a data matrix." Special report, Department of Statistics and Operations Research, University of South Africa (November).

Gorsuch, R. (1983). *Factor Analysis.* Lawrence Earlbaum, Hillsdale, N.J.

Hoffman, D. & Franke, G. (1986). "Correspondence Analysis: Graphical representation of categorical data in marketing research." *Journal of Marketing Research*, 23, 213-227.

Lebart, L., Morineau, A. & Warwick, K. (1984). *Multivariate descriptive statistical analysis.* John Wiley & Sons, New York.

Lindeman, R., Merenda, P., & Gold, R. (1980). *Introduction to bivariate and multivariate analysis.* Scott Foresman, Glenview, Illinois.

"Technical Report: A-108 Cubic Clustering Criterion." (1983). The SAS Institute, Cary, North Carolina.

Principal Components

Factor 1 — Risk	
Activities/dangerous	.79
Take risks/leisure	.79
Activities/effort	.59
Factor 2 — Thrift	
Bargain hunter	.73
Thrifty	.48
Activities/free	.43
Make things	.32
Pay for conven.	-.47
Factor 3 — Indulgence	
Fun/buy things	.69
1st try new things	.55
Time for me	.54
Want/buy even cost more	.54
Factor 4 — Conservatism	
Dangerous/phone/card	.77
Never buy on ads	.45
I'm stubborn	.43
Try on clothes	.40
Make thing	.32
Factor 5 — Advertising	
Enjoy TV commercials	.67
Enjoy magazine ads	.65
Factor 6 — Control	
Control over spending	.73
Work is rewarding	.52
Factor 7 — Busy	
Rarely time w/family	.71
Never worry/bills	.46

Figure 1

Means of Principal Components by Consumer Segment

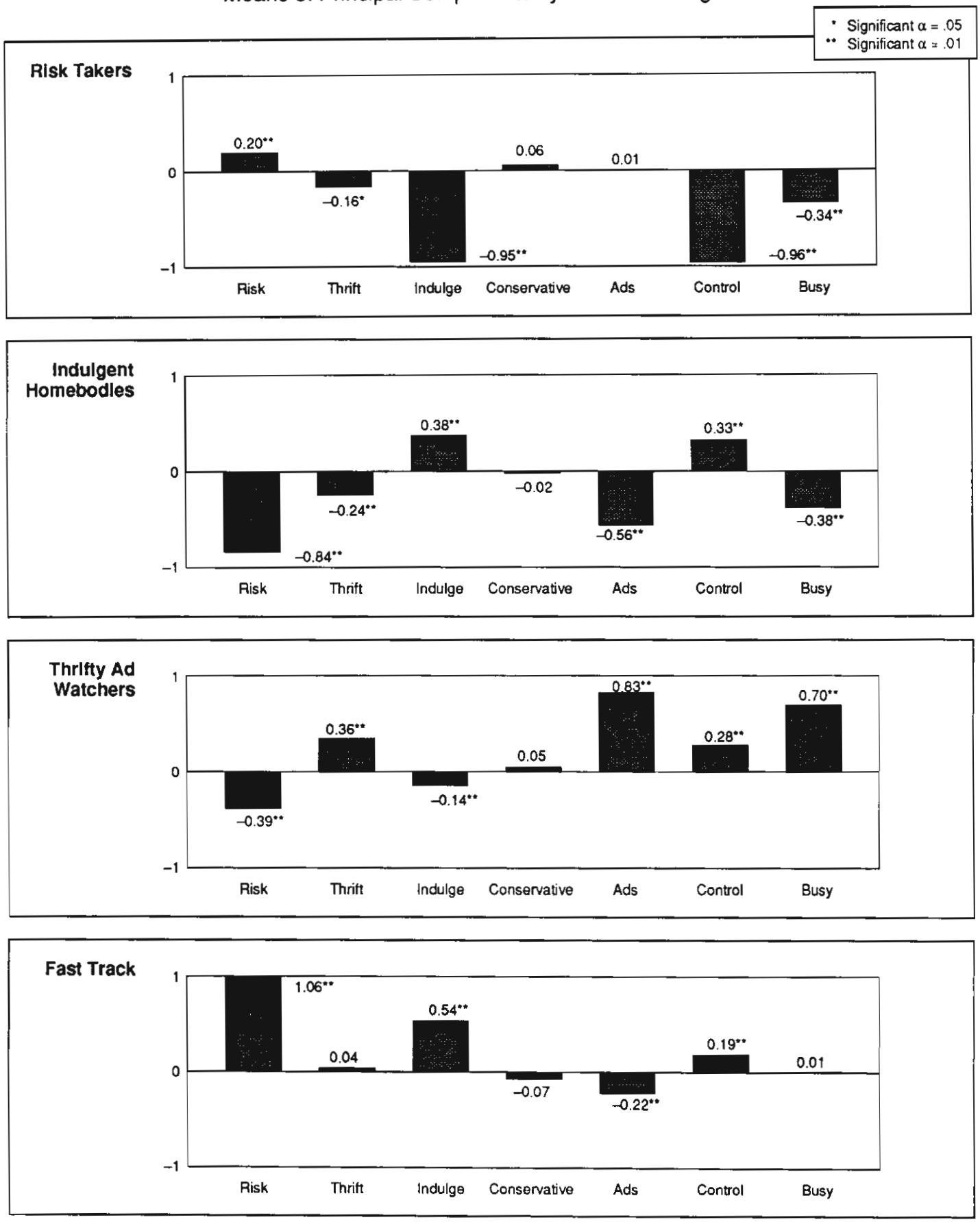
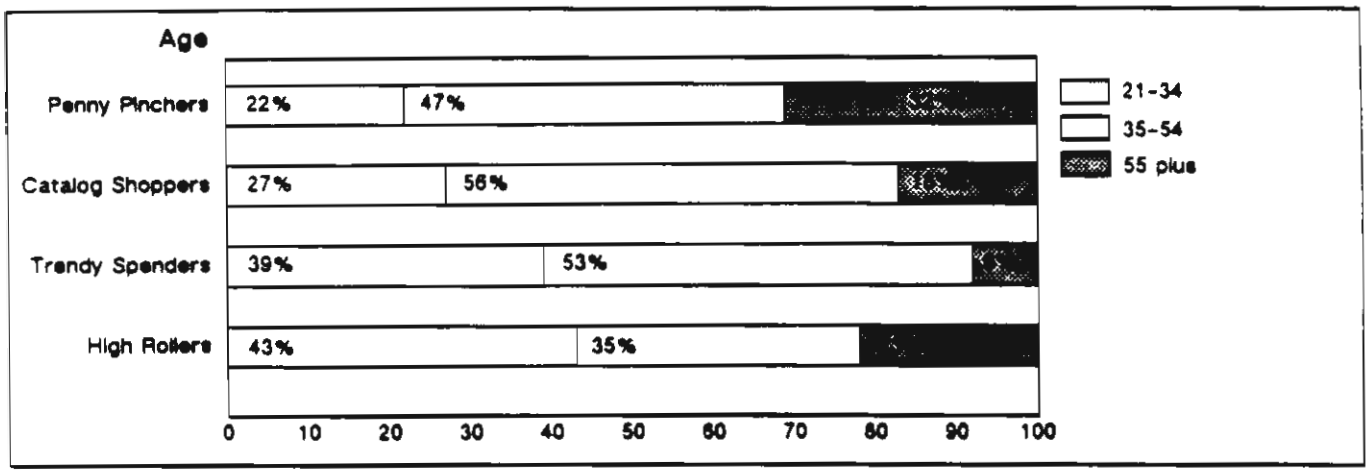
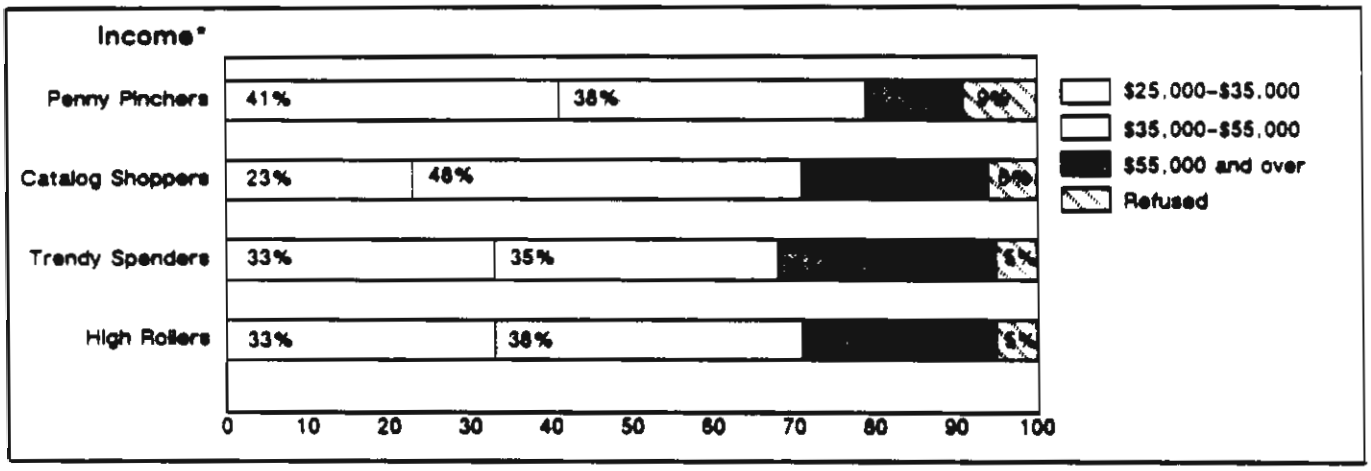


Figure 2A

Consumer Segments by Age, Income and Education





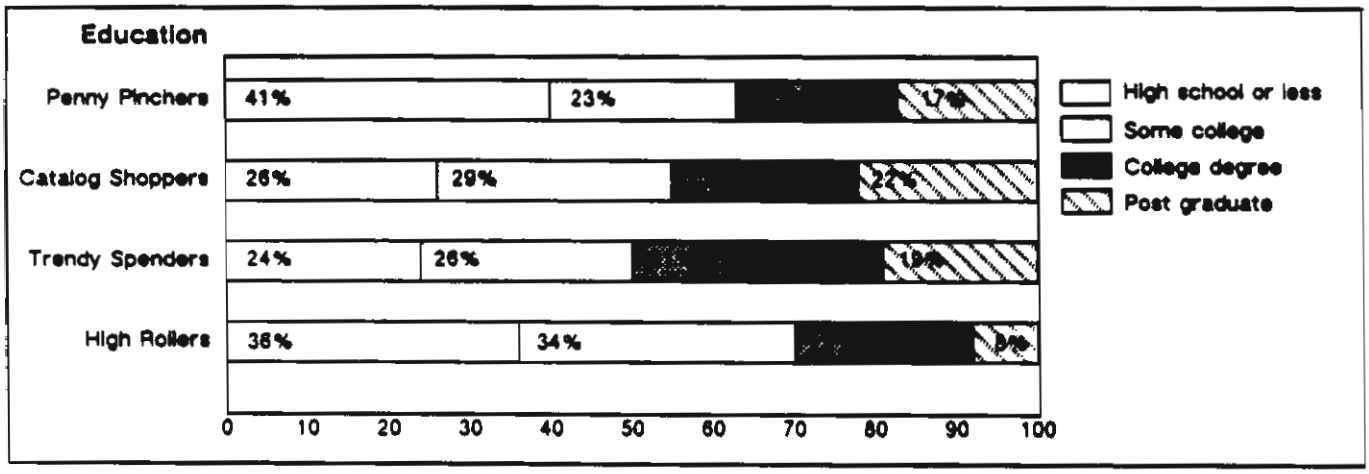
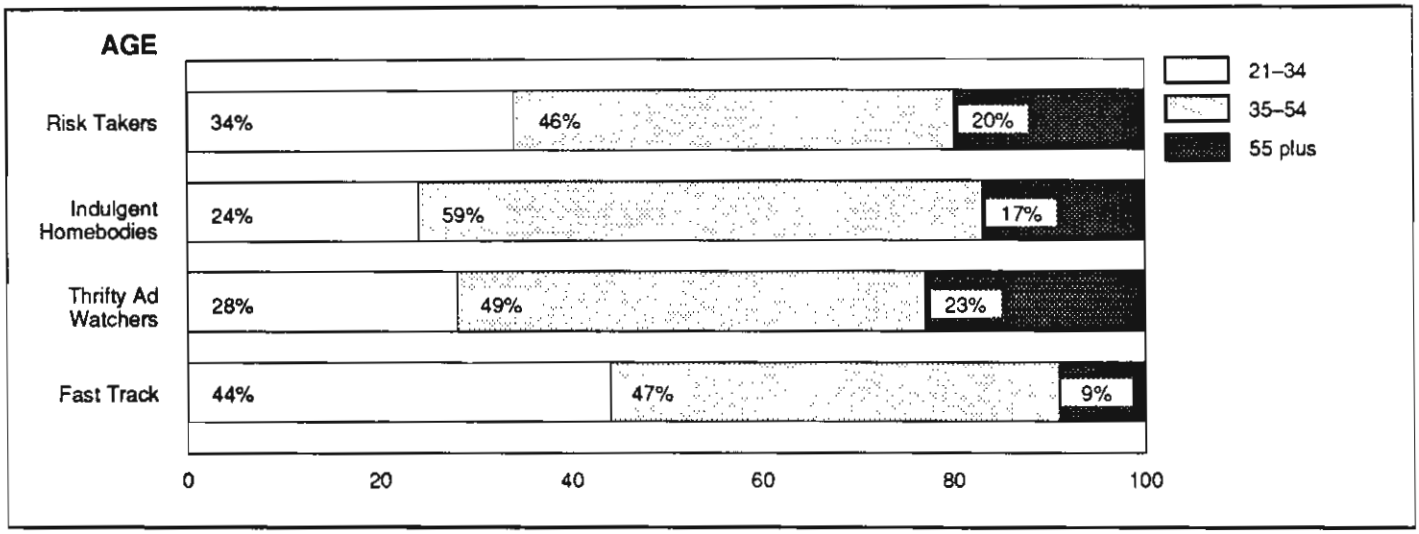
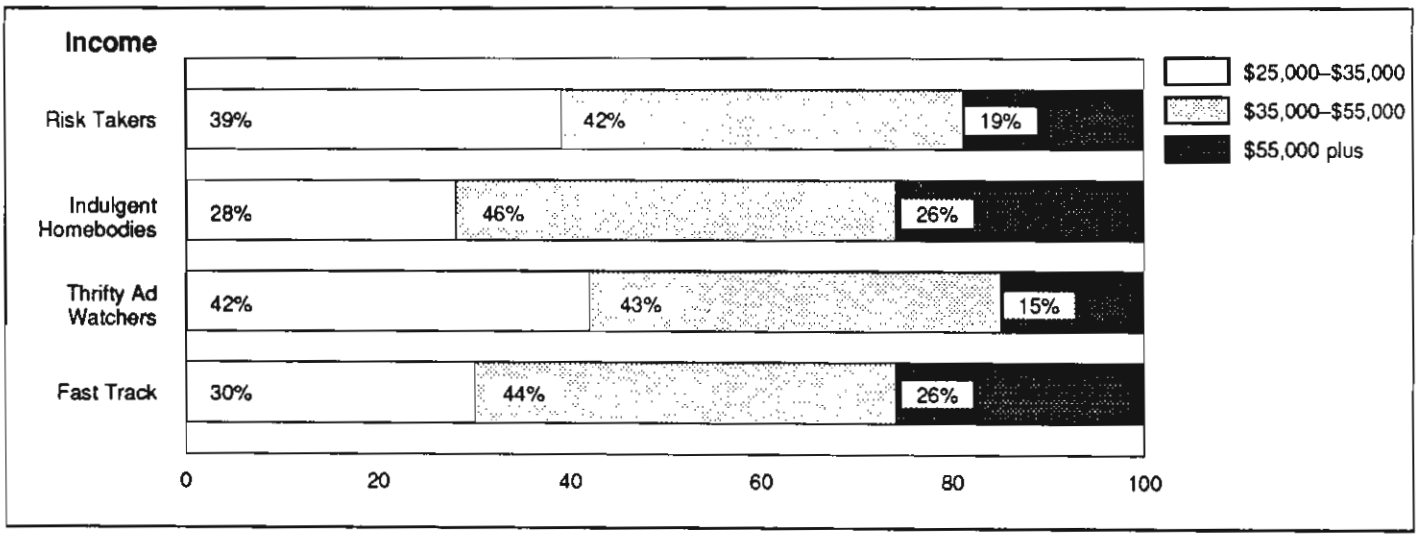


Figure 2B

Consumer Segments by Age, Income and Education





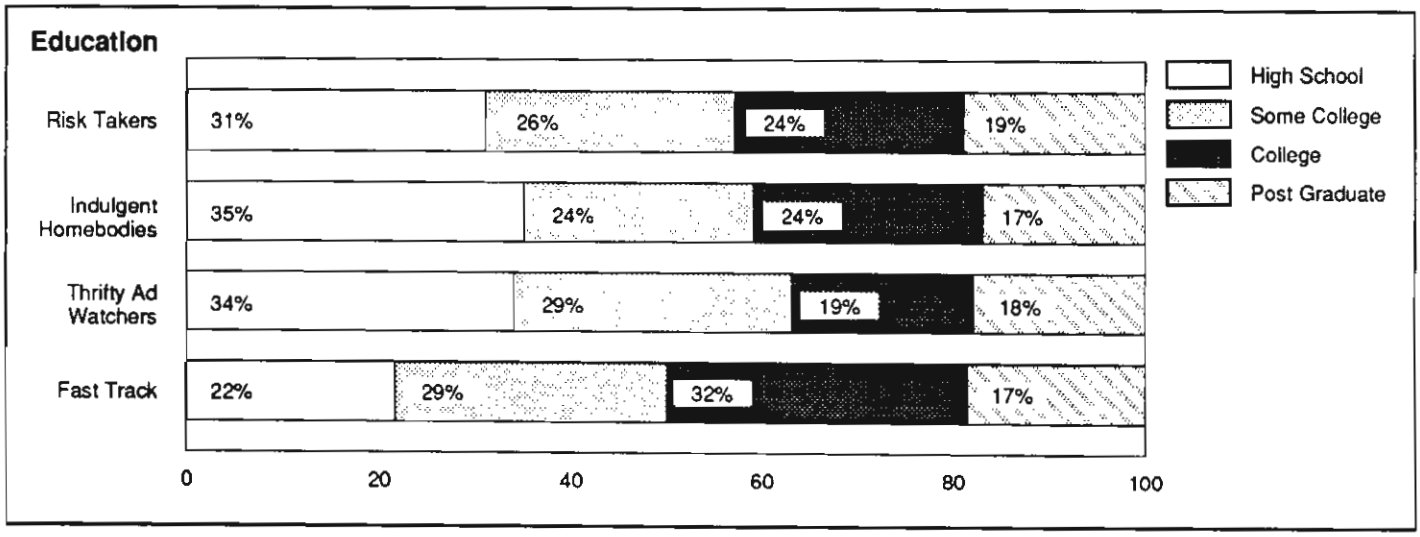


Figure 3A

Openness to Spending by Consumer Segment

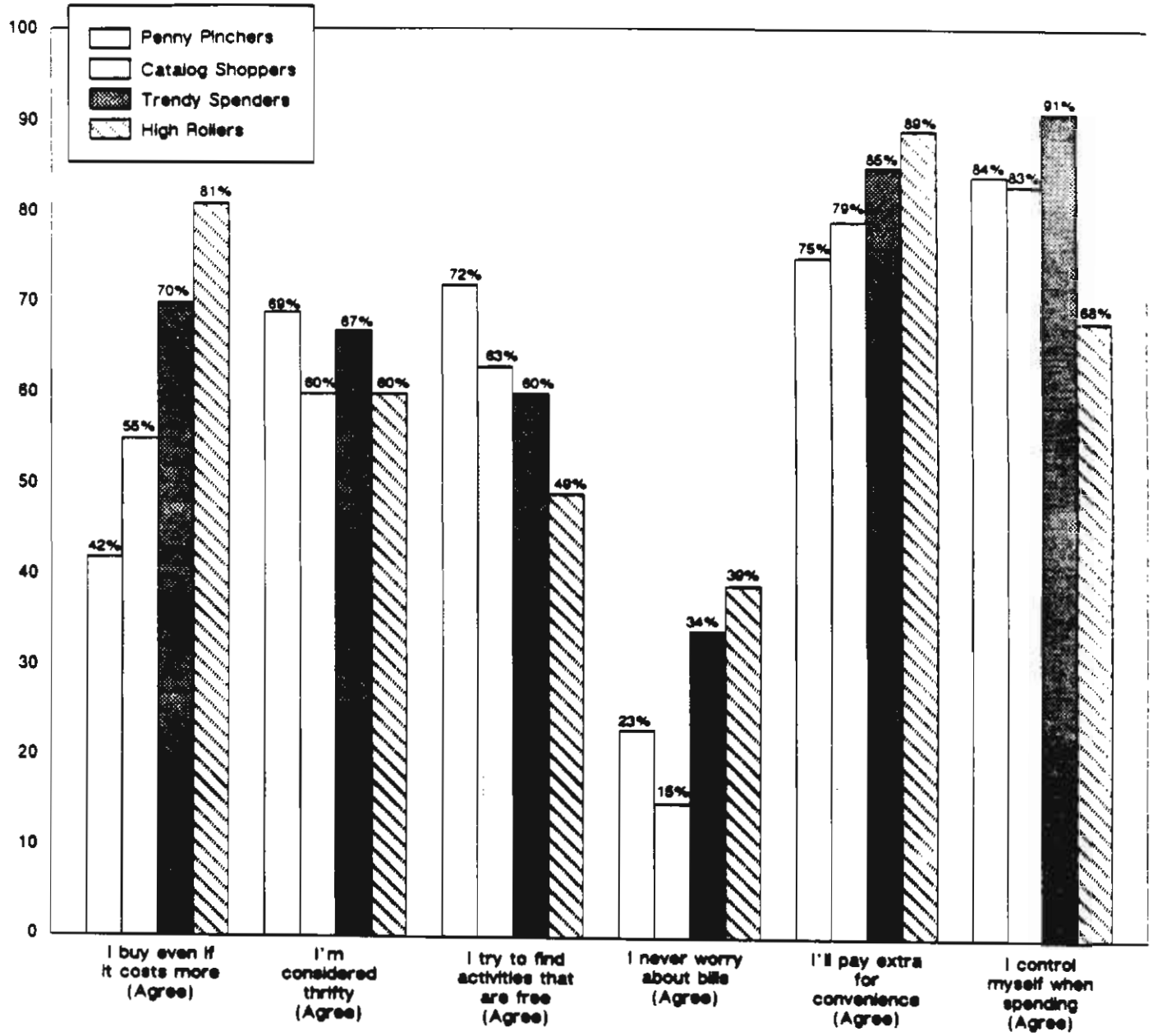


Figure 3B

Openness to Spending by Consumer Segment

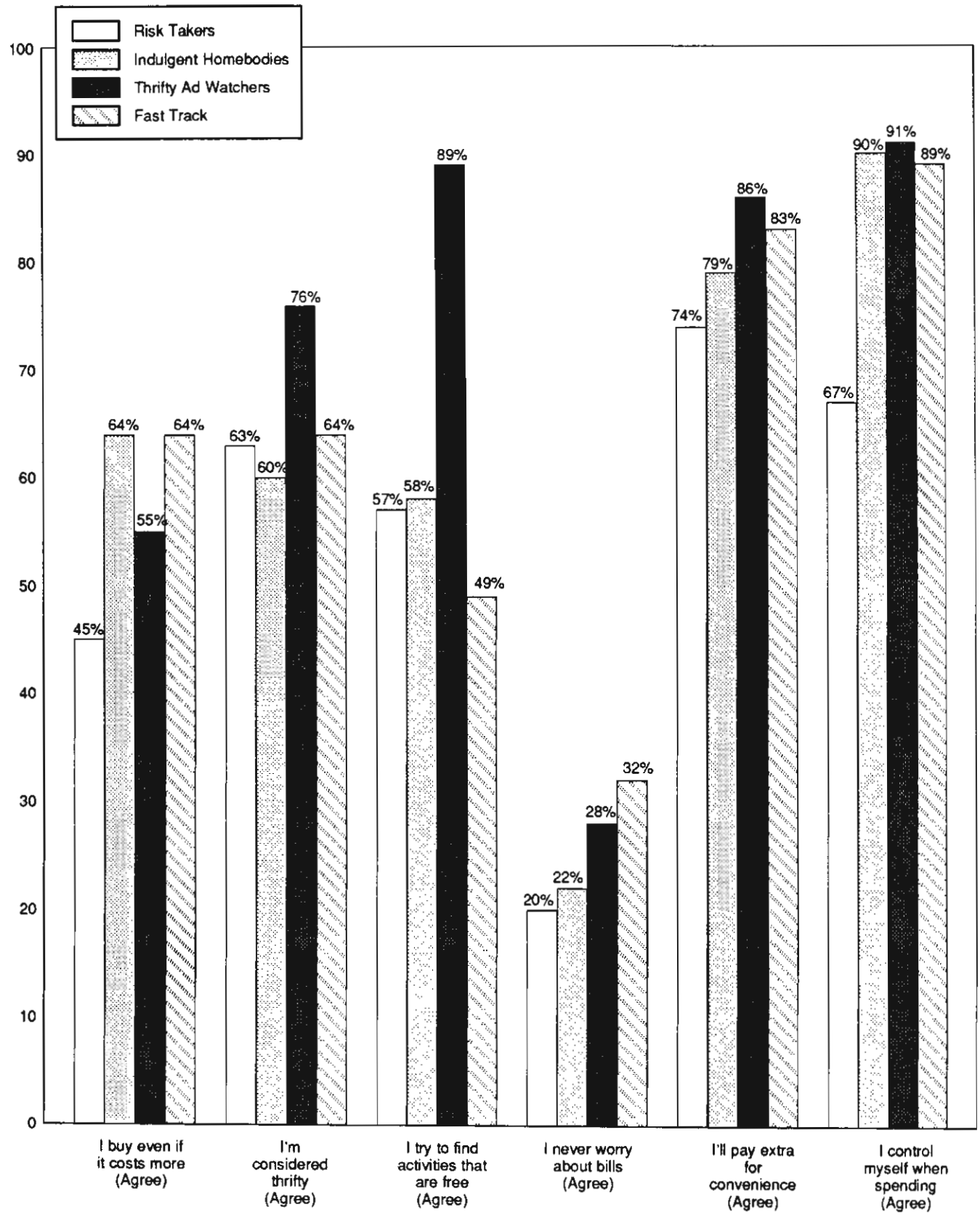


Figure 4A

Openness to Risk-Taking by Consumer Segment

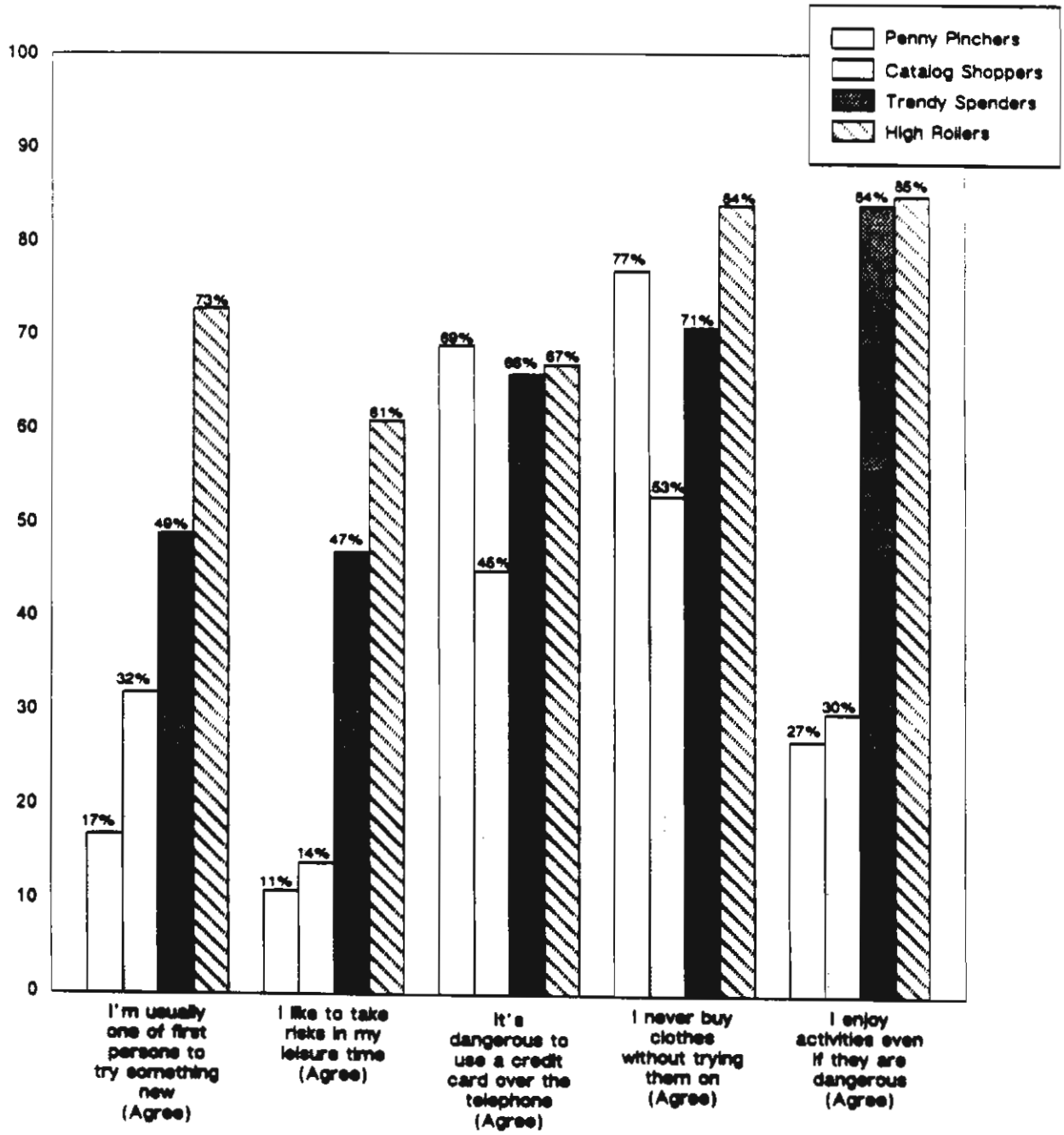


Figure 4B

Openness to Risk-Taking by Consumer Segment

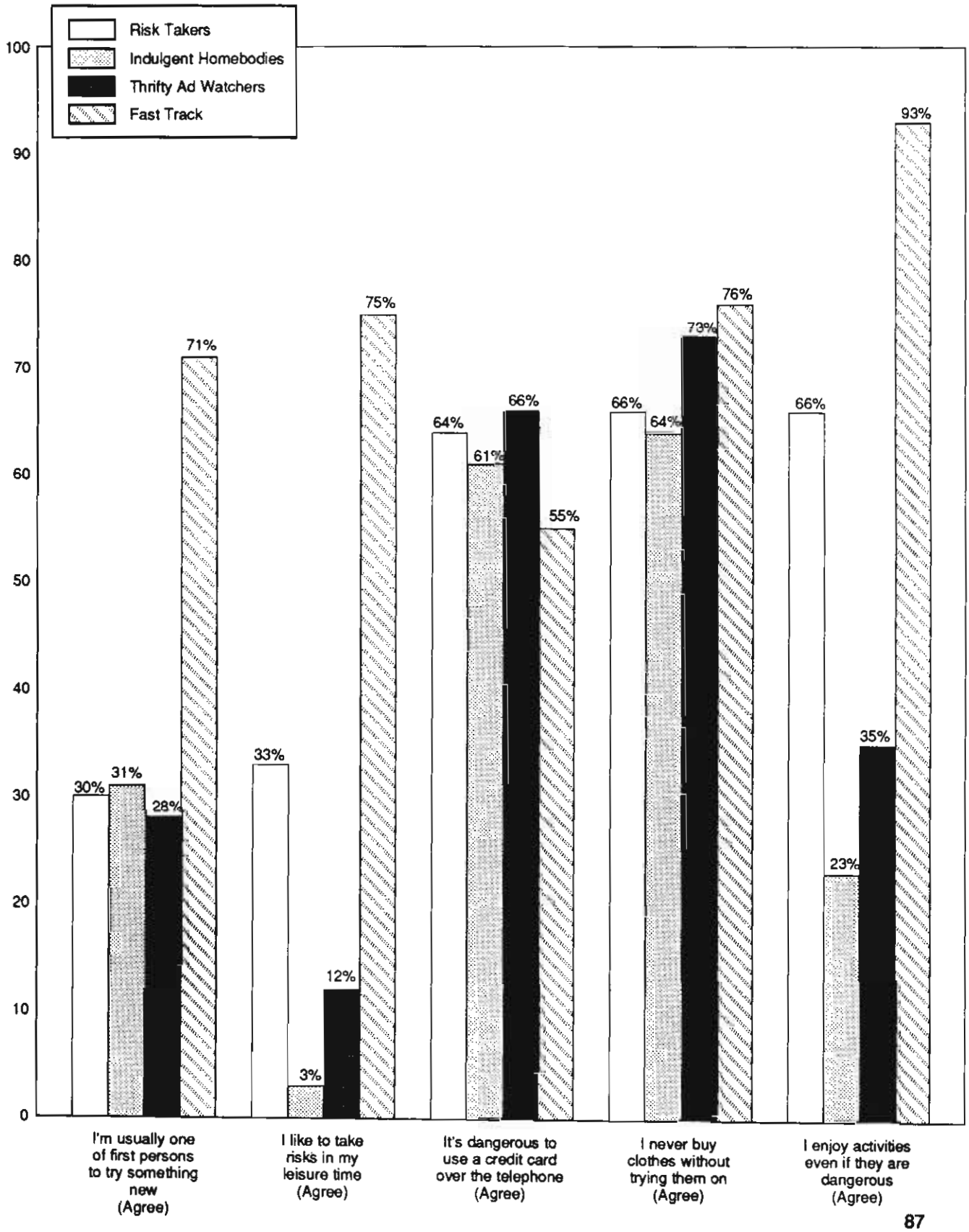


Figure 5A

Shopping Channels Used in Past Month by Consumer Segment

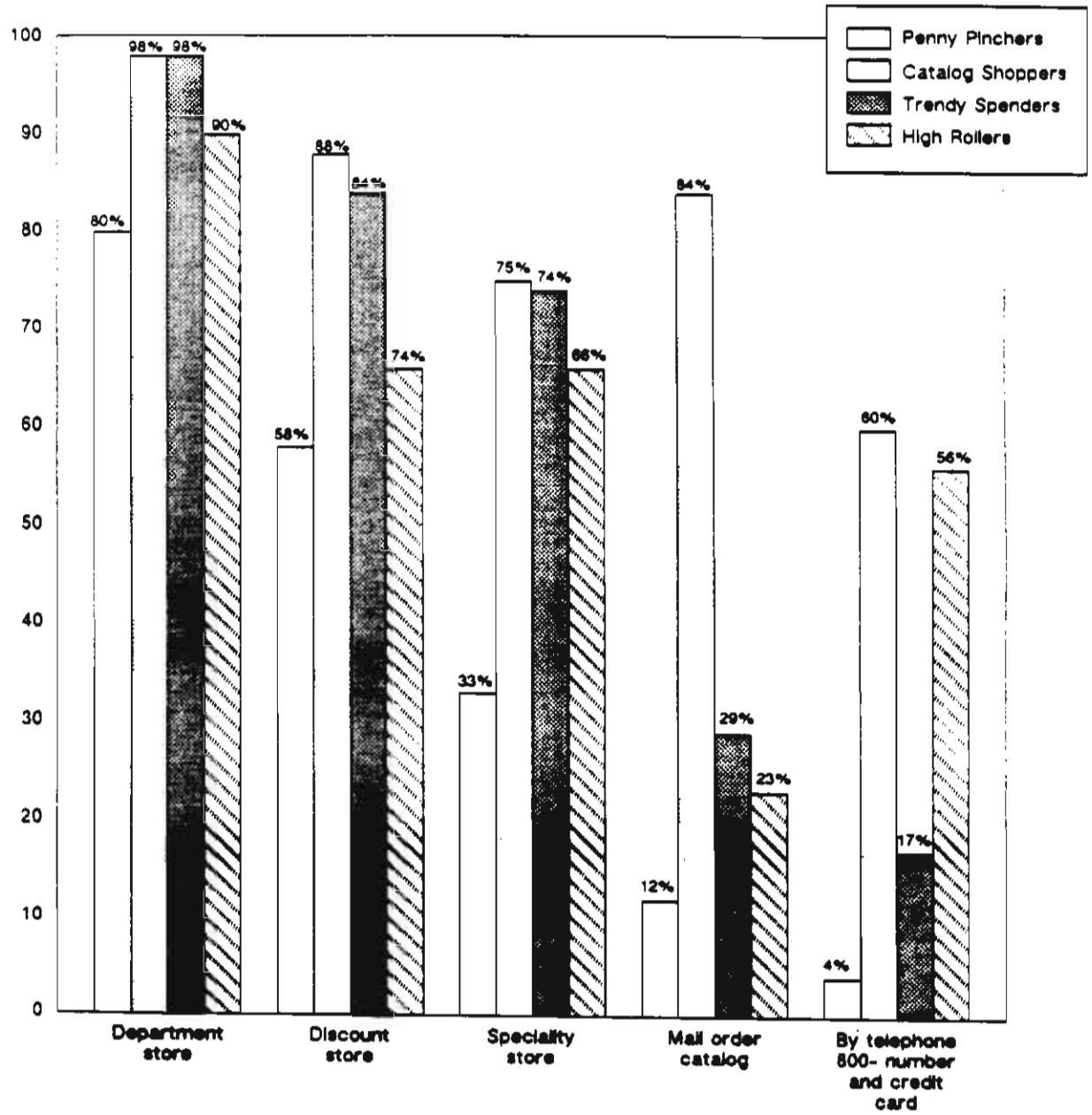


Figure 5B

Shopping Channels Used in Past Month by Consumer Segment

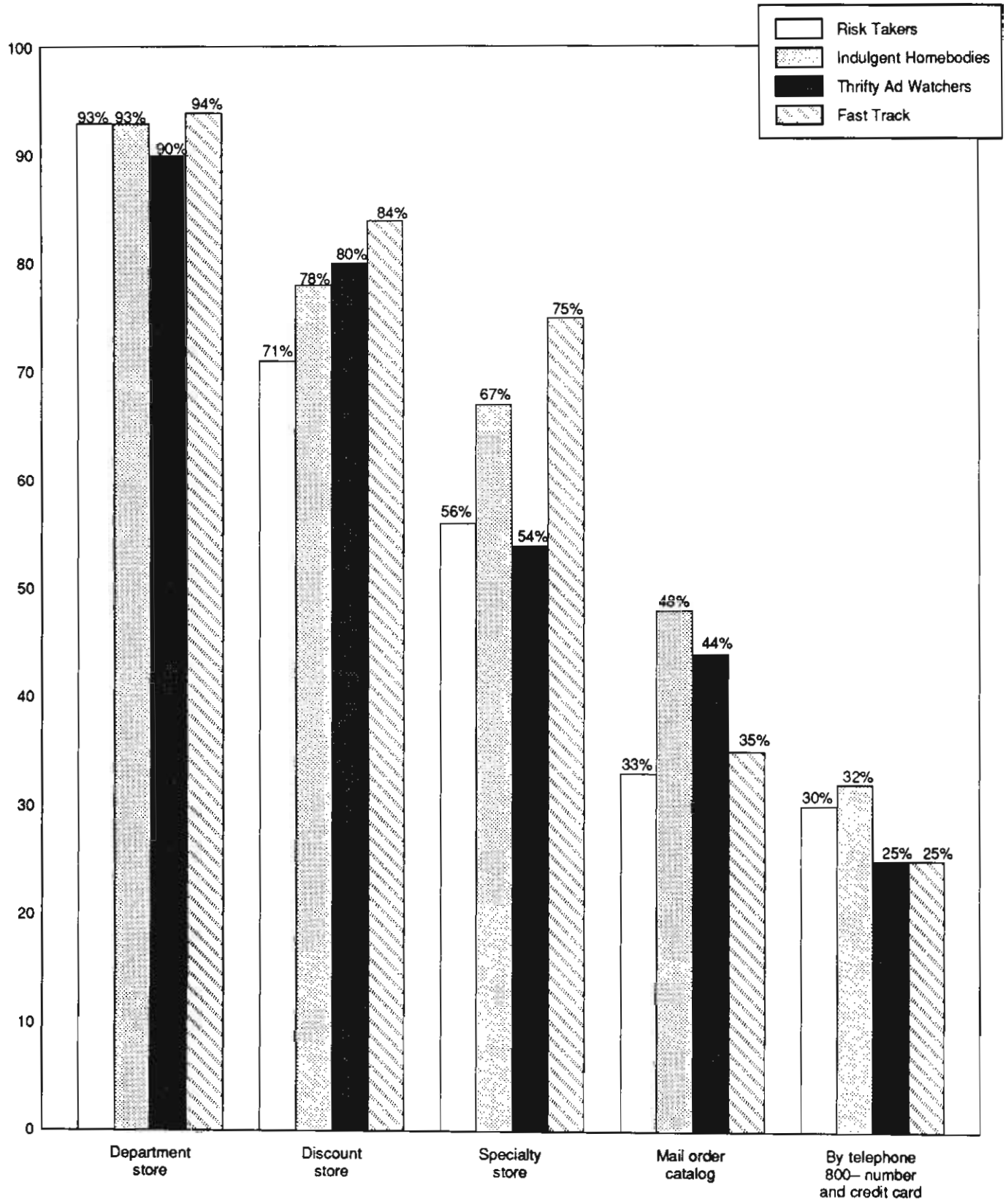


Figure 6A

Openness to Advertising by Consumer Segment

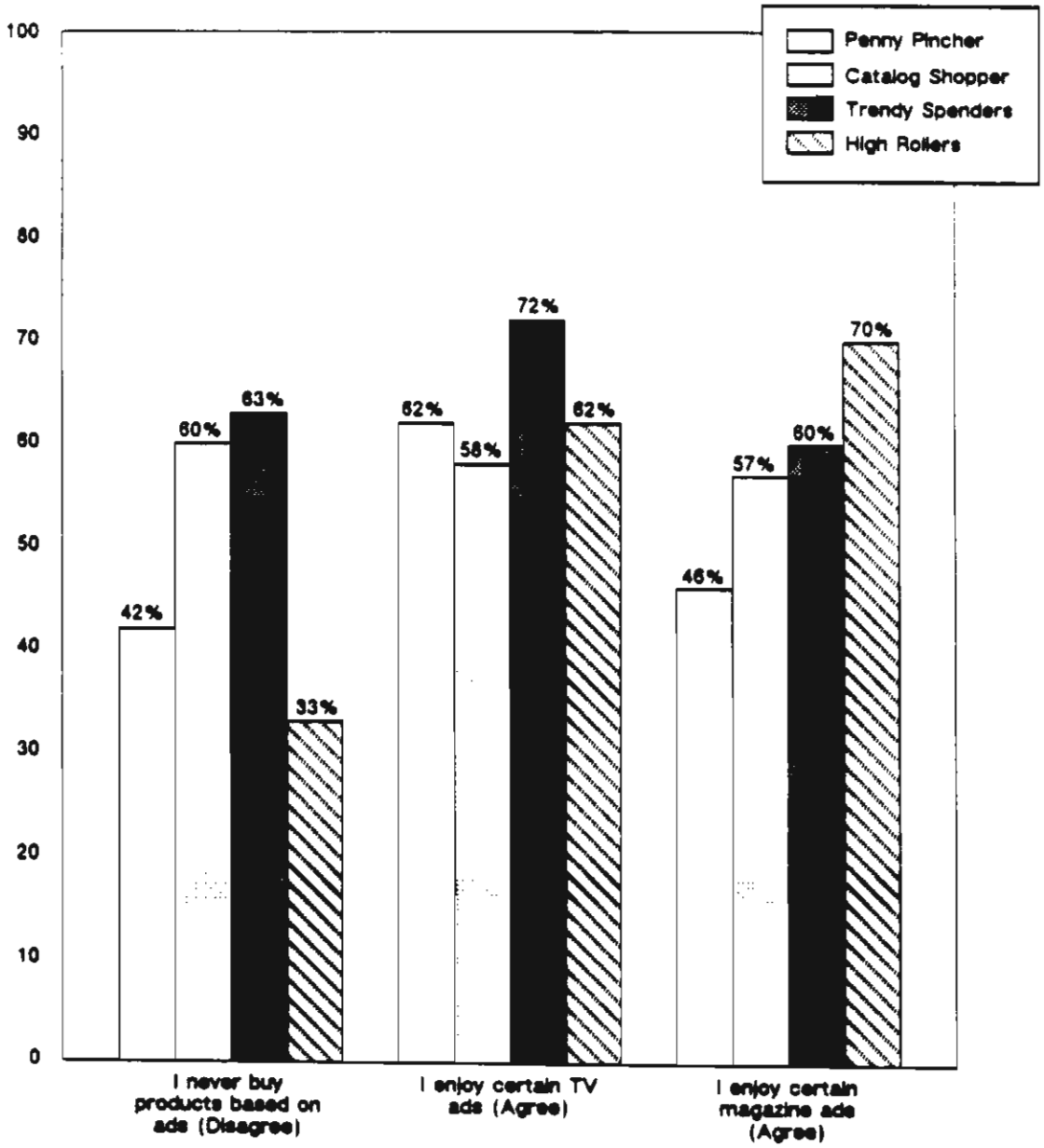


Figure 6B

Openness to Advertising by Consumer Segment

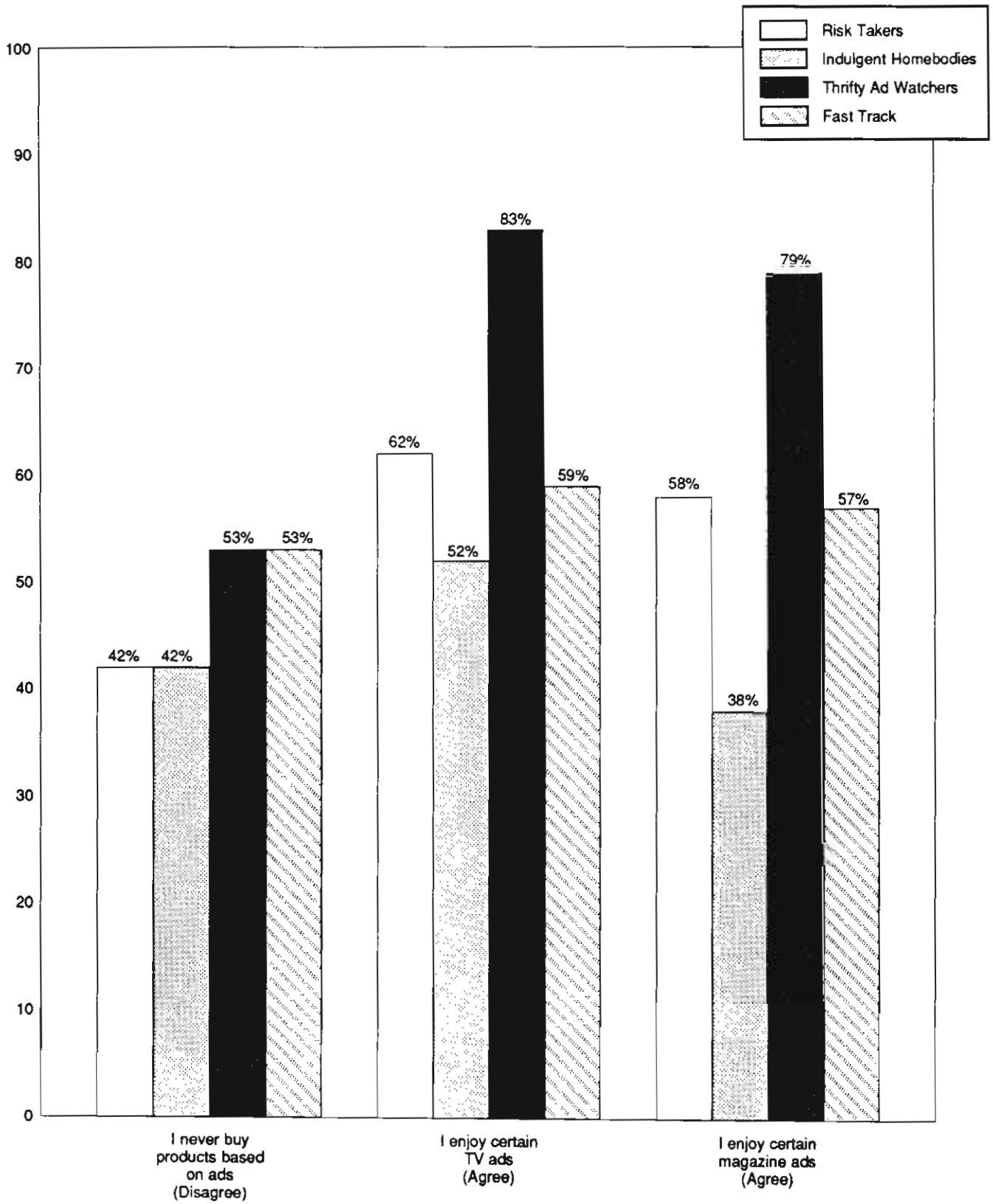


Figure 7A

Media Use by Consumer Segment

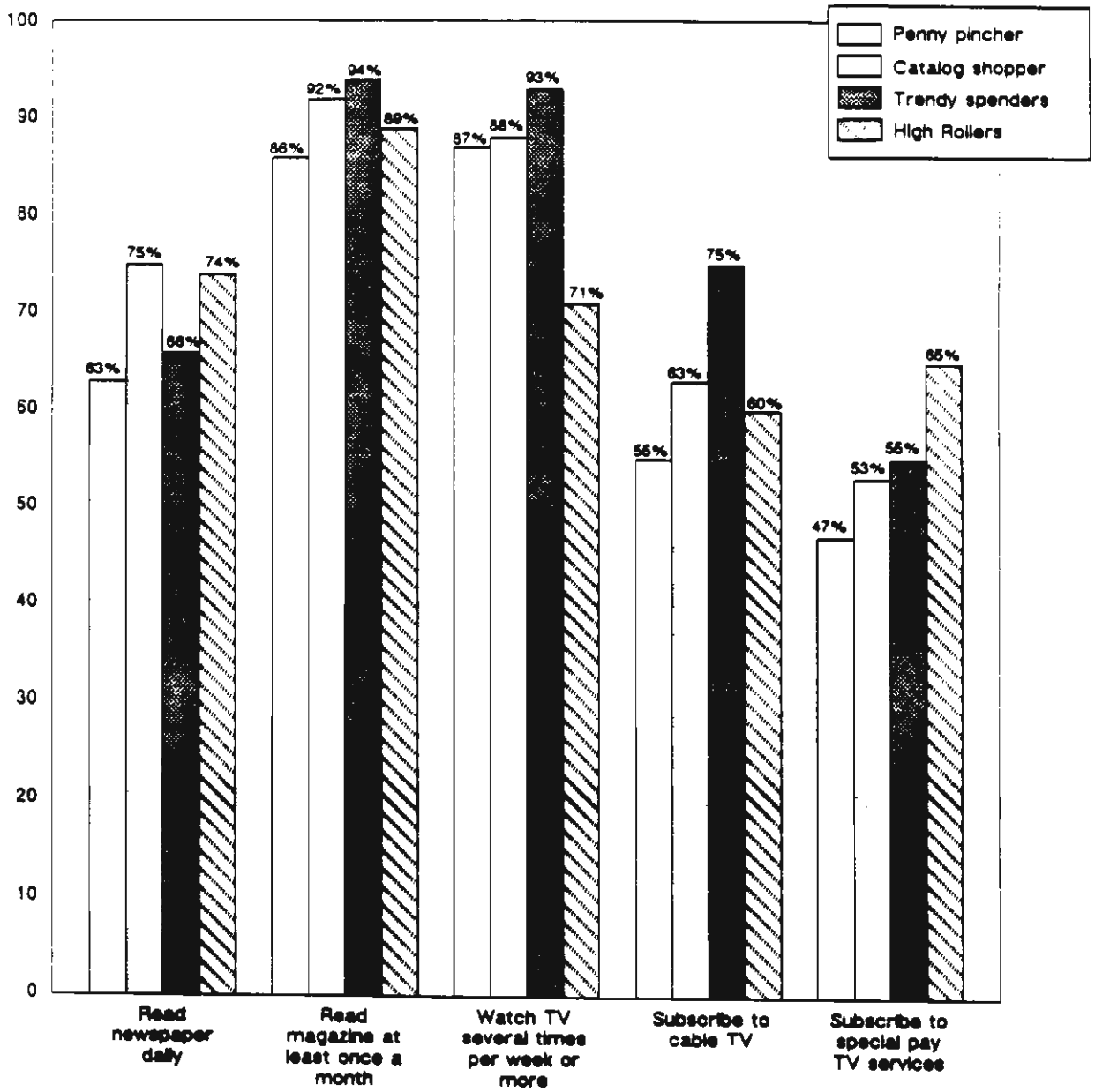


Figure 7B

Media Use by Consumer Segment

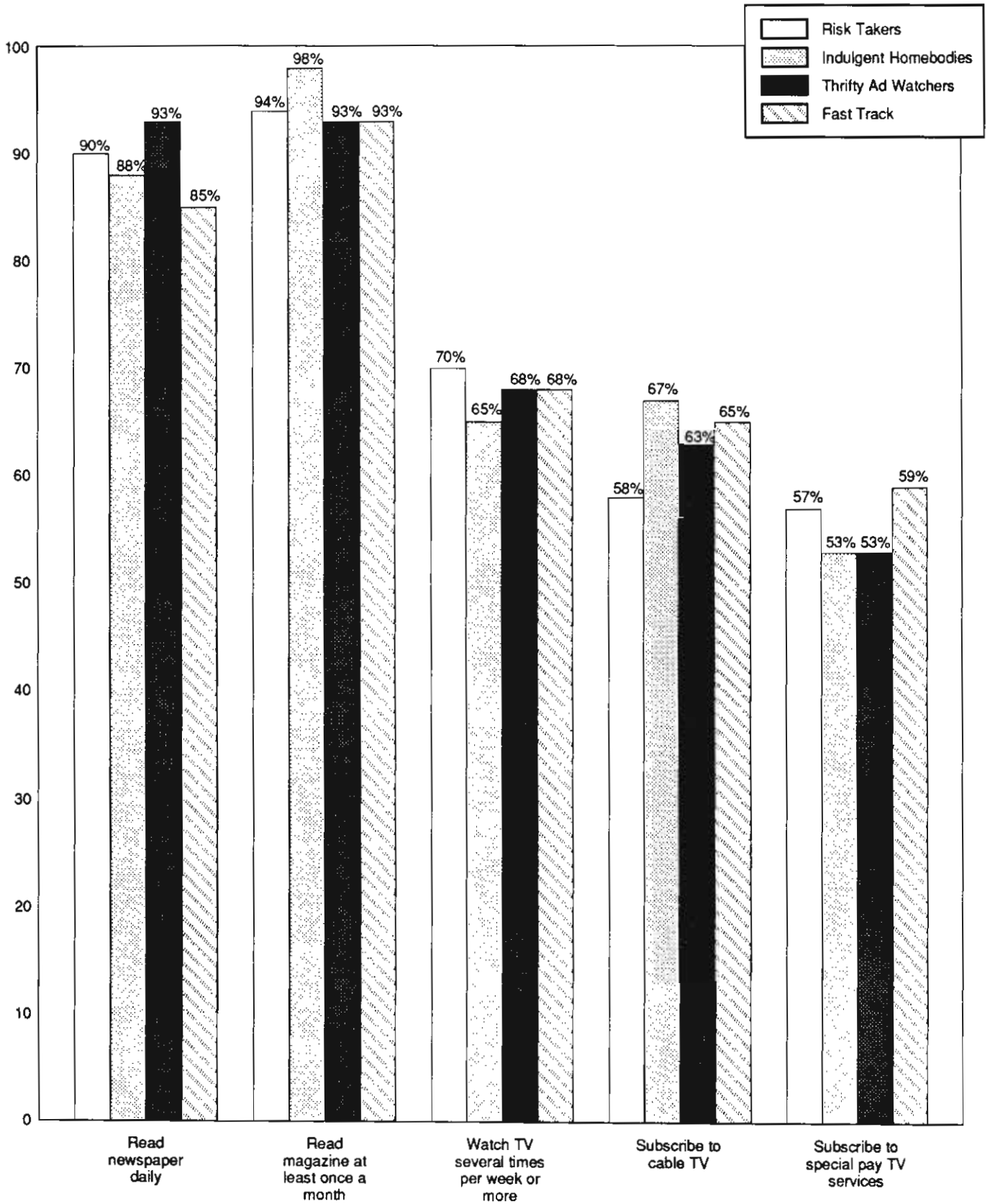
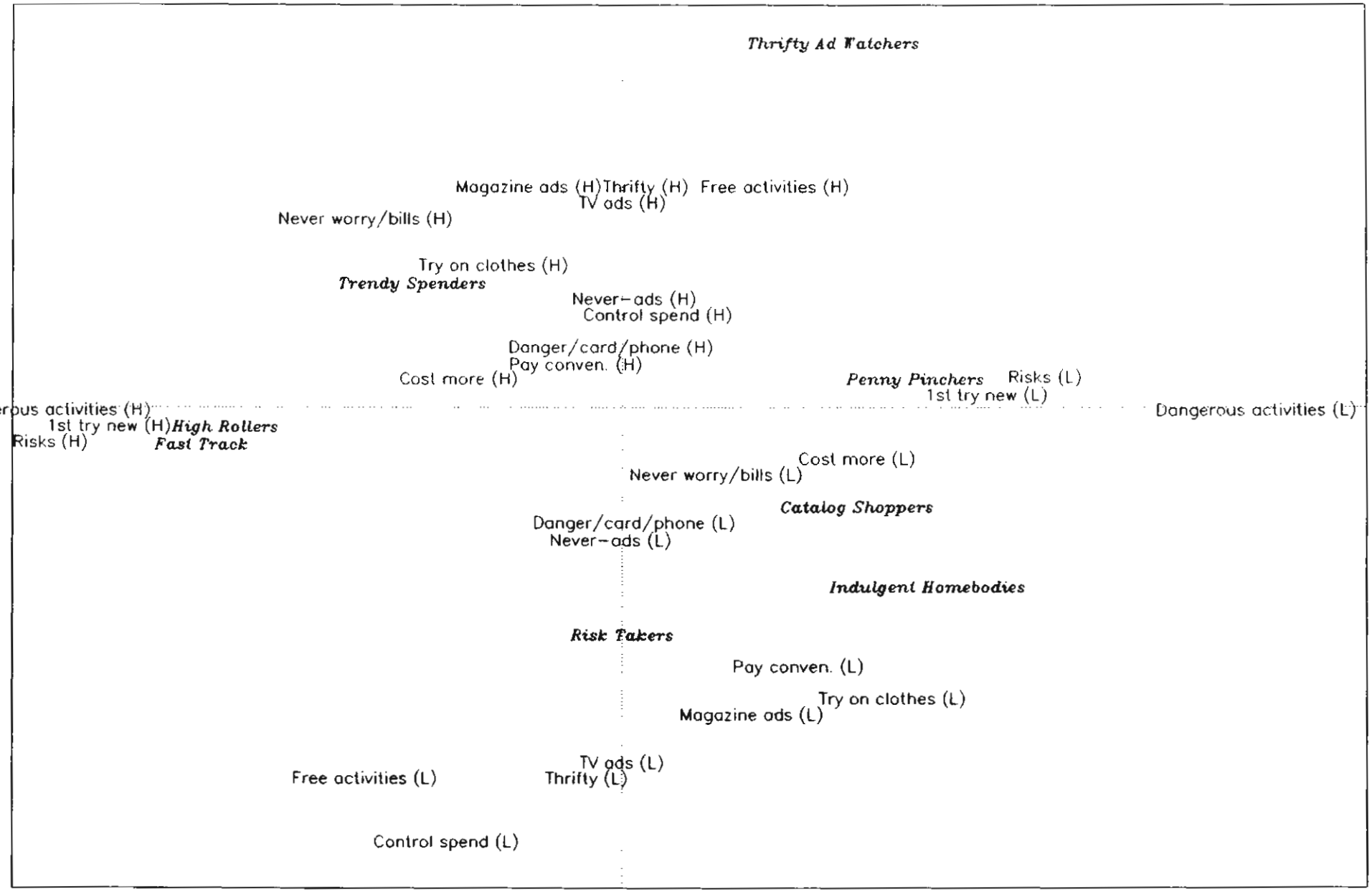


Figure 8

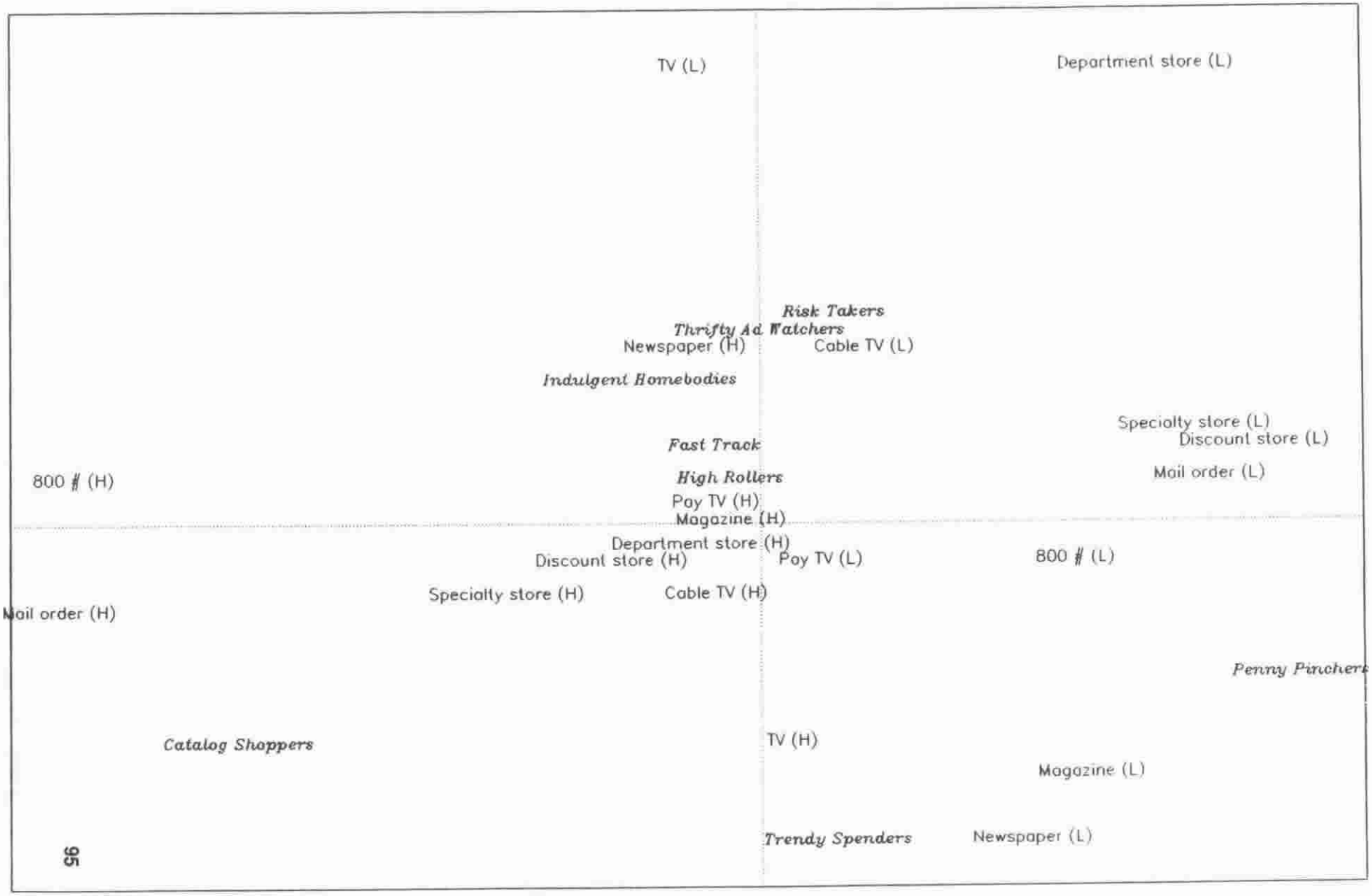
Attitude Variables by Cluster



(H) —High
(L) —Low

Figure 9

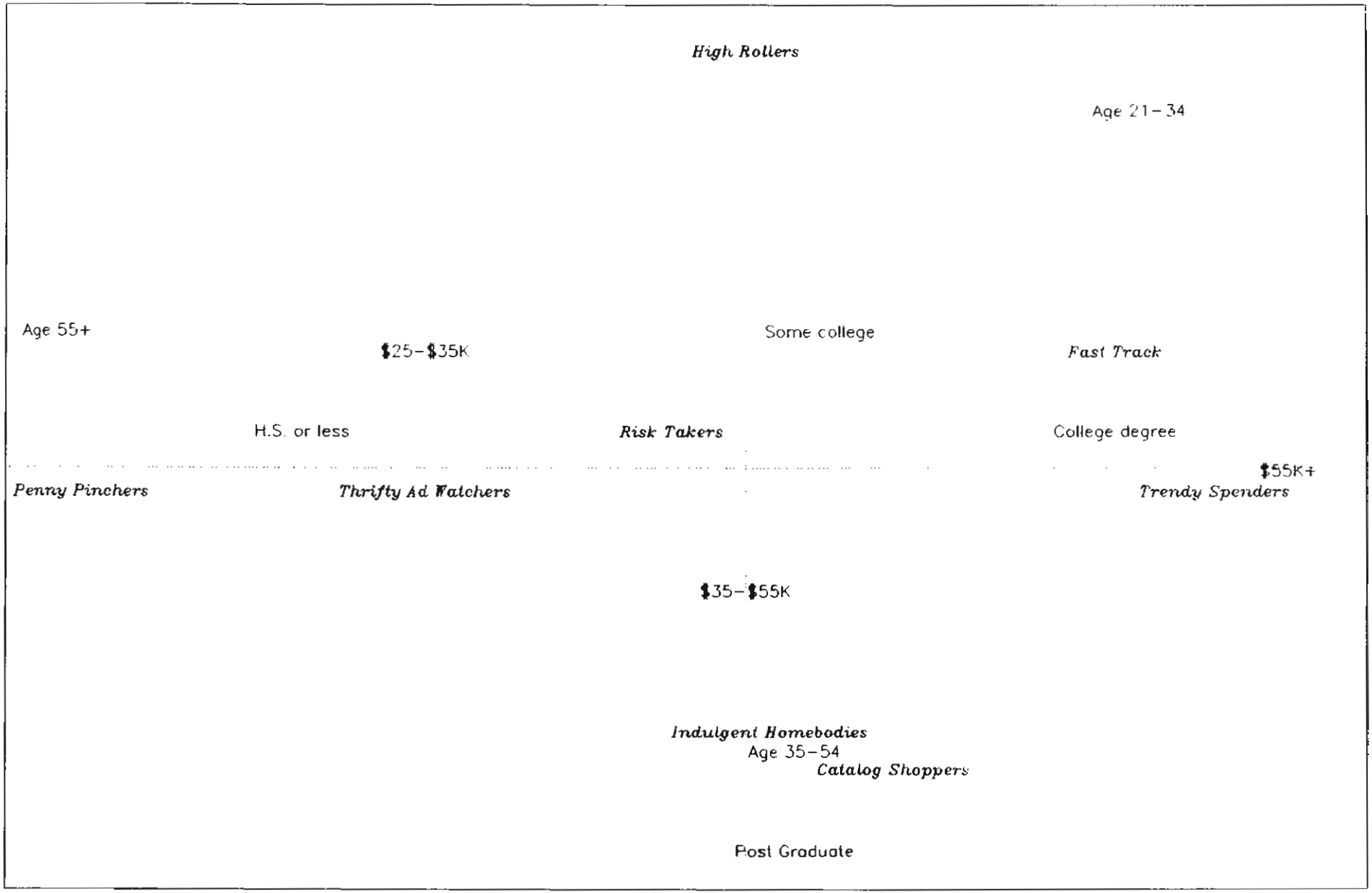
Behavior Variables by Cluster



(H) —High
(L) —Low

Figure 10

Demographic Variables by Cluster



MAKING CLUSTER-BASED NEEDS SEGMENTS ACTIONABLE

Thomas L. Pilon and ***Karlan J. Witt***
IntelliQuest, Inc.

An effective approach to market segmentation is to determine segments on the basis of a clustering of respondents on a set of variables which measure needs. Once these cluster-based needs segments are formed, it is necessary to derive an operational definition of these segments in order to make them actionable. This paper will describe and demonstrate alternative techniques for making cluster-based needs segments actionable. As background, a brief review of market segmentation research approaches will be provided.

MARKET SEGMENTATION

Since Wendell Smith's pioneering article in 1956, the concept of market segmentation has received considerable attention in both the literature and in practice. Landmark manuscripts include Frank, Massy & Wind (1972) and Myers & Tauber (1977). Haley's articles on "benefit segmentation" in 1968 and 1984 and Wind's overview article, which was the lead article in the Journal of Marketing Research's special issue on market segmentation in 1978, are also paramount.

Although there are numerous descriptions of "market segmentation" and of "a market segment," the following are among those most frequently cited. "Market segmentation is the subdividing of a market into distinct subsets of customers, where any subset may conceivably be selected as a market target with a distinct marketing mix" (Kotler). "Market segmentation is concerned with individual or intergroup differences in response to market-mix variables" (Green, Tull & Albaum). A market segment is "a group of customers anywhere along the distribution chain who have common needs and values--who will respond similarly to the company's offerings and who are large enough to be strategically important to their business" (Coles & Culley).

METHODS FOR FORMING SEGMENTS

Generally, there are two methods for forming segments: *a priori* segmentation and *post hoc* segmentation. *Post hoc* segmentation is also referred to as "response-based" segmentation (Myers & Tauber). Some researchers often refer to a third, "hybrid," approach in which the two above mentioned approaches are applied in one sequence or the other or in some recursive combination (Green, Tull & Albaum; Wind). It is beyond the scope of this paper to discuss specific other approaches such as "flexible segmentation," "component segmentation" (Wind), or "occasion-based segmentation" (Greenberg & McDonald). It should be noted, however, that Assael and Roscoe provide an interesting alternative classification of segmentation approaches.

A priori segmentation groups consumers in terms of some factor or factors that are known in advance to be related to purchase. Some examples would be demographics, favorite brand, heavy versus light usage, and media habits. In contrast, *post hoc* segmentation groups consumers according to the similarity of their multivariate profiles regarding such characteristics as product attribute importance or attitudes. Following this, the resulting segments are examined for differences in other characteristics, such as demographics, favorite brand, and heavy versus light usage.

Many researchers have found a particular type of *post hoc* segmentation – cluster-based needs/ benefits segmentation – to be the most useful in a wide variety of circumstances (Greenberg & McDonald; Moriarty & Reibstein).

A FOUR-STEP APPROACH FOR CLUSTER-BASED NEEDS/BENEFITS SEGMENTATION

Cluster-Based Needs/Benefits Segmentation can be described as a four-step procedure:

1. determine Key Buying Factors (KBFs),
2. measure relative importances of KBFs,
3. form homogeneous segments based on relative importance of KBFs,
4. profile segments to make actionable.

Each step will be described below.

1. Determine Key Buying Factors (KBFs)

Typically, industry knowledge, results from previous studies, or exploratory research such as in-depth interviews or focus groups are utilized to generate the list of key buying factors that will be examined in the next step.

2. Measure Relative Importances of KBFs

This step is concerned with the question of quantifying benefit importance. Although there are innumerable approaches, only the most widely used will be discussed here. Direct questioning (on a scale of 1 to 10, how important is ...) is widely used, but has well documented problems (Green, Wind & Jain). A more effective direct approach, termed "self-explicated," has been described by Srinivasan (Srinivasan; Srinivasan & Wyner). Perhaps the most popular approach is an indirect approach which involves derived importances from conjoint analysis. KBF (attribute) importances are obtained for each respondent by computing differences in utility between the most and least preferred level of each KBF, and then percentaging these against their sum. For an overview of conjoint analysis see IntelliQuest. Other indirect approaches include applications of probit analysis (Rao & Winter) and "benefit bundle analysis" (Green, Wind & Jain).

3. Form Homogeneous Segments Based on Relative Importance of KBFs

Typically, some type of cluster analysis is used to accomplish this step. For an example, see Moriarty and Reibstein. For an excellent review of cluster analysis with suggestions for application,

see Punj & Stewart. Neal presented a thorough empirical comparison of available software for cluster analysis (Neal, 1989). To date, nearly all applications of cluster analysis utilize metric-scaled variables. For those applications which require clustering of nominal and/or categorical responses, one should refer to Mullet.

4. Profile Segments to Make Actionable

There are many techniques which have been used to profile market segments. These techniques can be categorized as "variable reduction" and "detailed analysis" techniques (Magidson, 1988a). Variable reduction techniques are used to distill a large number of prospective descriptive variables down to a manageable number that will be examined in detail by the "detailed analysis" techniques. Another distinction is that the "variable reduction" techniques are multi-variable, while the "detailed analysis" techniques are multivariate. Allison and Forsyth classify these techniques similarly but refer to them as "hypothesis generation" and "hypothesis testing."

Variable Reduction Techniques

Crosstabulations are simple to interpret, particularly if accompanied by a chi-square statistic and/or a Cramer's V. A shortcoming is that they generate too much output, especially if one looks at 3-way up to n-way tables. Typically, the descriptor variables (demographics, etc.) are used in the stub and the cluster solution is the banner.

AID (Automatic Interaction Detector) is an *ad hoc* technique that provides a visual display of the results in tree-diagram form. AID first recodes all predictor variables into dichotomous variables so that the resulting dichotomy represents the lowest within-group sum of squared deviations for the criterion variable. AID then selects the dichotomized predictor variable which has the lowest within-group sum of squared deviations and splits the sample on it. The algorithm repeats for each subgroup and continues until a pre-specified stopping criterion is satisfied. Since each subgroup can (and often does) split differently, interaction effects are uncovered and displayed visually.

AID is of limited usefulness in this context since the criterion variable must be metric or dichotomous (a two-group cluster solution). The predictor variables must be categorical. Also, since AID is limited to dichotomous splits of the predictor variables and does not properly adjust for the likelihood of finding a dichotomization by chance due to sampling variability, it is biased in favor of variables containing many categories because there are more ways to dichotomize such variables. See Myers & Tauber, Magidson (1988a), and Green for applications of AID.

CHAID (Chi-Square Automatic Interaction Detector) is a statistical technique that also provides the results in tree-diagram form. CHAID first forms all two-way tables of criterion by the predictor variables. For each table, the predictor variables are recoded in all possible allowable ways and the particular recoding that results in the strongest relationship to the dependent variable is selected. CHAID then determines which recoded predictor variable is most significantly related to the criterion variable. CHAID then splits the sample on the best predictor and repeats the analysis on each

subgroup in turn. The analysis continues until a pre-specified stopping criterion is satisfied. Like AID, since each subgroup can (and often does) split differently, interaction effects are uncovered and displayed visually.

CHAID requires all variables to be categorical since it is based on the chi-square statistic, but includes a statistical algorithm that reduces the number of cells that is typical of a full multiway table. Unlike AID, CHAID is not limited to a dichotomous criterion variable or to dichotomous splits of the predictor variables. See Perrault & Barksdale and Magidson (1988a) for more on CHAID.

In this segmentation context, the cluster solution would be the criterion variable and the descriptor variables would be the predictor variables.

Detailed Analysis Techniques

Correspondence analysis has been described as "a multivariate statistical method that represents graphically the rows and columns of a categorical data matrix in the same low-dimensional space" (Hoffman & Franke). This graphical approach positions each category by its joint occurrences (crosstabulations) with other categories. Correspondence analysis positions categories on the map so that categories that have a strong positive relationship will be close to one another, whereas categories with strong negative relationship will be relatively far apart from one another.

In a segmentation context, the columns of the crosstabulation would be the cluster segments, while the rows of the crosstabulation would be the various segment descriptor variables. Note that since correspondence analysis treats rows and columns identically, the same solution would result if the rows and columns were transposed.

ANOVA, multiple regression and multiple discriminant analysis are all variations of the general linear model. ANOVA and multiple regression, like AID, are of limited usefulness in this context since the criterion variable must be metric or dichotomous (such as a two-group cluster solution).

There are three problems associated with the use of a linear regression model for dichotomous data. First, the predicted values may fall outside of the possible range of values. Although the predicted values can be constrained to values of 0 or 1, this is a theoretically unsatisfactory solution. The second problem with applying ordinary least squares (OLS) to dichotomous data is that the homogeneity of error variance will seldom be met. This problem can be addressed through the use of weighted least squares, but it does complicate the analysis. Finally, the assumption of normally distributed residuals that is required for significance tests of coefficients will not be met when using dichotomous response data. Incidently, a related but little used technique, multiple classification analysis (a type of dummy-variable regression model with some useful features for exploratory data analysis), shares these same problems (Green).

The principal assumption in multiple discriminant analysis is that the predictor variables follow a multivariate normal distribution. This assumption is violated when most predictors are categorical.

Finally, there is not a theoretically proper way of handling interactions in the general linear model. If the data are categorical, the definition of interaction terms is much easier. What is often the case, however, is that some of the predictor variables are metric while some are categorical, which makes the specification of interaction terms especially difficult.

Logit analysis (logistic regression) and multinomial logit analysis require far fewer assumptions than regression or discriminant analysis. Statisticians recommend logistic regression as a general alternative to discriminant analysis when the predictors do not follow the multivariate normal distribution. Since logistic regression is also theoretically justified under the assumptions of discriminant analysis, many statisticians recommend that logistic regression always be used in place of discriminant analysis (Magidson, 1988a).

Moreover, the logit model provides probability estimates as part of its output. The idea behind logit analysis is to model odds (specifically, the natural logarithm of the odds) instead of probabilities. An "odd" is simply the ratio of the probability of success over the probability of failure. Of course, the probability of failure is equal to 1 minus the probability of success. Although the logit model is based on odds, predictions from logit analysis can be transformed and presented in terms of predicted probabilities. Although probabilities must be between zero and one, odds have no finite maximum values. By taking the natural logarithm of the odds, we obtain the logit, which has no finite maximum.

Magidson claims that another advantage of the logit model is the assumption of no interaction terms. He goes on to say that the logit model will not only provide meaningful parameter estimates, but that logit also has both theoretical and computational advantages over all other models. "Efficient algorithms are available for computing maximum likelihood estimates under binomial, multinomial, and Poisson distributional assumptions. Moreover, the logistic distribution is the only distribution that is theoretically justified for modeling the probability of group membership under the assumptions of classical discriminant analysis" (Magidson, 1988b).

Log-linear modeling techniques, developed by Leo Goodman and others during the period of 1968 to 1972, provided a major advance in statistical modeling of categorical data. It should be noted that when all predictors are categorical, logit analysis is a special case of the general log-linear model. In addition, log-linear modeling includes features for identifying and statistically assessing interaction effects, a feature that is generally lacking in regression and logistic regression algorithms. Since log-linear modeling may be used to identify and assess higher order interactions, the full n-way table of frequency counts must be used as input. Because the size of an n-way table expands multiplicatively as a function of each variable's number of categories, it is important to screen the variables to reduce the number of categories prior to conducting log-linear modeling (Magidson, 1988a). The variable reduction techniques discussed above are effective ways to accomplish this task.

Probit analysis or multivariate probit analysis are first cousins to logit analysis and multinomial logit analysis. Probit and logit deal with the identical problem of predicting the level of a non-metric dependent variable. They differ solely in the assumption made about the frequency distribution of the criterion variable. In probit, the distribution is assumed to be normal, while in logit, a logistic

distribution is assumed. Empirically, the choice of model makes little difference. Since the results are similar, computational ease and availability of software are usually the primary considerations in model choice (Doyle).

Finally, canonical correlation, a technique that simultaneously classifies (Step 3) and discriminates (Step 4) should be mentioned (Myers & Tauber; Wind). Canonical analysis is a multivariate technique that can relate two (or more) sets of variables, measured across a group of respondents, in both a clustering and a predictive way. Other techniques do one or the other, but not both. If the researcher is not concerned with maintaining the distinction of a criterion variable set, canonical correlation should be considered.

Combination of Variable Reduction Techniques and Detailed Analysis Techniques

What has been found to work best in practice is a variable reduction technique to provide guidance on which predictors, which categories within predictors, and which types of interactions to include in a subsequent second-stage detailed analysis technique. Green suggested an AID/logit combination. More recently Magidson (1988a) suggested a CHAID/logit or a CHAID/log-linear combination.

CASE EXAMPLES

The use of profiling techniques beyond crosstabulations, ANOVA, discriminant analysis, and AID are quite rare in market research practice (Neal, 1988). The next section will demonstrate two of these traditional techniques, crosstabulations and discriminant analysis, along with a few relatively new and less widely used techniques, correspondence analysis, CHAID and multinomial logit analysis. AID will not be demonstrated because it is not applicable when there are more than two cluster-based needs segments (which is usually the case). Log-linear modeling and probit analysis will not be demonstrated because of their theoretical similarities (as discussed above) to logit analysis. Please note that the studies used in the following examples were not conducted for the purposes of market segmentation. They are presented here only for the purpose of demonstrating the techniques.

Example 1

The first example is from a study of a commercial service. The database is proprietary and has been disguised for use as an example. The 515 respondents in the survey were representative of known subscribers to a related service.

The ACA (Adaptive Conjoint Analysis) conjoint module was incorporated as a section of the questionnaire. From the resulting utilities, attribute importances were derived for each respondent by computing the differences in utility between the most and least preferred level of each attribute, and percentaging these to their sum. These derived attribute importances were used as the basis for the cluster procedure.

A cluster analysis using a unique k-means clustering procedure developed by Richard Johnson and published by Sawtooth Software, Convergent Cluster Analysis (CCA), was used to produce alternative cluster solutions. The four-cluster solution was selected as the most operational after examining the separation between groups, the reproducibility of each solution, and the profiled description of the clusters with respect to the variables used to form the clusters.

The variables selected for use in describing the cluster groups were demographic variables, such as age, income, ethnic background, employment status, and home ownership; some variables relating specifically to the service being studied, including the number of phones and phones lines installed at the respondent's home, whether the respondent owned a personal computer or a car phone; as well as several psychographic measures.

Crosstabulations

Typically the first step in profiling needs-based segments is to examine 2-way crosstabs. The most significant and other expected descriptive variables were run as banners and the probabilities from the chi-square statistic calculated. (Smaller numbers in the first column indicate the largest difference.) The output from P-STAT v2.13 statistical package has been summarized in Figure 1.

There are almost no significant variables in the study (four at the .05 level). Many variables expected to describe the needs-segments are not significant in the banner output. In order to uncover any interaction effects, many n-way crosstabs would need to be run. This is time consuming and somewhat inefficient. In some cases, the researcher would expect that collapsing several categories in the independent variables would increase probabilities given the lower degrees of freedom as well as the potentially more balanced distribution among categories in the independent variable, but the problem remains of finding the optimal recode for each predictor variable.

In an attempt to compensate for the sample size's impact on the significance of the chi statistic, the Cramer's V statistic was calculated for each of the variables as well. These statistics are as inconclusive as the chi-square in describing our segments. While only age, car phone, ethnic and income are significant, all of the variables listed will be included in CHAID in an effort to uncover more significant relationships.

CHAID

CHAID is especially useful as a variable reduction technique. Although the predictor variables must be categorical, the optimal recoding of the predictor variables within CHAID allows the researcher to input a recoded version of metric variables with a minimum of *a priori* coding. Also, as a variable reduction technique, the outcome of this analysis will simply be to determine which variables to carry forward to more descriptive (and predictive) techniques, where the power of the metric variables will be more fully utilized. It is often the case in applied market research that the majority of variables available for market segment description are categorical. While categorical variables are less powerful than metric variables, the plentifulness of categorical data warrants the use of techniques which extract the most information from these data as possible.

Figure 2 contains the condensed output from the SPSS-PC+ CHAID run. In addition to the p-value for the recoded variable in relation to the dependent variable, the collapsed categories for each variable are given. Note that the number of collapsed categories for each variable varies, as well as the way each is combined. Each variable is designated as either ordinal, where only adjacent categories may combine; ordinal with a floating category that is allowed to combine with any other category; or free format, which allows any categories to combine.

The number of resulting categories is determined by the significance level specified by the researcher. CHAID provides the option of setting different significance levels for the recoding of categories and the prediction of a relationship with the dependent variable. While the significance level for the predictors is the traditional probability from the chi statistic, the significance level for the categories means simply how different categories must be in order to be kept apart for the analysis. Both values in this analysis were set at .05.

Figure 3 shows the tree diagram that CHAID produces. It is important to note that while the tree is a nice visual display of the information in the table, there is a danger in looking only at the diagram output. Frequently — particularly with large data sets since there is no adjustment for sample size with the chi statistic — many variables will be highly significant, and only the most significant variable would be chosen for inclusion in the diagram. If the second most significant, although still very significant, variable were chosen for that split in the tree, the diagram would likely look quite different from that point on. It is useful, then, to examine the table in addition to the tree diagram for important variables that should be selected for additional analyses.

The four variables shown as significant in the crosstabs are notably more significant after the optimal recoding done in CHAID. Of the more significant variables from the crosstabs, the biggest increase in significance is in the variable with the largest number of categories, "premium," with nine categories, which increased from .09 to .03. The largest increase in significance overall is "numinhh," which had a probability of .55 in the crosstabs, .4 in the analysis of the total group in CHAID, and .1 in the sub-group analysis. This variable was reduced from nine to two categories in CHAID.

In addition to the four variables shown as significant in the banners, "premium," "forced," and "yrshome" are significant at the .05 level in CHAID. Several other variables significant at the .1 level also merit inclusion in subsequent analyses. Many variables may be more confidently eliminated from this point on, allowing for cleaner models in the final analyses.

Correspondence Analysis

Correspondence maps are relatively straightforward to interpret. If two points are relatively close to one another, the variables (or groups) represented by those points are considered to be relatively closely related to each other. Similarly, if two points are relatively far apart from one another, the variables (or groups) represented by those points are considered to be relatively unrelated to one another. As discussed above, all variables must be dichotomous. Categorical variables must be converted to dummy variables and metric variables must be recoded and then converted to dummy variables. In this example, all the variables that are closer to cluster group one than to any other cluster group can be considered to be most closely related to cluster group one, and so on.

By examining the correspondence map shown in Figure 4, one can see that "forced6-9," "frustrate1-2," and "homemaker" are among those variables that are relatively near to cluster group one. "Nownpc," "line1," "numinhh1-4," "etho," and "phone1" are relatively near group two. "Forced1-4," "premium6-8," "yrshome25-," "ethw," "I10K50K," "nhomemaker," "ncarphone," and "phone2+" are near group three. Finally, cluster group four is near "lines2+," "age44-," and "premium9."

Discriminant Analysis

Next, a discriminant analysis was run on those variables that were identified as being the most relevant in the crosstabulations and CHAID. The output from the SPSS-PC+ v3.1 statistical package is presented in condensed form in Figure 5. Of the 515 cases that were available, 273 were randomly selected to be used in the analysis, leaving 242 to be used for prediction. The classification results show that 32.2% of the holdout sample was correctly classified, which is a 126.5% improvement over the percentage correctly classified that one would expect from chance.

The linear discriminant function minimizes the probability of misclassification if in each group the variables are from the multivariate normal distribution and the covariance matrices for all groups are equal. Since several of the predictor variables are dichotomous, the assumption of multivariate normality is clearly violated. The highly significant Box's M test indicates that the covariance matrices are not equal. However, when the sample sizes in the groups are large, the significance probability may be small even if the group covariance matrices are not too dissimilar. The test also tends to call matrices unequal if the normality assumption is violated. In situations such as this where the predictor variables are a mixture of continuous and dichotomous variables, a variety of nonparametric classification procedures are available as well (see Hand or Goldstein & Dillon).

The significance of the canonical discriminant functions was tested by testing all the means simultaneously and then excluding one function at a time. Using such successive tests, it was found that the first two discriminant functions are significant, and the third was not.

The standardized canonical discriminant function coefficients (also known as weights) can be used to interpret the relative contribution of the variables to the functions. The interpretation of the coefficients is analogous to the interpretation of beta weights in regression and is therefore subject to the same criticisms. For example, a relatively small coefficient may indicate either that the variable is not relevant in determining a relationship or that is has been partialed out of the relationship because it is highly collinear with another variable(s).

As a result of potential problems with the standardized canonical discriminant function coefficients, the structure matrix (containing discriminant loadings) will be used instead. The discriminant loadings represent the simple correlation between each variable and the discriminant function. The variables "premium" and "age" correlate relatively highly with the first function while "carphone," "I50K75K," "ethw," and "numphone" correlate highly with the second discriminant function. The third function, while not quite significant, is correlated highly with "homemake," "numinhh," and "I25K35K."

By analyzing how the means of these variables (see first table in Figure 6) vary across cluster groups, a actionable profile of each cluster group can be determined. Cluster group one is very high on "ethw" and "I50K75K" while low on "homemake," "I25K35K," and "numinhh." Cluster group two is high on "age" and "numinhh" and low on "ethw," "carphone," "I50K75K," "premium," and "numphone." Cluster group three is high on "homemake," "carphone," "I25K35K," and "numphone." Finally, cluster group four is high on "I25K35K," "premium," and "numinhh" while low on "age."

Logit Analysis

As can be seen from the "convergence achieved" flag on the Systat Logit Module v4.1 output in Figure 6, the maximum likelihood routine did indeed converge.

As can be seen from Figure 6, the logit analysis had the identical overall classification accuracy as the discriminant analysis: 32.2% of the holdout sample was correctly classified, which is a 126.5% improvement over the percentage correctly classified that one would expect from chance. Interestingly however, there were non-trivial differences in terms of which particular cases were correctly or incorrectly classified (not shown).

The Wald tests on parameters across all choices are tests that all coefficients for each particular variable (as shown on the bottom of Figure 6, each variable has as many coefficients as there are groups minus one) are simultaneously zero. Note that "premium," "age," "homemake," and "numinhh" are all significant at the .05 level.

By analyzing the significance of the variables for each cluster group, an actionable profile of each cluster group can be determined. Cluster group one (the first set of parameters on the "Results of Estimation" table in Figure 6) was very significant on "premium" (negative) and "numinhh" (negative). Cluster group two is very significant on "age" (positive), "premium" (negative), and "numphone" (negative). Cluster group three is significant on "homemaker" (positive), "premium" (negative), and "numinhh" (negative). An artifact of logit analysis is that it cannot calculate the coefficients for all the groups at one time. To determine the coefficients for cluster group four, one would simply rerun the logit analysis excluding a different cluster group. The coefficients of the variables have been shown to be stable across various runs which each exclude a different cluster group.

Another way of profiling cluster groups in logit analysis is to analyze the cluster means on the variables that are significant in the Wald tests.

Summary

Among the variable reduction techniques, CHAID proved to be the most useful. Its optimal recoding algorithm shed light on relationships that would have been overlooked by examining the crosstabulations alone. However, the practice of making directional splits on the basis of minute differences in p-statistics limited CHAID's usefulness as a detailed analysis technique.

Correspondence analysis' usefulness as a detailed analysis technique is limited by the lack of statistics to guide interpretation as to which variables are <u>really closely</u> related to a group or merely more <u>closely related</u> to a particular group than to any other group.

In this example, discriminant analysis and multinomial logit analysis yielded very similar results. Logit's inclusion of significance levels for each variable for each cluster group lends itself to somewhat more precise interpretation than does relying on the discriminant loadings, which have no associated significance levels.

Example 2

The second example is from a survey of personal computer buyers. Again, the database is proprietary and has been disguised for use as an example. The 708 respondents in the study were representative of known prospects to purchase a personal computer.

Again, the ACA conjoint module was incorporated as a section of the questionnaire, and the derived attribute importances for each respondent were used as the basis for the cluster procedure.

CCA was used to produce the alternative cluster solutions. The three cluster solution was selected after examining the configurations of each of the cluster solutions.

The variables chosen for use in profiling the needs-based groups included demographic variables, including level of personal computer expertise, hours per week use a PC, job title, and level of decision-making authority for PC purchases; firmagraphic variables, including revenue, existence of an approved brand or source list, and installation of a mini or mainframe computer; as well as data relating to purchase preferences for personal computers, such as preferred brands (divided into three tiers for analysis), preferred purchase source, purchase decision responsibilities for personal computers, interest in buying mail-order for various PC needs, and the number of personal computers in the most recent order.

Crosstabulations

Again, the most significant and other expected descriptive variables were run as banners and the probabilities from the chi-square statistic calculated. The output has been summarized in Figure 7.

Unlike the previous example, there are many significant variables in the crosstabs for this study. The researcher faces a different problem here: how to narrow the set of variables to an operational description of the market segments. Although many variables are significant at this primary level, n-way crosstabs need to be run in order to uncover any interaction effects. Other problems from the previous data set still exist, such as how to optimally recode categories and which variables to include in subsequent analyses.

Again, in an attempt to compensate for the sample size's impact on the significance of the chi statistic, the Cramer's V statistic was calculated for each of the variables, with somewhat better, yet still inconclusive results in describing the segments. Hence, all variables from this banner run will be included in CHAID in an effort to eliminate some unnecessary variables before proceeding to further analyses.

CHAID

Figure 8 contains the condensed output from the SPSS-PC+ CHAID run. Again, both significance levels were set at .05 for this analysis. Figure 9 shows the tree diagram that CHAID produces. As before, this is shown as a visual aid, to be used in addition to the tables.

In this example, CHAID verified that the variables that were not significant in the crosstabs could be excluded from further consideration. It also demonstrated that while in the crosstabs "hr.wk.pc" had a probability of .1 and "depmis" a .11 in the crosstabs, the eight categories for "hr.wk.pc" collapsed to three groups, and increased the probability in the sub-group analysis to .008, while the dichotomous "depmis" was not significant enough to remain in the analysis.

Correspondence Analysis

By examining the map in Figure 10, one can see that "enduser," "purchmass," "hours<=20" and "mainmini<=1" are among the variables located relatively near to cluster group one. "Unix," "corp-"pcsupport," "pc21+," and "rev$500m+" are relatively near group two. Group three is nearest "tier3," "pc3-20," and "advanced."

Discriminant Analysis

Once again, a discriminant analysis was run on those variables that were identified as being the most relevant in the crosstabulations and CHAID. The output is presented in condensed form in Figure 11. Of the 708 cases that were available, 370 were randomly selected to be used in the analysis, leaving 338 to be used for prediction. The classification results show that 47.34% of the holdout sample was correctly classified, which is a 139.79% improvement over the percentage correctly classified that one would expect from chance.

As in the first example, several of the predictor variables are dichotomous, so the assumption of multivariate normality is clearly violated. The highly significant Box's M test indicates that the covariance matrices are not equal.

The significance of the canonical discriminant functions were tested in the same manner as in example one and only the first discriminant function was significant.

The structure matrix (containing discriminant loadings) shows that the variables "tier1," "appsourc," "appbrand" and "pursourc" were highly correlated with the first discriminant function.

By analyzing how the means of these variables (see first table in Figure 12) vary across cluster groups, an actionable profile of each cluster group can be determined. Cluster group one is very high on "appbrand," "appsourc," and "pursourc" while low on "tier1." Cluster group two is high on "tier1" and low on "appbrand," "appsourc," and "pursourc." Cluster group three is very similar to cluster group one on the variables that load highly on the first function, but is very different from cluster group one on the variables that load highly on the second function ("tier3" and "enduser").

Logit Analysis

Once again convergence was achieved.

As can be seen from Figure 12, the logit analysis had a slightly better classification accuracy than the discriminant analysis: 48.6% of the holdout sample was correctly classified, which is a 143.5% improvement over the percentage correctly classified that one would expect from chance.

The Wald tests show that "tier1" and "pursourc" are all significant at the .05 level. Cluster group one is marginally significant only on "enduser" (positive). Cluster group two is significant on "tier1" (positive), and "pursourc" (negative).

Summary

Once again CHAID proved to be the most useful variable reduction technique. Again, its optimal recoding algorithm identified relationships that would have been overlooked by examining the crosstabulations alone.

Once more, correspondence analysis was of limited usefulness as a detailed analysis technique. In this example, multinomial logit had a slightly higher classification accuracy than discriminant analysis.

SUMMARY

There are many useful and productive approaches to segmentation analysis. This paper focused on cluster-based needs/benefits segmentation since the authors have found this approach to be particularly useful in their work. The success of a cluster-based needs/benefit segmentation effort is often dependent on one's ability to make the cluster-based segments actionable through the use of cluster profiling techniques. This paper categorized and examined several techniques that can be used for this purpose. The authors found CHAID to be an extremely useful variable reduction technique. CHAID not only provided insight as to which variables to drop from subsequent analysis, but, through its optimal recoding algorithm, it also identified several variables that may have been otherwise overlooked.

Correspondence analysis was of marginal usefulness because of its lack of statistics at the variable level. Discriminant analysis and logit analysis performed almost identically with two exceptions. In example two, logit analysis had a higher classification accuracy than discriminant analysis. Also, logit's derivation of significance for each variable for each cluster group made interpretation more precise.

It is the authors' opinion that using CHAID followed by a logit analysis is an optimal approach. In most cases, discriminant analysis may classify as well or nearly as well as multinomial logit analysis, but logit analysis lends itself to more precise interpretation. The authors also believe that in some cases logit analysis will significantly out-classify discriminant analysis. We hope that this paper will inspire others to use logit analysis and to add to the comparisons presented above.

REFERENCES

Allison, Neil and John Forsyth (1990), "Describing Needs-Based Segments," *American Marketing Association Advanced Research Techniques Forum Proceedings*, in press.

Assael, Henry and A. Marvin Roscoe Jr. (1976), "Approaches to Market Segmentation Analysis," *Journal of Marketing*, 40 (4), 67-76.

Coles, G. J. and J. D. Culley (1986), "Not All Prospects Are Created Equal," *Business Marketing*, (May), 52-58.

Doyle, Peter (1977), "The Application of Probit & Logit In Marketing: A Review," *Journal of Business*, 5 (September), 235-248.

Frank, Ronald, William Massy and Yoram Wind (1972), *Market Segmentation.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Goldstein, M. and W. R. Dillon (1978), *Discrete Discriminant Analysis.* New York: Wiley & Sons.

Green, Paul E. (1978), "An AID/Logit Procedure for Analyzing Large Multiway Contingency Tables," *Journal of Marketing Research*, 15 (February), 132-136.

Green, Paul E., Donald S. Tull and Gerald Albaum (1988), *Research for Marketing Decisions,* Fifth Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Green, Paul E., Yoram Wind and Arun K. Jain (1972), "Benefit Bundle Analysis," *Journal of Advertising Research*, 12 (2), 31-36.

Greenberg, Marshall and Susan Schwartz McDonald (1989), "Successful Needs/Benefits Segmentation: A User's Guide," *Journal of Consumer Marketing*, 6 (3), 29-36.

Haley, Russell I. (1968), "Benefit Segmentation: A Decision-oriented Research Tool," *Journal of Marketing*, 32 (July), 30-35.

Haley, Russell I. (1984a), "Benefit Segments: Backwards and Forwards," *Journal of Advertising Research*, 24 (1), 19-25.

Haley, Russell I. (1984b), "Benefit Segmentation - Twenty Years Later," *Journal of Consumer Marketing*, 1, 5-13.

Hand, D.J. (1981), *Discrimination and Classification.* New York: Wiley and Sons.

Hoffman, Donna L. and George R. Franke (1986), "Correspondence Analysis: Graphical Representations of Categorical Data in Market Research," *Journal of Marketing Research*, 23 (August), 213-227.

IntelliQuest, Inc. (1989) *Conjoint Analysis: A Guide for Designing and Interpreting Conjoint Studies.* Austin, TX: IntelliQuest, Inc.

Kotler, Philip (1980), *Marketing Management: Analysis, Planning, and Control, 4th ed.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Magidson, Jay (1988a), "CHAID, LOGIT, and Log-Linear Modeling," Datapro Research Corp., (May).

Magidson, Jay (1988b), "Improved Statistical Techniques for Response Modeling," *Journal of Direct Marketing*, 2 (Autumn), 6-18.

Moriarty, Rowland T. and David J. Reibstein (1982), "Benefit Segmentation: An Industrial Application," Report No. 82-110 (November).

Mullet, Gary M. (1987), "Defining Market Segments by the Clustering of Nominal and Categorical Responses," *Journal of Professional Services Marketing*, 3(1/2), 47-66.

Myers, James H. and Edward Tauber (1977), *Market Structure Analysis.* Chicago, IL: American Marketing Association.

Neal, William D. (1988), "Market Segmentation Tutorial," *American Marketing Association Ninth Annual Marketing Research Conference.*

Neal, William D. (1989), "A Comparison of 18 Clustering Algorithms Generally Available to the Marketing Research Professional," *Sawtooth Software Conference Proceedings*, 299-324.

Perrault, W. and H. Barksdale Jr. (1980), "A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research," *Journal of Marketing Research*, 17 (August), 503-15.

Punj, Girish and David W. Stewart (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 20 (May), 134-148.

Rao, Vithala R. and Frederick W. Winter (1978), "An Application of the Multivariate Probit Model to Market Segmentation and Product Design," *Journal of Marketing Research*, 15 (August), 361-368.

Smith, Wendell R. (1956), "Product Differentiation and Market Segmentation as Alternative Marketing Strategies," *Journal of Marketing*, 21 (1), 3-8.

Srinivasan, V. (1988), "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences," *Decision Sciences*, 19, 295-305.

Srinivasan, V. and Gordon A. Wyner (1981), "CASEMAP: Computer-Assisted Self-Explication of Multiattributed Preferences," in Yoram Wind, Vijay Mahajan, and Richard Cardozo, eds., *New Product Forecasting*, Lexington, Massachusetts: D.C. Heath and Company.

Wind, Yoram (1978), "Issues and Advances in Segmentation Research," *Journal of Marketing Research*, 15 (August), 317-337.

SOFTWARE REFERENCES

ACA System for Adaptive Conjoint Analysis
Sawtooth Software, Inc.
P.O. Box 3429
208 Spruce North
Ketchum, Idaho 83340
(208) 726-7772

CCA System for Convergent Cluster Analysis
Sawtooth Software, Inc.
P.O. Box 3429
208 Spruce North
Ketchum, Idaho 83340
(208) 726-7772

Mapwise Graphic Software for Multiple Correspondence Analysis
Market ACTION Research Software, Inc.
Bradley University
Business Technology Center
Peoria, Illinois 61625
(309) 677-3299

P-STAT
P-STAT, Inc.
271 Wall Street
P.O. Box AH
Princeton, New Jersey 08542
(609) 924-9100

SPSS/PC+ CHAID
SPSS, Inc.
444 N. Michigan Avenue
Chicago, Illinois 60611
(312) 329-3300

LOGIT, A Supplementary Module for SYSTAT
SYSTAT, Inc.
1800 Sherman Avenue
Evanston, Illinois 60201
(708) 864-5670

Figure 1 - Summary Banner Statistics - Example 1

Variable	X^2 sig	Cramer's V
ETHNIC	0.00*	0.18
AGE.COD	0.02*	0.13
INCOME	0.02*	0.15
CARPHONE	0.03*	0.13
HOMEMAKE	0.09	0.11
PREMIUM	0.09	0.15
OWNPC	0.14	0.10
HOMESTAT	0.18	0.10
FORCED	0.31	0.13
LINE.COD	0.35	0.09
YRHM.COD	0.55	0.08
NUMINHH	0.55	0.12
FRUSTRAT	0.63	0.12
PHON.COD	0.67	0.14

*significant, $p<.05$

Figure 2 - CHAID Output - Example 1

Analysis of total group.

	Predictor	p-value	r-sq	groups
12	ETHNIC	0.0008	0.012	2 W O
7	AGE.COD	0.004	0.009	2 123 45
13	INCOME	0.005	0.012	2 123456 78
4	PREMIUM	0.03	0.014	3 12345 678 9
10	CARPHONE	0.03	0.006	2 Y N
1	PHON.COD	0.3	0.006	2 123456789 ABC
14	NUMINHH	0.4	0.011	3 1234 56 789
2	LINE.COD	1.0	0.000	1 12345
3	FORCED	1.0	0.000	1 123456789
5	FRUSTRAT	1.0	0.000	1 123456789
6	YRHM.COD	1.0	0.000	1 12345
8	HOMEMAKE	1.0	0.000	1 YN
9	OWNPC	1.0	0.000	1 YN
11	HOMESTAT	1.0	0.000	1 OR

Analysis of subgroup: ETHNIC : W (1/2)

	Predictor	p-value	r-sq	groups
7	AGE.COD	0.02	0.007	2 123 45
10	CARPHONE	0.03	0.006	2 Y N
13	INCOME	0.03	0.009	2 123456 78
4	PREMIUM	0.07	0.007	2 12345678 9
14	NUMINHH	0.2	0.006	2 123456 789
1	PHON.COD	0.3	0.006	2 123456789 ABC
3	FORCED	0.8	0.009	3 1234567 8 9

Analysis of subgroup: ETHNIC : W (1/2)
AGE.COD : 123 (1/2)

	Predictor	p-value	r-sq	groups
3	FORCED	0.04	0.008	2 1234 56789
10	CARPHONE	0.05	0.005	2 Y N
13	INCOME	0.09	0.007	2 123456 78
4	PREMIUM	0.2	0.006	2 12345678 9
14	NUMINHH	0.3	0.006	2 12345 6789

Analysis of subgroup: ETHNIC : W (1/2)
AGE.COD : 123 (1/2)
FORCED : 1234 (1/2)

	Predictor	p-value	r-sq	groups
13	INCOME	0.004	0.016	3 123 456 78

Analysis of subgroup: ETHNIC : W (1/2)
AGE.COD : 123 (1/2)
FORCED : 1234 (1/2)
INCOME : 78 (3/3)

	Predictor	p-value	r-sq	groups
10	CARPHONE	0.03	0.005	2 Y N
14	NUMINHH	0.1	0.003	2 12 3456789

Analysis of subgroup: ETHNIC : O (2/2)

	Predictor	p-value	r-sq	groups
6	YRHM.COD	0.004	0.006	2 1234 5
5	FRUSTRAT	0.07	0.005	2 12 3456789
7	AGE.COD	0.08	0.003	2 123 45
4	PREMIUM	0.4	0.007	2 12345 6789

Figure 3 - CHAID Output - Example 1

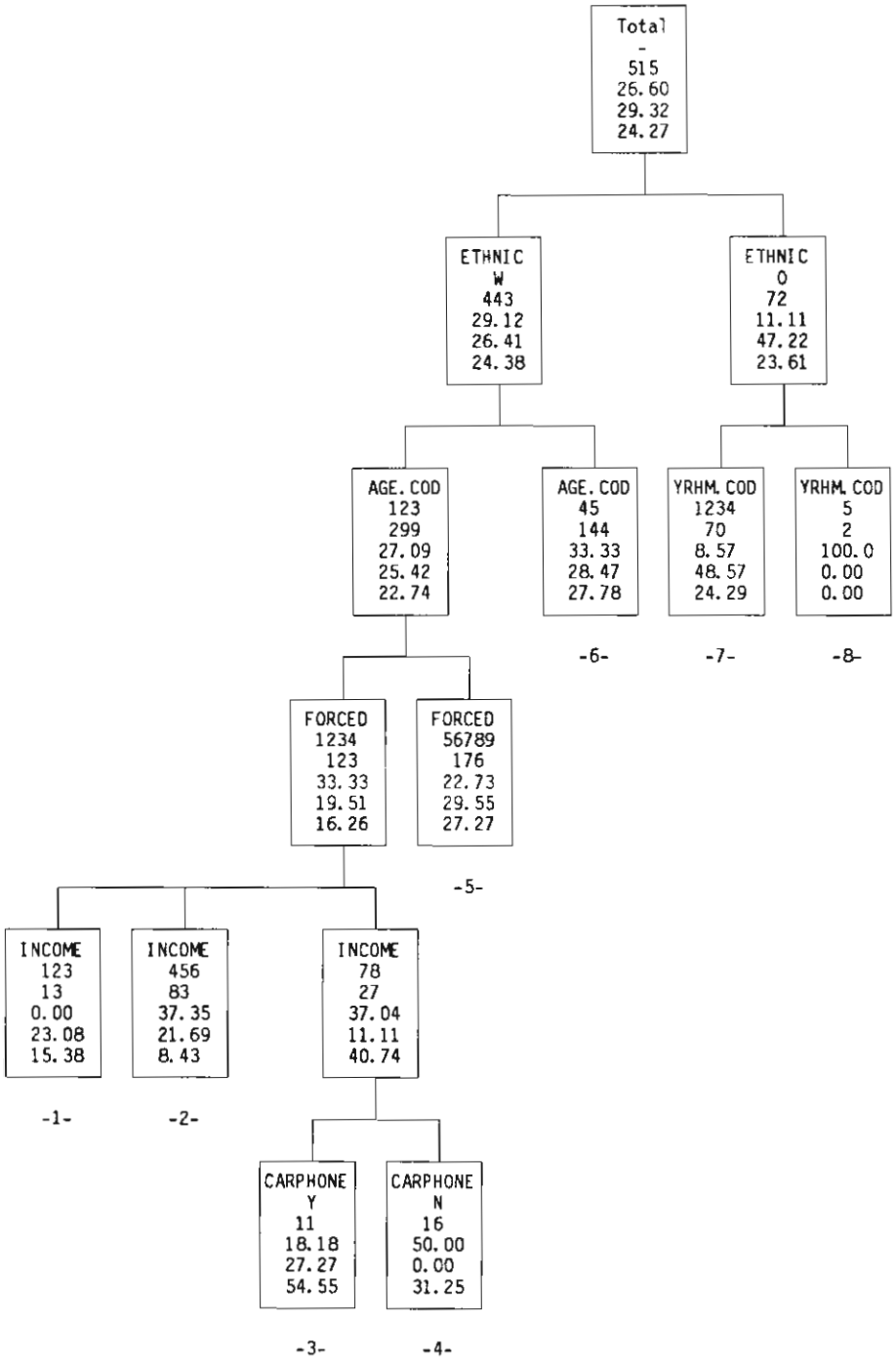


Figure 4 - Correspondence Output - Example 1

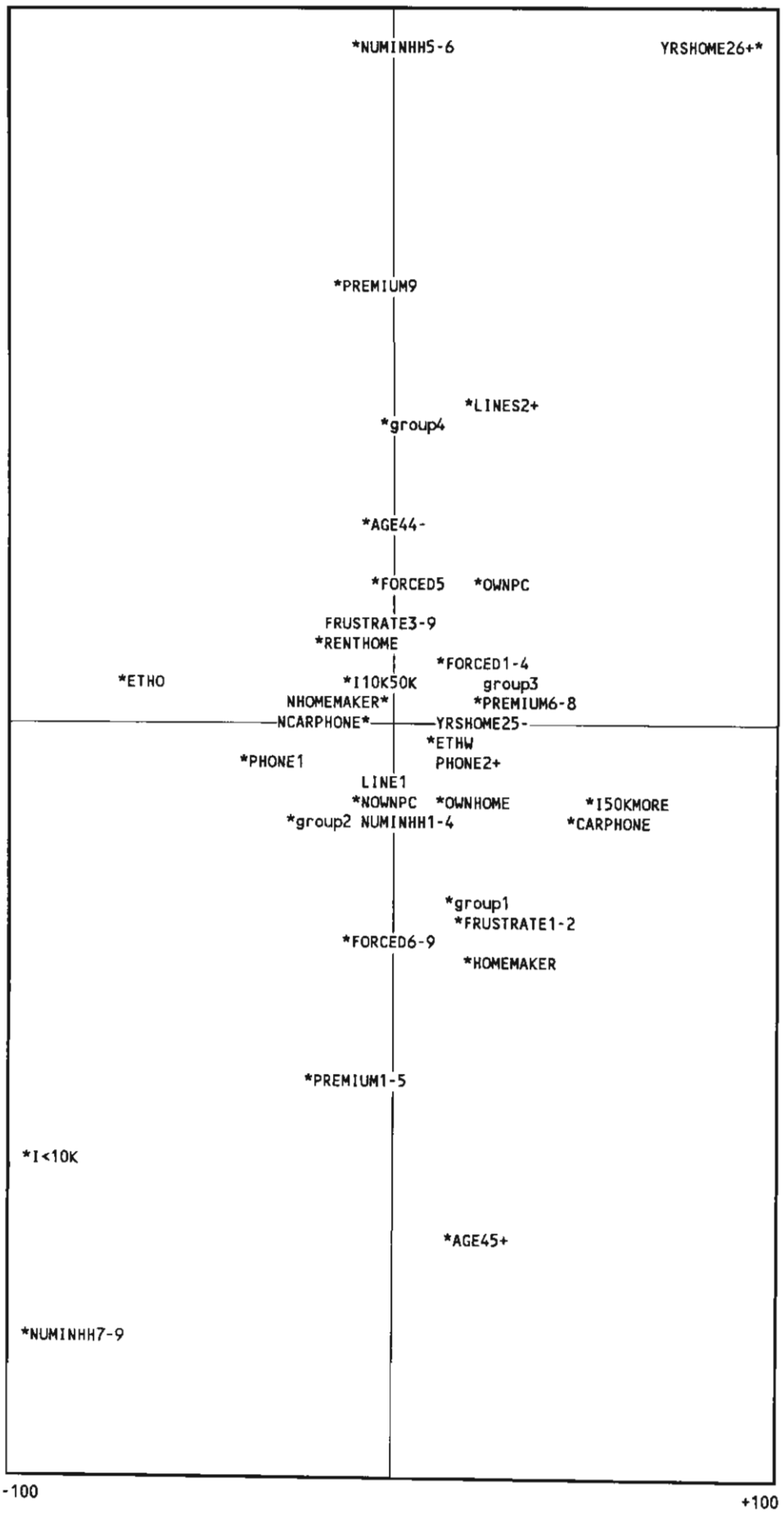


Figure 5 - Discriminant Analysis Output - Example 1

515 (unweighted) cases were processed.
242 of these were excluded from the analysis (used for prediction).
273 (unweighted) cases will be used in the analysis.

Number of Cases by Group

1	69
2	86
3	67
4	51
Total	273

Classification Results for cases not selected for use in the analysis -
Percent of "grouped" cases correctly classified: 32.2%
Classification improvement relative to chance: 126.5%

Test of equality of group covariance matrices using Box's M

Box's M	Approximate F	Degrees of freedom		Significance
581.48	1.9412	273,	122784.7	.0000

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chisquare	DF	Sig
					0	.7478	76.590	39	.0003
1*	.1467	47.79	47.79	.3576	1	.8574	40.528	24	.0188
2*	.0998	32.53	80.32	.3013	2	.9430	15.453	11	.1627
3*	.0604	19.68	100.00	.2387					

* marks the 3 canonical discriminant functions remaining in the analysis.

Standardized Canonical Discriminant Function Coefficients

	FUNC 1	FUNC 2	FUNC 3
HOMEMAKE	-.23207	.35676	.60762
I15K25K	.09232	.28287	.09077
I25K35K	.06759	.28737	.55527
I35K50K	-.20774	.44603	.22065
I50K75K	.09007	.71120	-.12889
MORE75K	-.02722	.40546	.05724
ETHW	.35044	.27325	-.39829
AGE	-.56824	-.05152	-.22767
CARPHONE	.07770	.32351	.27796
PREMIUM	.60673	-.25020	-.02191
NUMLINES	.16433	.11285	.11253
NUMPHONE	.35631	.15240	.10967
NUMINHH	.19317	-.56921	.16340

Structure Matrix:
Pooled-within-groups correlations between discriminating variables and canonical discriminant functions

	FUNC 1	FUNC 2	FUNC 3
PREMIUM	.59342*	-.20857	.00175
AGE	-.48791*	.21313	-.20273
I35K50K	-.25362*	-.00751	.10489
CARPHONE	.22169	.52836*	.21966
I50K75K	.10525	.49351*	-.22919
ETHW	.24868	.37794*	-.36074
NUMPHONE	.26237	.36942*	.17962
NUMLINES	.18135	.21731*	.19630
MORE75K	.11542	.17488*	-.00286
I15K25K	.06926	-.14181*	-.11110
HOMEMAKE	-.18856	.25361	.65038*
NUMINHH	.10614	-.28808	.38982*
I25K35K	.07928	-.10130	.37108*

Figure 6 - Mulitinomial Logit Analysis Output - Example 1

MULTINOMIAL LOGIT ANALYSIS

INDEPENDENT VARIABLE MEANS

PARAMETER	1	2	3	4
1 CONSTANT	1.000	1.000	1.000	1.000
2 HOMEMAKE	0.072	0.151	0.239	0.078
3 ETHW	0.942	0.779	0.866	0.843
4 CARPHONE	0.145	0.047	0.209	0.098
5 I15K25K	0.203	0.198	0.164	0.235
6 I25K35K	0.130	0.186	0.224	0.235
7 I35K50K	0.188	0.267	0.224	0.157
8 I50K75K	0.232	0.093	0.194	0.098
9 MORE75K	0.116	0.070	0.119	0.098
10 AGE	40.072	41.384	39.030	34.275
11 PREMIUM	6.275	5.744	6.149	7.216
12 NUMLINES	1.232	1.128	1.343	1.255
13 NUMPHONE	3.174	2.767	3.403	3.118
14 NUMINHH	2.870	3.174	3.179	3.431

CONVERGENCE ACHIEVED

OUT-OF-SAMPLE CLASSIFICATION TABLE

CORRECT	32.2 %
RELATIVE TO CHANCE	126.5 %

WALD TESTS ON PARAMETERS ACROSS ALL CHOICES

PARAMETER	WALD STATISTIC	CHI-SQ SIGNIF	DOF
1 CONSTANT	6.801	0.079	3.000
2 HOMEMAKE	8.306	0.040	3.000
3 ETHW	7.111	0.068	3.000
4 CARPHONE	3.529	0.317	3.000
5 I15K25K	1.011	0.799	3.000
6 I25K35K	3.345	0.341	3.000
7 I35K50K	3.419	0.331	3.000
8 I50K75K	5.039	0.169	3.000
9 MORE75K	2.177	0.536	3.000
10 AGE	10.679	0.014	3.000
11 PREMIUM	12.831	0.005	3.000
12 NUMLINES	1.079	0.782	3.000
13 NUMPHONE	4.501	0.212	3.000
14 NUMINHH	8.310	0.040	3.000

Figure 6 (cont) - Mulitinomial Logit Analysis Output - Example 1

RESULTS OF ESTIMATION

PARAMETER	ESTIMATE	S.E.	T-STAT	P-VALUE	ODDS RATIO	95.0% BOUNDS UPPER	95.0% BOUNDS LOWER
1 CONSTANT	1.311	1.355	0.967	0.334			
2 HOMEMAKE	0.147	0.754	0.194	0.846	1.158	5.079	0.264
3 ETHW	0.903	0.670	1.346	0.178	2.466	9.176	0.663
4 CARPHONE	0.298	0.703	0.423	0.672	1.347	5.344	0.339
5 I15K25K	0.173	0.654	0.264	0.791	1.189	4.285	0.330
6 I25K35K	-0.464	0.689	-0.673	0.501	0.629	2.427	0.163
7 I35K50K	0.716	0.722	0.991	0.322	2.046	8.429	0.497
8 I50K70K	1.149	0.798	1.440	0.150	3.156	15.081	0.661
9 MORE75K	0.722	0.882	0.819	0.413	2.060	11.599	0.366
10 AGE	0.034	0.018	1.878	0.060	1.035	1.073	0.999
11 PREMIUM	-0.250	0.105	-2.378	0.017	0.779	0.957	0.634
12 NUMLINES	-0.122	0.281	-0.434	0.665	0.885	1.536	0.510
13 NUMPHONE	-0.100	0.146	-0.689	0.491	0.904	1.203	0.680
14 NUMINHH	-0.418	0.161	-2.593	0.010	0.658	0.903	0.480
1 CONSTANT	3.157	1.262	2.502	0.012			
2 HOMEMAKE	0.881	0.671	1.312	0.190	2.412	8.992	0.647
3 ETHW	-0.662	0.514	-1.288	0.198	0.516	1.413	0.188
4 CARPHONE	-0.319	0.813	-0.393	0.694	0.727	3.575	0.148
5 I15K25K	-0.127	0.630	-0.202	0.840	0.880	3.025	0.256
6 I25K35K	-0.186	0.634	-0.294	0.769	0.830	2.874	0.240
7 I35K50K	0.826	0.681	1.213	0.225	2.285	8.688	0.601
8 I50K75K	0.105	0.818	0.129	0.897	1.111	5.520	0.224
9 MORE75K	0.329	0.899	0.366	0.714	1.390	8.096	0.239
10 AGE	0.056	0.018	3.085	0.002	1.058	1.096	1.021
11 PREMIUM	-0.357	0.101	-3.535	0.000	0.700	0.853	0.574
12 NUMLINES	-0.316	0.325	-0.971	0.332	0.729	1.380	0.385
13 NUMPHONE	-0.320	0.162	-1.979	0.048	0.726	0.997	0.529
14 NUMINHH	-0.237	0.154	-1.534	0.125	0.789	1.068	0.583
1 CONSTANT	1.900	1.333	1.425	0.154			
2 HOMEMAKE	1.514	0.661	2.289	0.022	4.544	16.609	1.243
3 ETHW	-0.145	0.565	-0.256	0.798	0.865	2.618	0.286
4 CARPHONE	0.817	0.684	1.193	0.233	2.263	8.654	0.592
5 I15K25K	0.514	0.737	0.697	0.486	1.672	7.094	0.394
6 I25K35K	0.793	0.720	1.101	0.271	2.211	9.074	0.539
7 I35K50K	1.433	0.778	1.842	0.065	4.192	19.254	0.913
8 I50K75K	1.384	0.867	1.596	0.110	3.992	21.857	0.729
9 MORE75K	1.266	0.941	1.345	0.179	3.546	22.431	0.560
10 AGE	0.022	0.019	1.153	0.249	1.022	1.060	0.985
11 PREMIUM	-0.297	0.105	-2.820	0.005	0.743	0.913	0.605
12 NUMLINES	-0.040	0.225	-0.178	0.859	0.961	1.493	0.618
13 NUMPHONE	-0.068	0.141	-0.485	0.628	0.934	1.231	0.709
14 NUMINHH	-0.390	0.161	-2.419	0.016	0.677	0.929	0.494

LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -373.921
2*[LL(N)-LL(0)] = 82.485 WITH 39 DOF, CHI-SQ P-VALUE = 0.000
MCFADDEN'S RHO-SQUARED = 0.110

Figure 7 - Summary Banner Statistics - Example 2

Variable	X^2 sig	Cramer's V
MAIL.UNI	0.00	0.12
CORPCSUP	0.00	0.12
PCEXPERT	0.00	0.13
TIER3	0.00	0.15
REVENUE	0.00	0.15
APPBRAND	0.00	0.18
APPSOURC	0.00	0.18
NUMPCPUR	0.00	0.18
PURSOURC	0.00	0.21
JOB.TITL	0.00	0.21
TIER1	0.00	0.30
ENDUSER	0.01	0.12
MAIL.NET	0.03	0.10
MINIMAIN	0.03	0.10
HR.WK.PC	0.10	0.12
DEPMIS	0.11	0.08
CEO	0.35	0.05
PCPURINV	0.38	0.07
TIER2	0.49	0.04
DEPMRG	0.51	0.04
DEPPCSUP	0.85	0.02

Figure 8 - CHAID Output - Example 2

Analysis of total group.

	Predictor	p-value	r-sq	groups
2	TIER1	3.6e-14	0.045	2 Y N
7	APPSOURC	7.6e-10	0.031	2 WO N
6	APPBRAND	4.5e-9	0.028	2 WO N
5	PURSOURC	8.7e-9	0.042	3 125 3468 7
19	NUMPCPUR	4.7e-7	0.031	3 123 4567 89A
15	MAIL.UNI	0.0002	0.014	2 1 234
4	TIER3	0.0004	0.010	2 Y N
20	JOB.TITL	0.003	0.039	4 12579AC 3468 B D
18	PCEXPERT	0.004	0.013	3 NI A E
12	CORPCSUP	0.004	0.008	2 Y N
8	END-USER	0.009	0.007	2 Y N
21	REVENUE	0.01	0.015	3 1 2347 56
14	MAIL.NET	0.05	0.007	2 1 234
16	MINIMAIN	0.07	0.007	2 AIN B
17	HR.WK.PC	0.10	0.011	3 12345 6 78
1	PCPURINV	1.0	0.000	1 1234
3	TIER2	1.0	0.000	1 YN
9	DEPPCSUP	1.0	0.000	1 YN
10	DEPMIS	1.0	0.000	1 YN
11	DEPMGR	1.0	0.000	1 YN
13	CEO	1.0	0.000	1 YN

Analysis of subgroup: TIER1 : Y (1/2)

	Predictor	p-value	r-sq	groups
7	APPSOURC	8.5e-6	0.020	2 WO N
6	APPBRAND	1.0e-5	0.020	2 WO N
12	CORPCSUP	0.002	0.010	2 Y N
18	PCEXPERT	0.004	0.010	2 NIA E
15	MAIL.UNI	0.005	0.011	2 1 234
19	NUMPCPUR	0.006	0.018	3 12 34567 89A
17	HR.WK.PC	0.008	0.011	2 1234 5678
14	MAIL.NET	0.02	0.009	2 123 4
20	JOB.TITL	0.2	0.026	3 125789AC 34B 6D
21	REVENUE	0.2	0.006	2 1 234567
5	PURSOURC	0.8	0.008	2 1258 3467

Analysis of subgroup: TIER1 : Y (1/2)
APPSOURC : WO (1/2)

	Predictor	p-value	r-sq	groups
15	MAIL.UNI	0.002	0.012	2 12 34
14	MAIL.NET	0.003	0.010	2 12 34
6	APPBRAND	0.06	0.004	2 WN O
17	HR.WK.PC	0.08	0.007	2 12345 678
20	JOB.TITL	1.0	0.012	2 12589AC 3467BD
21	REVENUE	1.0	0.006	4 1 2 345 67

Analysis of subgroup: TIER1 : N (2/2)

	Predictor	p-value	r-sq	groups
4	TIER3	0.02	0.006	2 Y N
18	PCEXPERT	0.03	0.005	2 N IAE
6	APPBRAND	0.08	0.004	2 W ON
19	NUMPCPUR	0.3	0.006	2 123 456789A
17	HR.WK.PC	0.4	0.009	3 12345 6 78
5	PURSOURC	0.4	0.009	2 13468 257
20	JOB.TITL	1.0	0.015	3 15AC 234688 79D

Figure 9 - CHAID Output - Example 2

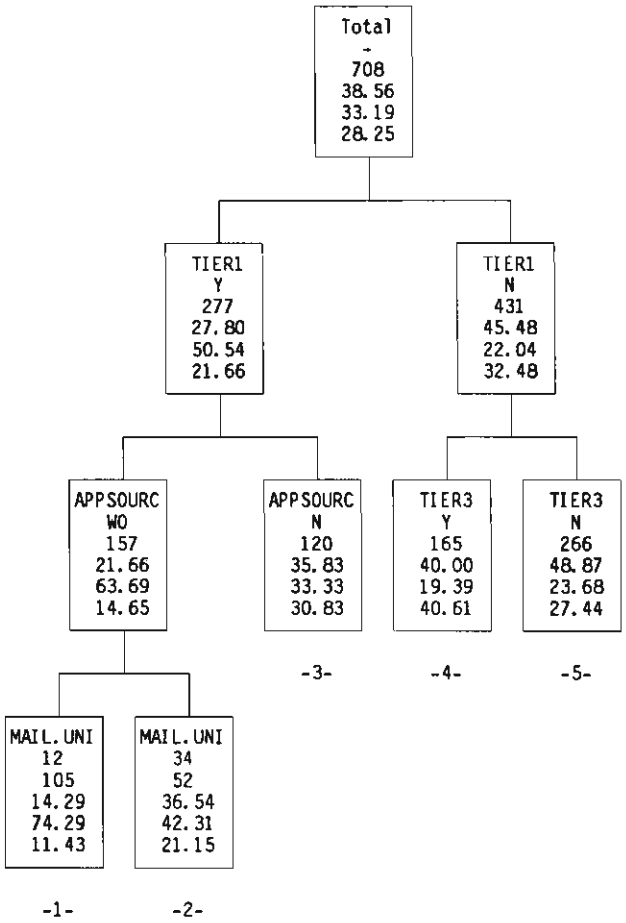


Figure 10 - Correspondence Output - Example 2

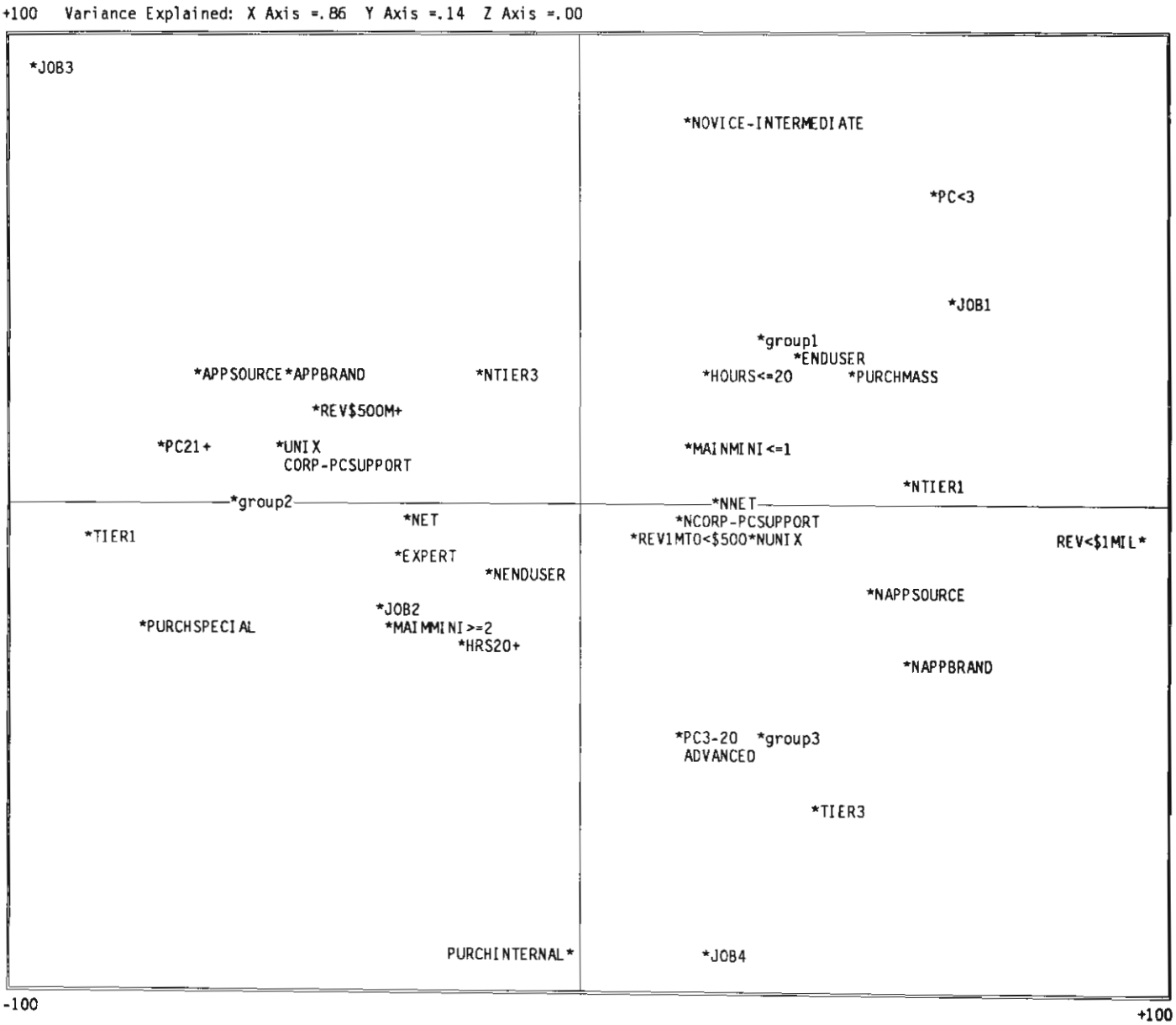


Figure 11 - Discriminant Analysis Output - Example 2

```
          708 (unweighted) cases were processed.
          338 of these were excluded from the analysis (used for prediction).
          370 (unweighted) cases will be used in the analysis.

Number of Cases by Group
         1          146
         2          117
         3          107
     Total          370


Classification Results for cases not selected for use in the analysis -
Percent of "grouped" cases correctly classified:  47.34%
Classification improvement relative to chance:   139.79%


Test of equality of group covariance matrices using Box's M

   Box's M      Approximate F  Degrees of freedom  Significance
    1276.6          5.0108      240,      316402.0      .0000


                    Canonical Discriminant Functions

                Pct of   Cum  Canonical  After  Wilks'
 Fcn Eigenvalue Variance  Pct     Corr      Fcn  Lambda  Chisquare    DF  Sig
                                          :    0   .7910     84.403    30  .0000
   1*     .2246   87.41   87.41    .4283 :    1   .9687     11.462    14  .6494
   2*     .0324   12.59  100.00    .1770 :

   * marks the   2 canonical discriminant functions remaining in the analysis.


Standardized Canonical Discriminant Function Coefficients

             FUNC  1    FUNC  2
APPBRAND      .25943     .24247
TIER1        -.48931     .37647
APPSOURC      .29013     .21977
NUMPCPUR     -.17657    -.08873
TIER3         .09527     .49630
PURSOURC      .36108     .15145
PCEXPERT      .22198     .02804
JOB.TITL      .06996     .22424
ENDUSER       .08196    -.60518
CORPCSUP     -.02683     .11685
REVENUE       .00904     .03898
MAIL.UNI      .20955    -.12960
HR.WK.PC     -.31860     .15271
MINIMAIN      .03240     .30634
MAIL.NET     -.03992    -.07090


Structure Matrix:
Pooled-within-groups correlations between discriminating variables
                                and canonical discriminant functions
             FUNC  1    FUNC  2
TIER1        -.68115*    .20824
APPSOURC      .58867*    .22589
APPBRAND      .57276*    .27621
PURSOURC      .48176*    .11040
REVENUE      -.24771*   -.10040
NUMPCPUR     -.23212*    .00081
MAIL.NET      .22053*   -.18169
CORPCSUP     -.21116*    .00220
ENDUSER       .24303    -.58585*
TIER3         .30618     .50101*
HR.WK.PC     -.25860     .30640*
MINIMAIN      .08218     .24697*
MAIL.UNI      .18577    -.20882*
PCEXPERT     -.13357     .20690*
JOB.TITL     -.08097     .09187*
```

Figure 12 - Multinomial Logit Analysis Output - Example 2

MULTINOMIAL LOGIT ANALYSIS

INDEPENDENT VARIABLE MEANS

PARAMETER	1	2	3
1 CONSTANT	1.000	1.000	1.000
2 APPBRAND	2.322	1.889	2.411
3 TIER1	0.288	0.632	0.336
4 APPSOURC	2.432	1.966	2.505
5 NUMPCPUR	44.404	294.530	49.271
6 TIER3	0.288	0.188	0.383
7 PURSOURC	2.822	2.145	2.879
8 PCEXPERT	3.014	3.162	3.093
9 JOB.TITL	4.144	4.402	4.252
10 ENDUSER	0.432	0.265	0.308
11 CORPCSUP	0.185	0.274	0.187
12 REVENUE	4.877	5.368	4.794
13 MAIL.UNI	2.445	2.162	2.327
14 HR.WK.PC	21.582	25.761	23.402
15 MINIMAIN	2.774	2.735	2.879
16 MAIL.NET	2.096	1.829	2.009

CONVERGENCE ACHIEVED

OUT-OF-SAMPLE CLASSIFICATION TABLE

CORRECT 48.6 %
RELATIVE TO CHANCE 143.5 %

WALD TESTS ON PARAMETERS ACROSS ALL CHOICES

PARAMETER	WALD STATISTIC	CHI-SQ SIGNIF	DOF
1 CONSTANT	7.154	0.028	2.000
2 APPBRAND	2.293	0.318	2.000
3 TIER1	12.324	0.002	2.000
4 APPSOURC	2.185	0.335	2.000
5 NUMPCPUR	0.691	0.708	2.000
6 TIER3	2.321	0.313	2.000
7 PURSOURC	7.515	0.023	2.000
8 PCEXPERT	2.614	0.271	2.000
9 JOB.TITL	0.698	0.705	2.000
10 ENDUSER	4.225	0.121	2.000
11 CORPCSUP	0.170	0.918	2.000
12 REVENUE	0.042	0.979	2.000
13 MAIL.UNI	2.499	0.287	2.000
14 HR.WK.PC	5.442	0.066	2.000
15 MINIMAIN	1.061	0.588	2.000
16 MAIL.NET	0.170	0.918	2.000

Figure 12 (cont) - Multinomial Logit Analysis Output - Example 2

RESULTS OF ESTIMATION

	PARAMETER	ESTIMATE	S.E.	T-STAT	P-VALUE	ODDS RATIO	95.0% BOUNDS UPPER	LOWER
1	CONSTANT	1.765	1.127	1.566	0.117			
2	APPBRAND	-0.145	0.241	-0.601	0.548	0.865	1.387	0.540
3	TIER1	-0.353	0.303	-1.163	0.245	0.703	1.273	0.388
4	APPSOURC	-0.132	0.239	-0.555	0.579	0.876	1.398	0.549
5	NUMPCPUR	-0.000	0.001	-0.272	0.786	1.000	1.002	0.998
6	TIER3	-0.412	0.298	-1.383	0.167	0.663	1.187	0.370
7	PURSOURC	-0.043	0.092	-0.467	0.641	0.958	1.148	0.799
8	PCEXPERT	0.000	0.175	0.002	0.998	1.000	1.409	0.710
9	JOB.TITL	-0.037	0.053	-0.698	0.485	0.964	1.069	0.868
10	ENDUSER	0.542	0.287	1.891	0.059	1.720	3.017	0.980
11	CORPCSUP	-0.129	0.357	-0.360	0.719	0.879	1.770	0.437
12	REVENUE	-0.013	0.067	-0.187	0.851	0.988	1.126	0.866
13	MAIL.UNI	0.049	0.120	0.405	0.685	1.050	1.330	0.829
14	HR.WK.PC	-0.007	0.011	-0.594	0.552	0.993	1.015	0.972
15	MINIMAIN	-0.136	0.138	-0.986	0.324	0.873	1.144	0.666
16	MAIL.NET	0.026	0.145	0.179	0.858	1.026	1.364	0.772
1	CONSTANT	3.350	1.255	2.669	0.008			
2	APPBRAND	-0.394	0.263	-1.499	0.134	0.675	1.129	0.403
3	TIER1	0.717	0.324	2.216	0.027	2.048	3.862	1.087
4	APPSOURC	-0.367	0.252	-1.453	0.146	0.693	1.136	0.423
5	NUMPCPUR	0.000	0.001	0.462	0.644	1.000	1.002	0.999
6	TIER3	-0.425	0.358	-1.187	0.235	0.654	1.319	0.324
7	PURSOURC	-0.282	0.108	-2.600	0.009	0.755	0.933	0.610
8	PCEXPERT	-0.269	0.195	-1.380	0.167	0.764	1.120	0.522
9	JOB.TITL	-0.045	0.061	-0.744	0.457	0.956	1.077	0.848
10	ENDUSER	0.110	0.332	0.333	0.739	1.117	2.139	0.583
11	CORPCSUP	-0.012	0.374	-0.031	0.975	0.989	2.057	0.475
12	REVENUE	-0.013	0.080	-0.163	0.871	0.987	1.154	0.844
13	MAIL.UNI	-0.145	0.134	-1.080	0.280	0.865	1.126	0.664
14	HR.WK.PC	0.019	0.012	1.630	0.103	1.019	1.043	0.996
15	MINIMAIN	-0.120	0.157	-0.762	0.446	0.887	1.207	0.652
16	MAIL.NET	0.068	0.166	0.410	0.681	1.070	1.482	0.773

LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -403.223
2*[LL(N)-LL(0)] = 82.946 WITH 30 DOF, CHI-SQ P-VALUE = 0.000
MCFADDEN'S RHO-SQUARED = 0.103

MEASURING BRAND EQUITY WITH CONJOINT ANALYSIS

Douglas L. MacLachlan
University of Washington

Michael G. Mulhern
Mulhern & Associates

INTRODUCTION

In an era when mergers, acquisitions, and strategic alliances are commonplace and marketing costs to develop consumer franchises continue to increase, it is not surprising that properly managing and measuring the intangibles associated with a brand name is of increasing interest to managers at all organizational levels. Indicators of this interest are several. The Marketing Science Institute has sponsored two conferences in the last three years to address the issue of brand equity. Managing and measuring brand equity was voted second highest among the top ten research priorities for 1990-1992 by MSI trustees. In addition, the American Marketing Association's 1990 Marketing Research Conference devoted a session to this issue. Finally, academic papers focusing on this topic are beginning to be published.

Much of the published material has concentrated on proper positioning and management of a brand (Homa, 1988; Park, *et al.*, 1986; Pecorella, 1988; Speirs, 1988) and optimal approaches to extending brands to new categories (Tauber, 1981; Farquhar, *et al.*, 1990). Several approaches to measuring brand intangibles have also been published. Aaker and Keller (1990) addressed the issue of attitude formation toward brand extensions to a new product class. Using quality of the extended product as the dependent variable, they found that the perceived quality of the extended product was influenced by the quality of the original product, perceived ability to transfer manufacturing skills to the new product class, the perceived complementarity of the new product class, and the difficulty with which the firm could make the extension. Louviere and Johnson (1988) used conjoint analysis to estimate the brand image or position of retailers.

Perhaps the most important use of brand equity measurement would be to assess the intangible value of marketing activities. Firms spend considerable resources building the strength of brands through media advertising and other marketing efforts. For the most part, little is known regarding the effectiveness of these activities. If a reliable way were developed to measure brand equity, its period-to-period changes could be used to evaluate the effectiveness of such expenditures.

The purposes of this paper are the following. First, we investigate whether brand equity can be measured with conjoint analysis. Second, we describe and illustrate one such measurement method. Third, we test some hypotheses regarding brand equity findings and differences in brand equity observed among people with different likelihood of purchasing a product. Fourth, we examine the "dollar value" of brands. Fifth, we discuss some limitations of our proposed method. Finally, we suggest some directions for further research on the topic.

DEFINING BRAND EQUITY

Brand equity can be considered the sum of the intangible values that are associated with a product identified by a brand name or trademark. This equity can be the added value that a brand name endows the product (Farquhar, 1989) or the transfer of a product's heritage as embodied in the associations and behaviors related to the product (Chay, 1990). Its impact on the success or failure of the firm is contained in its contribution to financial performance via margins or volume and to competitive strategy via the creation of sustainable advantage. (See Exhibit 1)

Brand equity can be viewed from the perspective of the firm, the channel (trade), or the consumer (Shocker and Weitz, 1988). From the firm's point of view, brand equity may be viewed as the incremental cash flow generated by the use of a brand name. The firm may also consider it as a leading indicator of financial performance (Allan, 1990). From the trade's perspective, a strong brand implies increased leverage in entering markets and negotiating terms. From the consumer's perspective, brand equity is a utility or value not explained by tangible product attributes and a source of loyalty that goes beyond preference for the tangible product.

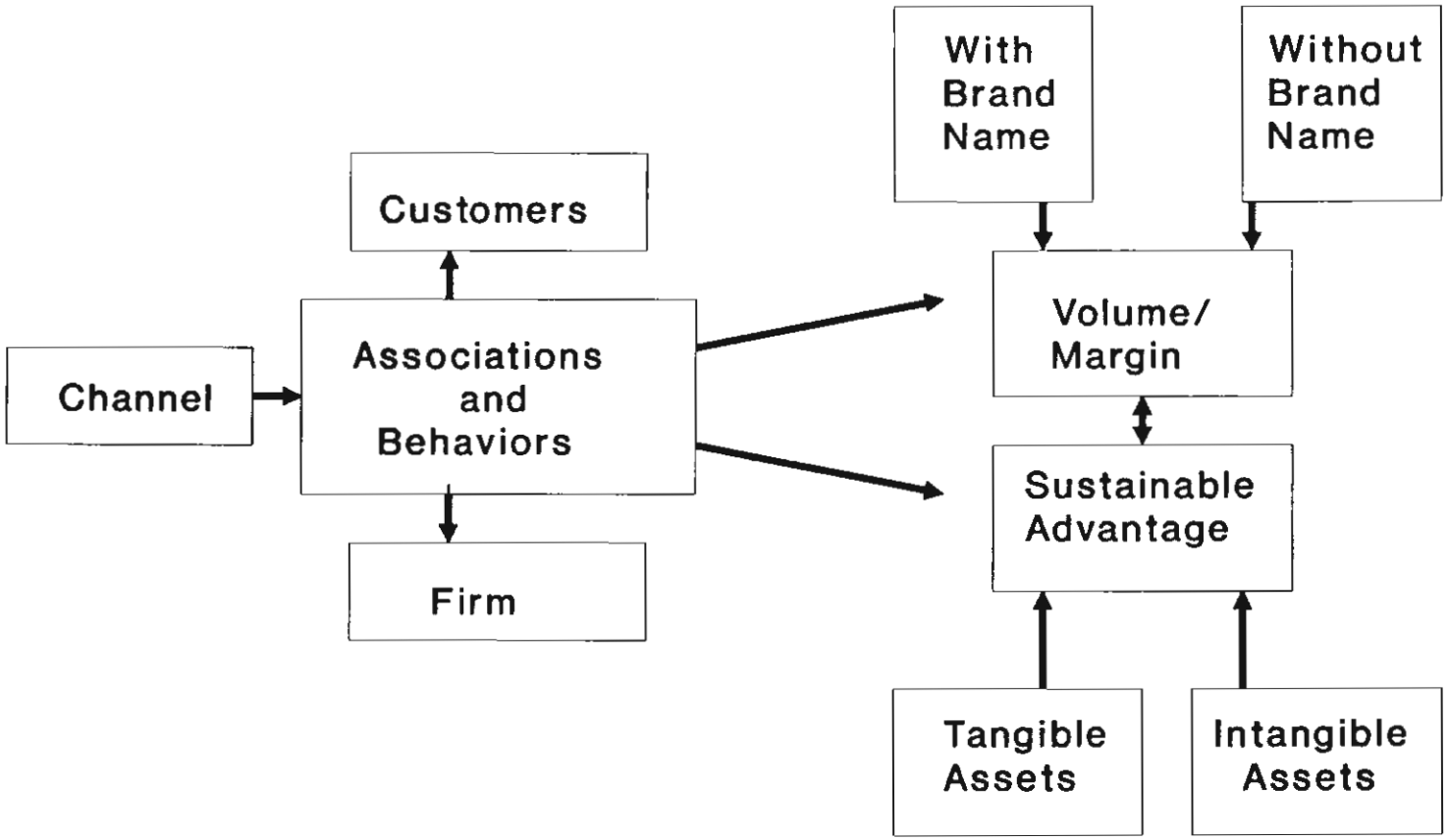
MEASURING BRAND EQUITY

Assignment of value to physical assets is replete with problems and typically involves subjectivity. Common methods of asset valuation include book value, replacement value, liquidation value, acquisition value, and discounted cash flow (Brasco, 1988). Valuation of intangibles such as brand associations is necessarily of even greater difficulty. Sometimes, as in the case of evaluation of the good will of a company or product line, it is possible to infer the value of an intangible from an offer price. Thus, by analogy, if one is able to assess the total value of a product to a prospective purchaser, the value of a brand would be the total value minus the value of the product without the brand. This approach is financially oriented, often at the firm or division level of analysis, and is most appropriate for merger, acquisition and strategic alliance decisions.

An alternative approach is to use some relative market index to infer brand equity. For example, Longman-Moran Analytics, Inc. have developed a copyrighted Brand Equity Index (BEI) which is a ratio of a brand's market share to the market share estimated to be correct given the brand's price elasticity relative to other brands (Moran, 1990). The latter index can be calculated from scanner data. The BEI approach is market-oriented, at the brand level of analysis, and combines all non-price related aspects of the brand. This approach is only applicable to mature consumer products, however, since accurate share and price cross-elasticity estimates of branded products are necessary for index calculation. Another characteristic of this approach is its focus on actual behavior (via market share) rather than on behavioral intentions.

Because of the widespread use of conjoint analysis to infer values consumers associate with the components of a product (Green and Srinivasan, 1990), it seems natural to explore the possibility of measuring some aspects of brand equity with this technique. If one considers brand to be an attribute of a product, several measures of brand equity come immediately to mind. First, one might

Exhibit 1
A Conceptual Model of Brand Equity



Modified from Chay, 1990

use estimates of partworth of the brand as direct measures that can take on positive or negative values relative to other brands or to other product attributes. Second, one could measure the importance of the brand attribute as a function of the difference in partworths between highest- and lowest-valued alternatives. In the latter case, if the conjoint study only includes a given brand and a nonadvertised or generic brand, the attribute importance becomes a measure of brand equity. Third, if price is an attribute traded-off in the conjoint task, it may be possible to determine how much money it would take to make people indifferent between branded and nonbranded items. Finally, it is possible to simulate market shares for branded and nonbranded alternatives and thus create a brand-equity index analogous to the Longman-Moran index, but based on preference shares rather than actual market shares. Price can also be introduced into the simulations to evaluate the prices necessary to have equivalent preference shares for branded and nonbranded products.

Obviously there are a number of limitations connected with the use of conjoint analysis for measurement of brand equity. For example, there are a variety of measurement models, data collection methods, and estimation algorithms that might be employed, each of which will have impact on the validity and reliability of the results. Nevertheless, we feel it is an important enough issue to begin the dialogue. Whatever measures are obtained will be valuable at least in a relative sense for comparing brand equity at different points in time or across market segments. Further, the conjoint approach obtains brand value assessments directly from consumers at the individual level of analysis. From a marketing strategy perspective, this is an important advantage. However, it is obtained at the cost of using preference or intention rather than actual behavior as the measurement basis.

STUDY DESCRIPTION

We viewed our task as providing an exploratory means of assessing the ability of conjoint analysis to measure brand equity. Therefore, when designing the research, we had three primary objectives: (1) to simplify the procedure, (2) to isolate brand effects, and (3) to minimize error. To simplify the procedure, we used a student population (business school students taking an introductory marketing class) and found a simple product that was familiar to them. We wanted to use a product class that was not gender specific, contained a small number of important brands, and had a relatively few easily described attributes. It was decided to use products that had been rated recently by a consumer rating service (Consumer Reports) in order to provide relevant anchor points and a validity check on the accuracy of self-reported importance measures. After selecting four possible products reflecting various decision-involvement categories (sunglasses, laptop computers, compact disc players, and automobiles), we conducted a preliminary study with marketing research students to ascertain familiarity with the products and specific brands. Also, following the recommendations of MacLachlan, *et al.* (1988), we selected attributes that described the products.

The preliminary study resulted in our selection of sunglasses for the conjoint task. The most important attributes (on the basis of self-explicated ratings) were brand, price, ultraviolet radiation protection, lens composition, lens color, frame composition, and whether or not the surface was mirrorlike. These attributes were corroborated by the consumer rating service.

Several factors dictated the conjoint methodology employed in the study. Since it was desirable to have all respondents evaluate the same sets of stimuli, we decided to use a full-profile card sort approach. To isolate the impact of a single brand, an experimental design was developed. Respondents were asked to evaluate profiles containing only a single familiar brand versus an unknown, nonadvertised brand. Two such sets of stimuli were created for the most familiar brands of sunglasses, Ray-Ban and Vuarnet.

Other steps were taken to minimize complexity and reduce error in the design. To avoid the upward importance bias in attributes with a greater number of levels (Wittink, 1990), all attributes contained the same number of levels (two). Also, both attribute order and stimuli presentation sequence were varied to control for any potential order effects. Attribute levels were selected that reflected the typical choices available to the students. In particular, prices were chosen that were not out of line with real-world choices, to avoid artificially inflating or deflating the importance of price in respondent judgments.

For seven attributes at two levels each, a minimum set of eight profiles is adequate for an orthogonal main effects design. We created two such sets as judgment tasks for our respondents, one for each of the two familiar brands of sunglasses. In the non-brand conditions, the product was described as a nonadvertised brand ("one you are not likely to have heard of"). In addition to sorting the profiles (printed on 4 x 6 inch cards) by preference, respondents rated the profiles on a likelihood of purchase scale ranging from 0% to 100% in 10% increments.

The study was administered in four discussion sections of a large introductory marketing class at a major public university. We obtained responses from 132 students (one student failed to complete the task for one of the brands, resulting in sample size of 131 in that condition).

The data were analyzed using Bretton-Clark's Conjoint Analyzer, which provided individual level OLS estimates of each attribute's partworths, normalized estimates of each attribute's importance, and simulated market shares (given a first choice model).

Hypotheses. Before describing the study findings, it is useful to anticipate various results. According to the preliminary study results, for this product class students say that brand is a very important consideration in their purchase choice. Therefore, we hypothesize that brand will have nonzero importance when measured as the normalized range of the partworths for this attribute. Furthermore, major brands will have partworths greater than nonadvertised brands. Hence,

H1. Brand will have nonzero importance.

H2. Partworths of major brands will exceed those of nonadvertised brands.

Preference shares for major brands ought to be higher than those for nonadvertised brands. Thus,

H3. The ratio of major brand to nonadvertised brand preference shares should exceed one.

It is likely that individual differences among respondents would yield different brand equity evaluations. For example. Kapferer and Laurent (1988) suggest that consumers have different brand sensitivities. In this regard, we hypothesize that our respondents will provide different levels of brand attribute importance depending on their likelihood of purchase of the product. Also, since people with higher likelihood of purchase are more likely to have become acquainted with the brands and their advertising, we would expect them to place greater value on the brand relative to other features. Thus,

H4. Higher brand-attribute importance will be found for people with high likelihood of purchase than for those with low likelihood of purchase.

FINDINGS

Exhibits 2 and 3 show the attribute importances and average partworths, respectively, obtained from the respondents for each of the conjoint tasks. Hypotheses H1 and H2 were strongly supported, since average importance of the brand attribute is considerably higher than zero and the partworths of the major brands were positive, while those of the nonadvertised brands were negative. Statistical tests of these hypotheses were not conducted, although it is clear that their null-hypothesis counterparts would have been rejected handily.

Simulation of preference shares using the first choice model resulted in preference shares for Ray-Ban of 75.2% and Vuarnet of 73.5%, relative to preference shares of 24.8% and 26.5% for the nonadvertised brands, respectively (see Exhibit 4). Hence, H3 was supported by the data. Again, it served little purpose to test the hypothesis statistically.

Respondents were divided into high and low purchase likelihood categories according to median scores on average purchase likelihood rated across all eight stimuli. This was done separately for each of the conjoint tasks. Exhibit 5 shows the result of estimating partworths and attribute importances separately for respondents with high purchase likelihoods versus low purchase likelihoods. It is clear that H4 is supported, since average brand importance was significantly higher for the high purchase likelihood group. Here standard errors of estimation were calculated and t-tests of difference between means were significant at the .001 level.

COMPUTING THE DOLLAR VALUE OF BRANDS

In principle, provided the utility-price relationship has constant slope, it is possible to compute the change in price required to make a consumer indifferent between branded and nonadvertised alternatives. One merely takes the difference in partworth values between brand and nonbrand conditions (in units of utils, say), and divides it by the slope of the utility-price curve (in utils per

Exhibit 2
Brand Equity Results:
H1: Brand Importance > 0

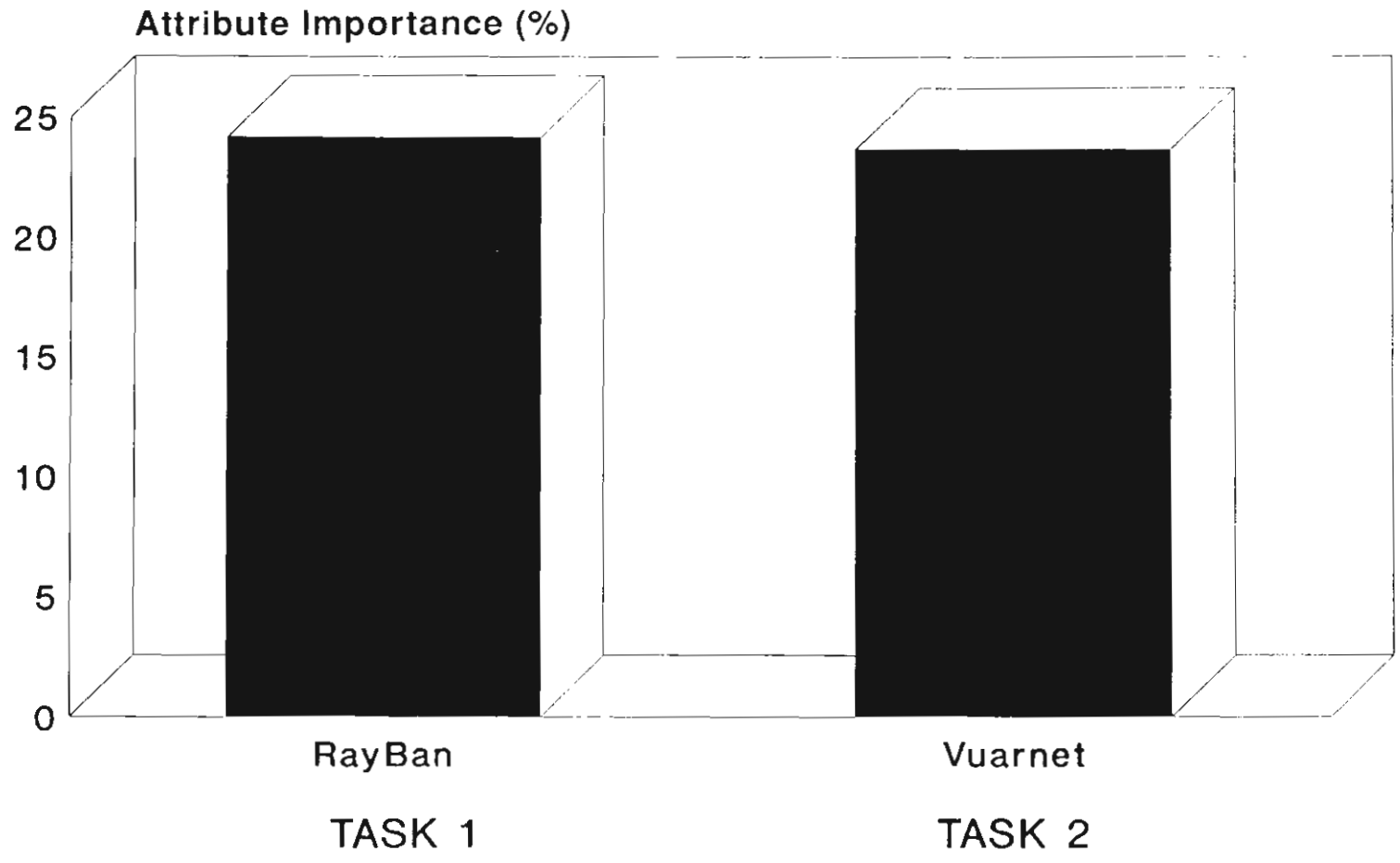


Exhibit 3
Brand Equity Results:
H2: Brand Partworth > Unadvertised Partworth

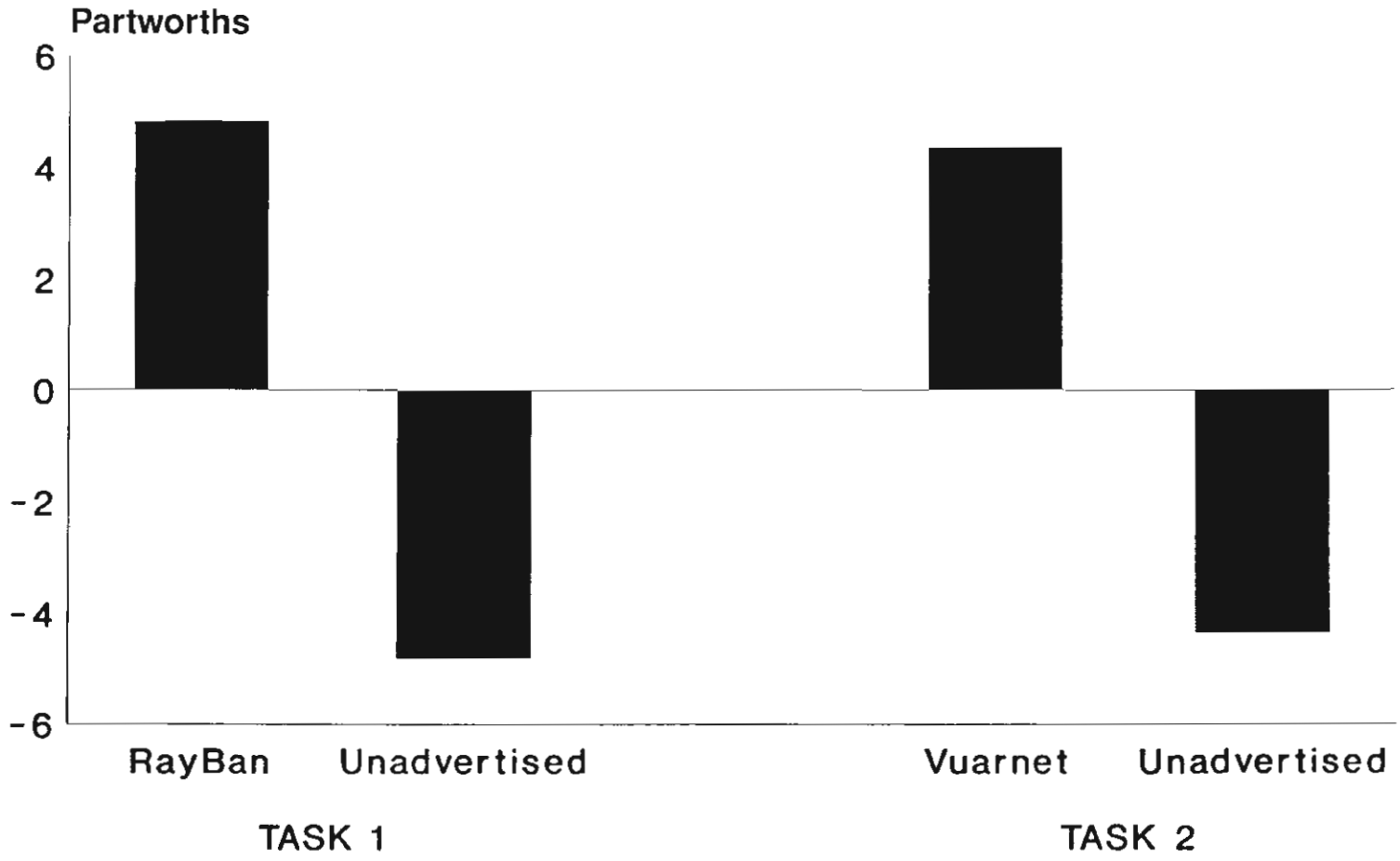


Exhibit 4
Brand Equity Results

H3:

$$\frac{\text{Major Brand Preference Share}}{\text{Unadv Brand Preference Share}} > 1.00$$

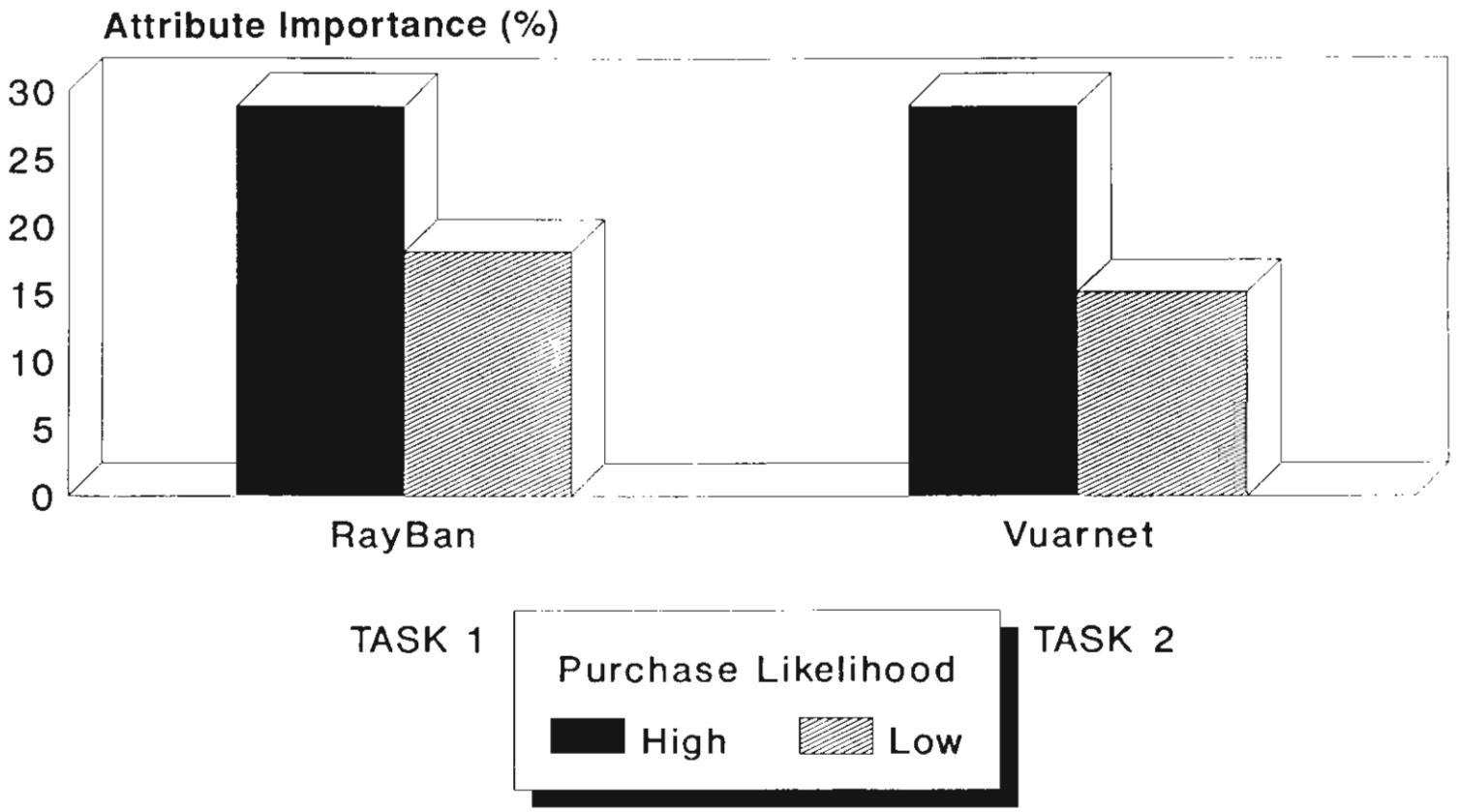
TASK 1

$$\frac{\text{RayBan Preference Share}}{\text{Unadvertised Preference Share}} = 3.03$$

TASK 2

$$\frac{\text{Vuarnet Preference Share}}{\text{Unadvertised Preference Share}} = 2.77$$

Exhibit 5
Brand Equity Results:
H4: If Hi Likelihood, Then Hi Importance



dollar). This is, of course, impossible when price is of no importance to the customer (one cannot divide by a zero value). Also, in some cases (particularly when price is viewed as a relatively unimportant attribute), the slope of the estimated utility-price curve for many respondents will be positive, suggesting higher utility for higher price. Finally, when the slope is estimated to be a very small number, the ratio becomes very unstable, resulting in extreme outliers.

In the present study, we only employed two levels of price, so we must assume constant slope of the utility-price curve. Almost half the utility price slopes were zero or positive, however. Ignoring the data from those respondents with zero or positive utility-price slopes, we found the median "dollar value" of Ray-Ban sunglasses to be $12.00 and the similar value for Vuarnet to be $7.45. That is, for example, if Ray-Ban were to charge a price premium of 34 percent (i.e., $12 over an assumed price of $35), the median respondent (in this reduced set of respondents) would be indifferent between Ray-Ban and the nonadvertised alternative.

Rather than eliminating respondent data, another way to deal with the zero or positive slope problem is to estimate the partworths with constraints on the price attribute (i.e., constraining the slope of the utility-price curve to be negative). Both LINMAP and CVA (Conjoint Value Analysis) have the ability to provide such constrained estimates. We recalculated the partworths using CVA with constraints on price, brand, and one other attribute (UV Protection) which had obvious directionality to its utility function. Unfortunately, the result was an even greater number of utility-price slopes near zero, making the ratio of brand utility to utility-price slope extremely unstable and unreliable.

In future studies, we suggest using a wider range of prices in the conjoint task, which will assure that people will view price as an important attribute. This is likely to make the "dollar value" of brand calculation more meaningful.

DISCUSSION AND MANAGERIAL IMPLICATIONS

The implications for the researcher and manager are several. Importantly, we have demonstrated that conjoint analysis is, in combination with the proper experimental design, a viable tool for measuring the value added by a brand name. It is possible to estimate partworths and brand importances for isolated brands within a simple product category. It is also possible to estimate preference share (a surrogate for market share) for a given brand over and above what would be estimated for the product if a customer did not know the brand. Further, we have shown that brand importance varies across subjects who have different probabilities of purchasing the product (that is, those who have different brand sensitivity). Finally, we have suggested a method for interpreting the brand value to customers in terms of the dollar change in price that would make them indifferent between the given brand and the nonadvertised alternative.

We do not argue that we have completely captured the brand equity concept with our conjoint measures, but we believe that we have measured important aspects of the concept.

Brand equity measurement can serve as a benchmark from which to assess effectiveness of marketing programs (for example, brand image advertising) over time and/or across market segments. In addition, brand equity measures can serve as leading indicators of consumer opinion. Other marketing and financial indicators such as market share, incremental cash flow, and the like often lag or follow important information for marketers, i.e., brand preference and purchase intention. Since they occur after the purchase, market share and financial measures are more accurate but retrospective. Conversely, consumer preference measures such as those generated by conjoint analysis are inherently less accurate but prospective. This latter feature is critical to development of strategic plans and the creation of sustainable competitive advantage implied by a strong brand franchise.

STUDY LIMITATIONS

The study was clearly exploratory, examining only one product, using a student sample, and employing a very simple conjoint design. On the other hand, we feel it was appropriate to begin this undertaking modestly and the product was carefully chosen to be relevant to the sampled population.

It is important to recognize that the specific values obtained for partworths and brand-attribute importance are completely dependent on the set of attributes and their ranges employed in the study. This implies that any practical use of the method requires that one be careful to include all critical attributes used for purchase decisions and include appropriate attribute levels for the product class. However, so long as the major use of the measures is period-to-period or segment-to-segment comparison, any reasonable set of attributes can be used to assess differences.

We have not determined which of the suggested measures (partworths, brand importance, preference share, or dollar value) is best for assessing brand equity; it will probably depend on the nature of the study.

A major limitation of the study is that we have used an orthogonal data-collection design in a situation where there are undoubtedly environmental correlations between some of the attributes (for example, brand and price). Again, we fall back on the need to begin simply and leave possible solutions to the problem for additional research. Also, it might not make much difference in the typical comparative situation where all else remains the same.

SUGGESTIONS FOR FURTHER RESEARCH

Since this was an exploratory study, there are many directions for follow-up research. For example, it would be desirable to examine additional products such as industrial products and those with greater complexity of attribute structure.

The concept of brand equity needs refinement. The questions of what elements constitute brand equity, their relative contribution to the perceived value added by the brand, and the relationship

among brand equity, brand loyalty, and brand sensitivity are directions for future research. Many methodological issues remain, such as the choice of appropriate conjoint measurement procedures (for example, respondent tasks, preference model, and analytical algorithm). Investigation needs to be made of the sensitivity of the measures to assumptions about the inclusion of all critical attributes and levels, orthogonality of design, and underlying preference model. If weakness in assumptions leads to unreliable or inaccurate results, methods need to be devised to overcome them.

The conjoint-measurement indices of brand equity need to be compared with other measures of brand equity for validation and contrast. To that end, it might be interesting to determine customer preferences for bundles of physically measured attributes, so that preference shares can be accurately gauged for products with and without brands.

There are difficulties associated with computing "dollar value" of the brands. In addition to the above-mentioned problems of zero or positive sloped utility-price functions, the latter function might not have constant slope. In such a case, it will be necessary to approach the problem piece-wise for different parts of the utility-price curve. Much research needs to be done to determine the best ways to introduce price into brand-equity measurement.

Perhaps the most interesting research would be to experiment with the ability of the measurement approach to capture changes in brand equity that occur as a result of marketing activities such as brand-image advertising. These could be conducted as either field investigations (say between isolated market segments) or as laboratory experiments.

REFERENCES

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

Allan, Elizabeth A. (1990), "Branding: The Sum of the Parts. . ." Presentation at the American Marketing Association Marketing Research Conference, Chicago, Illinois.

Brasco, Thomas C. (1988), " How Brand Names Are Valued for Acquisition," in *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Chay, Richard F. (1990), "Marketing Research: Banking on the Future of Brand Equity," Presentation at the American Marketing Association Marketing Research Conference, Chicago, Illinois.

Farquhar, Peter H. (1989), "Managing Brand Equity," *Marketing Research*, 1 (September), 24-33.

Farquhar, Peter H., Paul M. Herr, Russell H. Fazio (1990), "A Relational Model for Category

Extension of Brands," in M. E. Goldberg, G. Gorn, and R. W. Pollay, eds., *Advances in Consumer Research*, 17, 856-860.

Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments With Implications for Research and Practice," *Journal of Marketing*, 54 (October), 3-19.

Homa, Kenneth E. (1988), "Extending the Black & Decker Brand Name: The G.E. to Black & Decker Transition in Housewares," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Kapferer, Jean-Noel and Gilles Laurent (1988), "Consumer Brand Sensitivity: A Key to Managing and Measuring Brand Equity," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Louviere, Jordan and Richard Johnson (1988), "Measuring Brand Image with Conjoint Analysis and Choice Models," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

MacLachlan, Douglas L., Michael G. Mulhern, and Allan D. Shocker (1988), "Attribute Selection and Representation in Conjoint Analysis: Reliability and Validity Issues," *Sawtooth Software Conference Proceedings*, 37-49.

Moran, William T. (1990), private correspondence and company literature, Longman-Moran Analytics, Inc.

Park, C. Whan, Bernard J. Jaworski, and Deborah J. MacInnis (1986), "Strategic Brand Concept-Image Management," *Journal of Marketing*, 50 (October), 135-145.

Pecorella, Patricia A. (1988), "Keeping a Brand Strong with Strategic Marketing Decisions," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Shocker, Allan D. and Bart Weitz (1988), "A Perspective on Brand Equity: Principles and Issues," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Speirs, John R. (1988), "Thoughts on Brand Equity," in Lance Leuthesser, ed., *Defining, Measuring & Managing Brand Equity*, Marketing Science Institute Report No. 88-104, Cambridge, Massachusetts.

Tauber, Edward M. (1981), "Brand Franchise Extension: New Product Benefits from Existing Brand Names," *Business Horizons*, 24, 36-41

Wittink, Dick R. (1990), "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions," in *Advanced Research Techniques Forum Proceedings*, American Marketing Association.

Comment on MacLachlan and Mulhern

Thomas L. Pilon
IntelliQuest, Inc.

GENERAL COMMENTS

Researching approaches to measuring brand equity is timely and important. As documented in the paper, there has been considerable interest in this topic in the past two years. As an additional indicator of interest, I have received several requests from clients to help them research this topic. The request is often phrased as "given my brand name, how much more can I charge than (less well-known) brand A?" or "given my brand name, how much less do I have to charge than (more well-known) brand B?" In the present recessionary economy, corporations are looking to redeem some of the investment that they have made in establishing a strong brand name. Additionally, in an economic downturn pricing becomes more salient.

Measuring brand equity is an appropriate application of conjoint analysis. There are other equally promising ways of approaching the brand equity measurement problem such as more general choice modeling and various types of perceptual mapping and mapping simulators. Since, in my opinion, conjoint analysis is a viable approach for measuring brand equity, I will limit my comments to this approach.

SPECIFIC COMMENTS

The authors state in the "Study Limitations" section of their paper:

> A major limitation of the study is that we have used an orthogonal data-collection design in a situation where there are undoubtedly environmental correlations between some of the attributes (e.g. brand and price). Again, we fall back on the need to begin simply and leave possible solutions to the problem for additional research.

While complimenting the authors on their commendable work in documenting this first step in measuring brand equity with conjoint analysis, and with complete recognition of the need to limit the depth and breadth of any research endeavor, I will comment on and suggest some possible ways around the limitation described above.

Although very little has appeared in the literature, there has been much debate in the research community about the inclusion of (whether or not as well as how to include) brand name as an attribute in conjoint analysis.

One of the issues is that a respondent's perception of a brand may cause certain profiles to seem inconsistent and confusing in the mind of the respondent. For example, a respondent may not be aware that Ray Ban makes sunglasses with mirrored glass. When this respondent thinks of Ray Ban, he thinks of their classic aviator-style green lenses. When confronted with a profile in a conjoint study that has both the "Ray Ban" and "mirrored glass" levels of their respective attributes, it may confuse the respondent. Even though the respondent may have been instructed to "just think about the brand name and not the specific characteristics of products with that brand name," it is often difficult or impossible for a respondent to do that, since a brand name is such a potent and multi-faceted phenomenon.

A seemingly simple way of getting around this problem would be not to allow "Ray Ban" and "mirrored glass" to appear together in the same profile. However, this "prohibited pair" technique is not appropriate when there are many brand(s) and/or attribute(s) linkages. This technique also cannot get around the fact that different respondents have different (some incorrect) perceptions about brands. Finally, there are problems associated with using this technique to prohibit combinations that may later appear together in a simulator.

A second brand inclusion issue centers around using the price attribute to compute the dollar value of brands. Typically several other attributes are included in a study along with brand and price. The price attribute reflects the average price sensitivity (or elasticity) of all the attributes in the study. One would expect that each attribute (including brand name) would have a unique and sometimes quite different price sensitivity. If one uses the price attribute to compute the dollar value of a brand, it is likely that the results will be at least somewhat misleading since the average price sensitivity is being used instead of a brand-specific price sensitivity.

Along with a few research colleagues and clients, I have been working on a way to circumvent both of these problems. Since a complete description of this method would be a paper in itself, I will provide a summary overview. The method involves splitting the attributes into two separate conjoint measurement tasks. One task would include all attributes except brand and price. The second task includes brand and price along with at least one attribute (or a composite attribute) from the first task. Each of these tasks is conducted separately by each respondent. After estimating the utilities for each conjoint separately, the resulting utilities are rescaled or "bridged" together so that the attributes from the two separate conjoint tasks can be made comparable and combined for input into one simulator. Using Sawtooth Software's Conjoint Value Analysis (CVA) software is a particularly effective way of conducting the second conjoint task when measuring the dollar value of a brand is an issue.

A final, very minor, comment about the paper is that Hypothesis H2 and H3 are somewhat redundant and I am not sure that both of them needed to be included. If all other things are held equal in the simulations (as was the case here), H3 would always yield results identical to H2 (either both rejected or both not rejected).

In summary, I think this area of brand equity measurement is tremendously important and I am eager to see future work on this topic.

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

I. PREFACE

State and provincial lotteries are entrusted with the role of producing revenue for their respective jurisdictions and increasing it from year to year, just as private corporations must do. To accomplish this, lotteries must continually advertise and offer new and exciting lottery products.

The characteristics of lottery products can differ greatly by game and market segment. Lottery game attributes may include:

Game theme
Method of play
Prize structure
Number of drawings
Price of ticket
Bonus options
Scratch vs. On-line
Color of ticket
Odds of winning

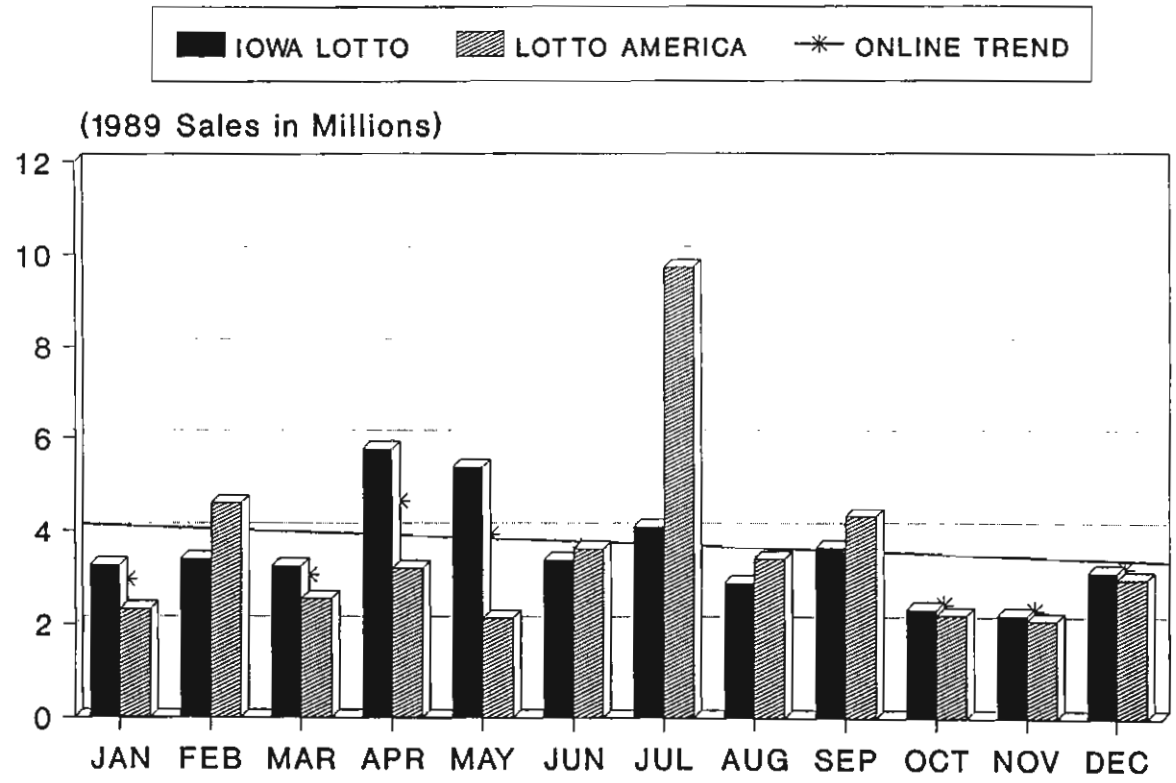
Given the cost to develop and produce each game, coupled with the limited number of games the lottery may offer at any one time, it is the lottery's challenge to maximize the revenue by optimizing the games to satisfy the population's preferences.

The Iowa Lottery commissioned a conjoint analysis study to evaluate proposed new lottery games and game options. The following presents a summary of the study.

II. THE PROBLEM

In the fall of 1989, the Iowa Lotto game and the Lotto America game had experienced declines in both their player bases and in revenues.

LOTTO SALES



Kopel Research Group, Inc.

A recently completed baseline study isolated "jackpot sensitivity" and "perceptions regarding the ability to win the games" among the reasons for the games' declines.

The Iowa Lottery needed to evaluate its options and investigate new products which could be introduced to turn around the sales decline.

III. FOCUS OF THE RESEARCH

The research objective was to evaluate the "likelihood of purchase" of a variety of proposed on-line game concepts. Each proposed new game offering was to be evaluated for its market potential.

As reference points from which to evaluate the proposed game concepts, both the existing Iowa Lotto and Lotto America games were to be included as benchmarks.

The model also needed to be able to isolate respondent demographics and current player behavior.

IV. METHODOLOGY

The Adaptive Conjoint Analysis (ACA) data collection and statistical technique was selected because it is able to identify which attributes affect a lottery purchase decision (prior to this technique it was virtually impossible). ACA offers the ability to create a preference score (utility value) for each of the game variables produced by each respondent. Further, by using the ACA simulator, the preference and strength of various choices is measurable. Also a "likelihood of purchase" factor may be calculated by respondent subgroup.

In this case, the ACA interview was conducted in a mall intercept format. This format allowed the information to be presented orally and visually so that misunderstandings about jargon and concepts would be kept to a minimum. The lottery industry uses terms such as:

Prize structure
Jackpot
Payout
Odds
Bonus balls
Parimutuel
Box, Split, Wheel
Keno

The mall intercept format let the Iowa Lottery offer the one-on-one detailed interactive process that the survey required.

We took great care, for example, to make sure that the prize structure comparisons were perceived as a function of the odds of winning. To do this we needed to present the prize structure as a function of the number of winners with equal dollars paid out in prizes. As a result, each prize structure/odds of winning "pyramid" became a measurable variable.

To make the concepts understandable and realistic, we developed color-coded "flash cards" by attribute type. These "flash cards" helped the respondent physically conceptualize the game attributes. Respondents had the capability to physically make new games by sorting the attributes according to their preferences. Respondents were allowed to answer the questions at their own pace with no time constraints.

The Adaptive Conjoint Analysis interview module runs within Ci2 (the Ci2 System for Computer Interviewing) on the personal computer. As part of the game design testing we developed an extensive questionnaire outside of the ACA module but within the Ci2 portion of the interview. In this module we were able to collect information about the respondents' game playing habits and attitudes toward the lottery.

Further, the lottery had several specific new game options and marketing plans they were evaluating as add-on features to the existing lottery products. In this Ci2 section we were able to ask the respondents about which options they preferred as well as reasons for preference of various game features and options, in an open-ended format.

The fielding was performed in 3 metropolitan locations in the state. In brief, we screened our targeted sample to attain the desired quotas. All subjects were over 18 and played the on-line lottery games either frequently (5+ times per month), infrequently (1-4 times per month) or were lapsed players (had played but not in the past month).

V. SCOPE OF STUDY

Integral to the proposed on-line games were the method of play and prize structure. Since each game had a wide variety of prize payouts, the prize structure could not be separated from the game as a separate attribute ("Prize structure" and "method of play" are factors that can be separated when testing the instant scratch tickets).

The lottery tested 8 potential on-line games, 4 different play frequencies and 3 different price points as follows:

I. PROPOSED GAMES

A. Choose 1 Card (from each suit)

This game was designed around a card game theme and required a player to pick 4 cards, one from each suit. The player won $10,000 for matching all 4 cards to the lottery's draw, $100 for matching 3 cards, $10 for matching 2 cards and a free ticket for matching 1 card.

B. Choose 4 Card (from 52)

This game is a variation of the previous game. The variation here is that a player can pick any 4 cards from the deck of 52 cards. The player would win $20,000 for matching all 4 cards, $100 for matching 3 cards, $2 for matching 2 cards and a free ticket for matching 1 card.

C. Pick 3

This game was a simple pick 3 game, played in many of the Eastern states, and is derived from an illegal street game. The object of this game is to choose 3 numbers from 0 to 9, the player wins $500 for matching all 3 numbers in exact order and $80 for matching the 3 numbers in any order.

D. Keno 7/20/80

This game is a lottery derivative of the Keno game played in the casinos in Las Vegas. In this game, players pick 7 numbers from a field of 80 numbers. The lottery picks 20 numbers from the field of 80 numbers. The player wins $10,000 for matching any 7 numbers, $500 for matching any 6 numbers, $10 for matching any 5 numbers and a free ticket for matching any 4 numbers.

E. Keno 10/20/80

This game is a variation of the 7/20/80 Keno game above. In this case, players pick 10 numbers from a field of 80 numbers. The lottery picks 20 numbers from the field of 80 numbers. The player wins $100,000 for matching any 10 numbers, $10,000 for matching any 9, $1,000 for matching any 8, $100 for matching any 7, $10 for matching any 6, $2 for matching any 5 and $2 for matching 0 numbers.

F. Sports Action

This is a sports pool lottery game patterned after the weekly illegal football cards widely available during football season. In this game, a player can pick from 4 to 14 games from the current week's football lineup. In order to win the player has to correctly pick the winners of all games he or she chooses to play including adding in the "point spread" stated on the card. Prize payout is "pari-mutuel" and is based on the number of games played. Approximate prize payout ranges from $8, 192 for picking all 14 games correctly to $8 for picking 4 games correctly.

G. Iowa Lotto

This game is currently played in Iowa. The player picks 6 of 39 numbers. If the player matches all 6 numbers the player wins a share of the jackpot which averages $2,000,000. Matching 5 numbers pays $600, and matching 4 numbers pays $30.

H. Lotto America

This game is a Lotto game that included 11 states at the time of the study. Players pick 6 numbers of 54 numbers and have 2 plays for $1. The payout is pari-mutuel with the jackpot averaging $5 - $10 million dollars, match 5 paying about $1,000 and match 4 about $40.

II. PLAY FREQUENCY

A. Play Daily
B. Play 2 times per week
C. Play 3 times per week
D. Play Weekly

III. PRICE

A. $1 per Play (Prize structure as shown on cards)
B. 2 plays per $1 (Prize payout is 1/2 of that shown)
C. 3 plays per $1 (Prize payout is 1/3 of that shown)

VI. RESULTS

ACA develops utility values as a means of weighting the relative value of attributes toward the purchase decision process. These are numerical scores which are derived mathematically and define a level of preference for one attribute as compared with another.

These values are normalized to permit comparison and aggregation across various combinations of attributes and subjects. It is the product of combined utility values which generates the "Likelihood of Purchase and "First Choice" models.

I. GAME THEME

The current lottery games, Iowa Lotto and Lotto America, were the highest scoring themes among all tested. This indicated a high acceptance among those surveyed for the existing games, and indicated a benchmark by which all other concepts could be measured.

MALES

The frequent males scored Iowa Lotto and Lotto America as the most preferred lottery games. Sports Action was solidly in third place among frequent male players. The Keno 10/20/80 game was a distant fourth with the other Keno 7/20/80 game close behind. Pick 3 was in sixth place. The two card games, liked by women, tied for last place among the frequent males.

After scoring the existing games the highest, the infrequent males showed little preference among other concepts, including Sports Action which was well liked by the frequent male players.

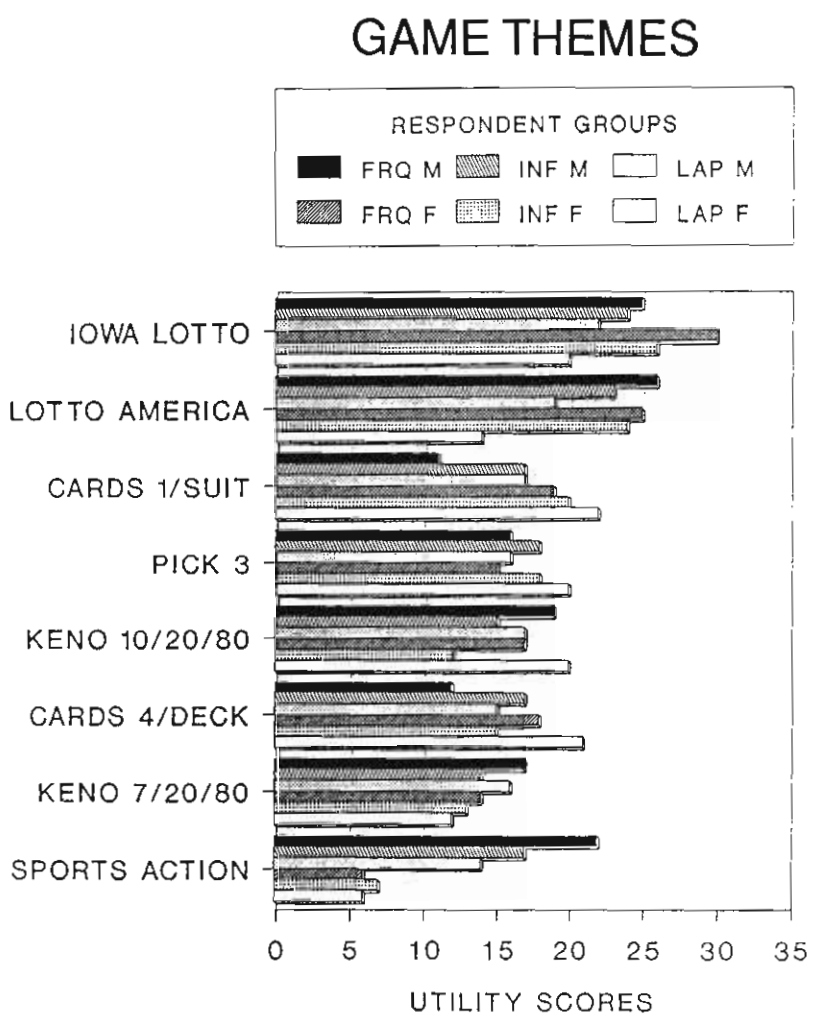
The lapsed males preferred Iowa Lotto to Lotto America, but showed no clear preference after that. Sports Action actually scored lowest of all concepts among lapsed males.

FEMALES

The frequent females clearly preferred Iowa Lotto to Lotto America. In turn, Lotto America was substantially preferred to the next closest games, the 2 card games. Keno 10/20/80 was liked fifth, followed by Pick 3 and Keno 7/20/80. Sports Action was a distant last.

The infrequent females also preferred Iowa Lotto and Lotto America. They clearly liked the Pick 1 from each Suit third, the Pick 3 Game fourth and the Pick 4 From The Deck card game fifth. Both Keno Games followed closely with Sports Action a distant last.

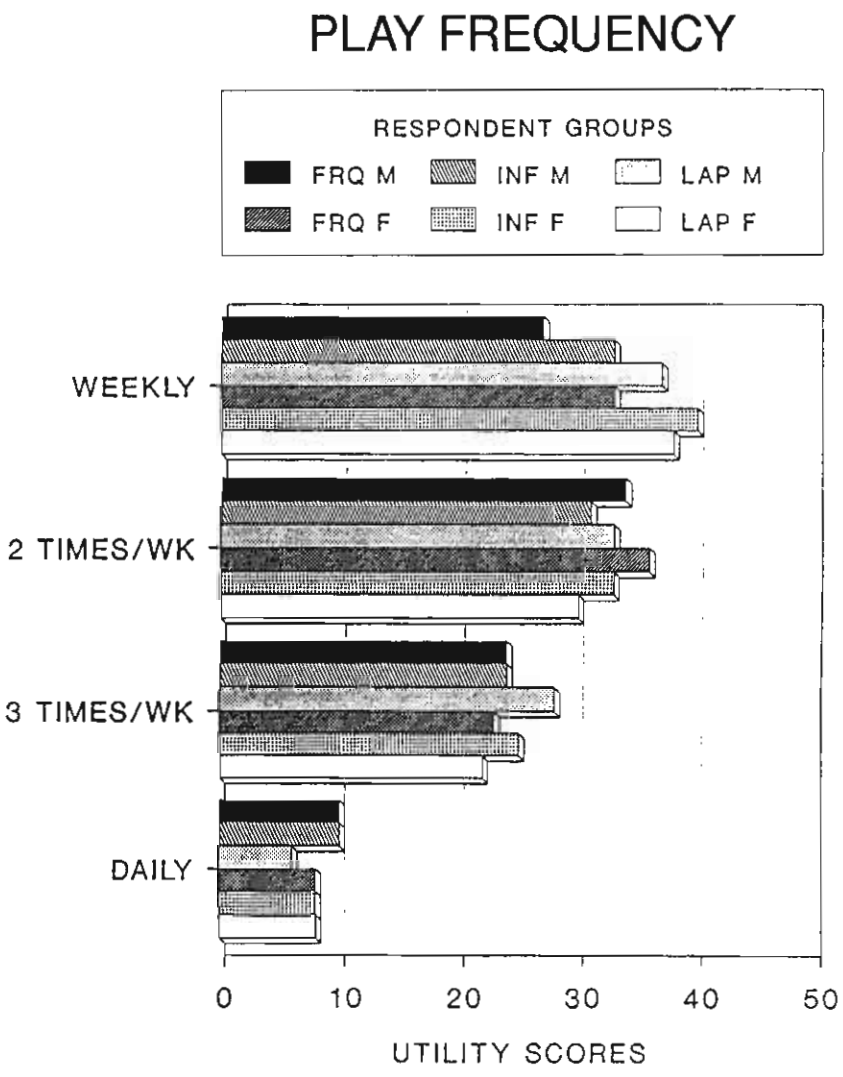
The lapsed females scored Pick 1 Card From A Suit first, followed closely by Pick 4 Cards From A Deck, Pick 3, and Keno 10/20/80. Next to last was Keno 7/20/80 and last was Sports Action.



II. PLAY FREQUENCY

There was very little difference between men and women or among frequent, infrequent and lapsed players in terms of play frequency preferences.

* Weekly and 2 times per week were the most preferred draw frequencies.
* Respondents were less enthusiastic about playing 3 times per week.
* A strong negative reaction was given about daily lottery play, among all respondents.



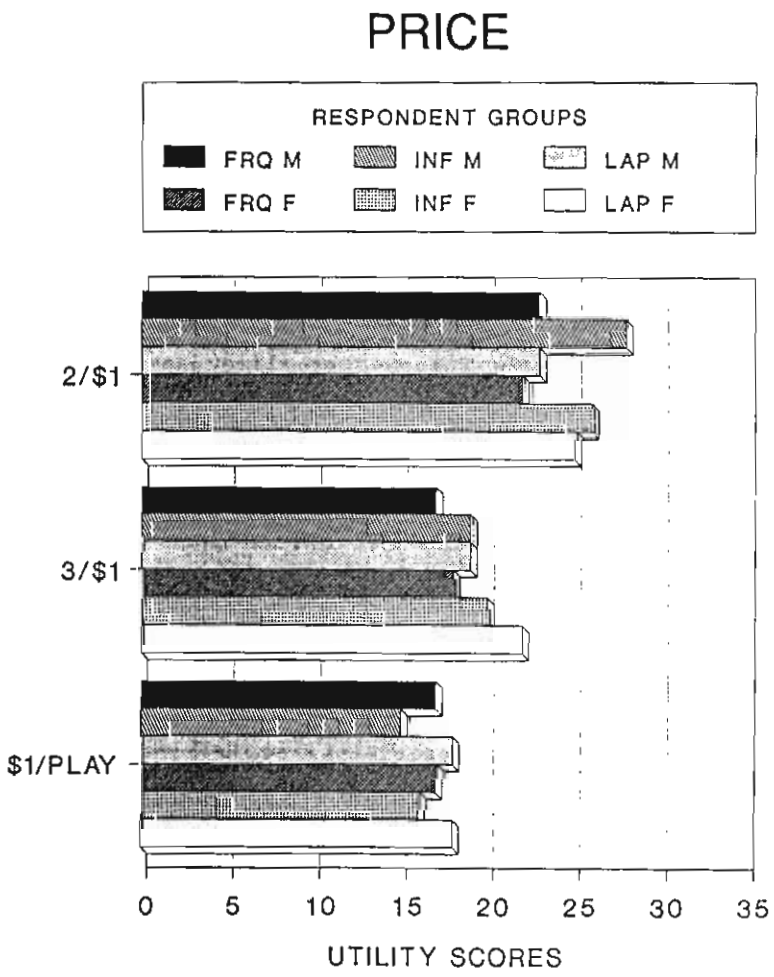
III. PRICE PER PLAY

The respondents were told that they could have a choice on the price of the ticket but also that it would affect the prize payout proportionally. In other words, 2 plays for a dollar pays out 50% of $1 per play, and 3 plays for $1 pays out 33% of $1 per play as shown on the cards.

With this in mind respondents chose:

* 2 plays per $1 as the clear favorite,
* followed by 3 plays for $1 second, and;
* $1 per play last.

This distribution was consistent among all player groups.



VII. CONCEPT GAMES — FOLLOW UP QUESTIONS

To validate findings from the ACA interview, respondents were asked follow-up questions in the Ci2 portion. This section, when evaluated in context with ACA, played a major part in shaping the appropriate new game and game option strategies.

HIGHLIGHTS FROM Ci2

The current lottery games, **Iowa Lotto** and **Lotto America**, were the highest scoring themes among all tested. This indicated that these games were a benchmark by which all other concepts are to be measured.

Among the new concepts tested, the **Pick 1 From Each Suit** card game scored the highest among women.

Sports Action was a game that garnered strong positive and negative reaction. Of the proposed games, frequent males preferred the **Sports Action** game the most. Yet, this game was also liked least by all women and some men.

PICK 3, a game liked by many, also had high least-liked ratings.

BOTH CARD GAMES had low "like least" ratings. The **KENO 10/20/80** game performed slightly better than the **7/20/80** game, the former scoring fairly high for likelihood of play among frequent males.

Weekly and 2 times per week were the most preferred draw frequencies. There was very little difference between men and women or among frequent, infrequent and lapsed players in terms of play frequency preferences. All respondent sub-groups were consistently against a daily game.

The favorite price point was 2 plays per $1.

VIII. CONCLUSION

A new game needed to support and enhance Iowa Lotto's market position, not offer more alternatives. Conjoint research indicated that daily games and games that appealed to highly segmented player groups would not be prudent at this time. Iowa Lotto was generally well liked, however; it was in a competitive situation with its sister Lotto America game. Therefore, instead of producing a new and different game at this time, it was decided to offer an option on Iowa Lotto to help differentiate it.

The lottery applied the utility scores to formulate "Wildcard" (the card game themes tested had low least liked utility scores). The Wildcard game is played as an option that costs an additional $1 for Iowa Lotto players (it was decided that the difference in revenue for $1 per game versus $.50 offset the more positive utility score 2/$1 generated). A playing card number and suit (1 of 52) is pre-printed on the Iowa Lotto Ticket. The player decides before the ticket is printed whether or not to activate it. When activated, the ticket shows that Wildcard is activated. If it is not activated, the Wildcard is still shown, but is marked as inactivated. (Its presence on the ticket creates a psychological motivation to play it the next time.) During each Iowa Lotto drawing, a separate drawing is held for Wildcard.

The subsequent baseline results were impressive. Approximately 1/4 of Iowa Lotto tickets have Wildcard activated. Iowa Lotto sales increased from an average of $800,000 per week to $1.2 million per week including Wildcard.

Further, with the new positioning of Iowa Lotto, the "current" player base for the Iowa Lottery increased. The increased player base had an increase effect on the sales of all lottery games. Since Iowa Lotto is the most widely played game in Iowa, as the current player base increased for Iowa Lotto, players "cross-merchandised" and played other games as well. Further, subsequent baseline research indicated that those playing multiple games tended to spend more on each game played, compared with those only playing one game.

REASONS FOR ACA SUCCESS

ACA allowed us to run simulations on small market segments. At this point in the lottery's maturity, it was recognized that it needed to maximize sales among the strongest core of players; not fragment the player base and alienate the core player base. Thus games such as Sports Action were seen as too risky at this time.

The use of "Flash Cards" so that respondents could visualize games on the desk worked quite well; many of the complex concepts needed to be laid out in front of the respondents.

We maximized the use of Ci2 as a check on the answers. This allowed us to be able to read more into the responses than we otherwise could have. For example, we asked questions on such concepts as Bonus Balls, annuity versus cash payout, and optional add-on prize matrices for additional costs.

WHAT WOULD WE DO DIFFERENTLY NEXT TIME?

This research incorporated a wide variety of game types and therefore the precision was compromised where we tried to compare apples with oranges.

If, for example, we knew that the next game was going to be a Keno game, we could maximize our effort toward optimizing the Keno game and its component parts.

We did not have the luxury of conducting a baseline study prior to the conjoint research. By conducting a preliminary baseline study prior to carrying on a conjoint study, we could more clearly identify the needs, and focus more toward optimizing the future products.

Comment on Kopel and Kever

Keith Chrzan
Walker: Research and Analysis

The Iowa Lottery Case Study turned out the way we all hope our conjoints will: it allowed the design of a new product that succeeded in its own right and increased the size of the overall market as well. Moreover, the data lend themselves to further uses; most importantly, the authors have ACA-generated utilities for the new product and two previously existing products (the old Iowa Lottery and Lotto America) and actual sales data for the three products as well. This allows them to test the predictive validity of their conjoint model by checking the correlation between the modeled utility of the three products and their actual sales.

One thing that leaps out of the utility data is respondents' relatively higher utility for familiar products than for unfamiliar ones. Lottery formats like the Iowa Lottery and Lotto America have high utilities and are familiar to respondents. Similarly, the Sports Action game format gets a high utility from male respondents and is something many of them will have known from firsthand experience. Do respondents just respond more positively to familiar formats or have lotteries already discovered the best formats? If one of the Keno games had the advertising support of the Iowa Lottery, might it have scored as highly? Even if familiarity affects utility, is this a problem: will not the same thing occur out in the marketplace? These questions may deserve further attention, because if conjoint analyses can be biased by respondents' familiarity with some attributes or levels and not with others, they may not treat truly novel ideas fairly.

My final point is to iterate some of the points extensively covered in Louviere and Gaeth's paper in the *1988 Sawtooth Software Conference Proceedings*: as with many conjoint studies, this study could probably have been better modeled as a choice task. While the reader is referred to the Louviere and Gaeth paper for details, the choice study could have been performed with as few as 27 choice sets consisting of two game options and a no play option each.

1) One of the major drawbacks of choice studies is that they use aggregate, not individual level information. The Iowa Lotto Case Study did not rely upon disaggregate data, so this potential problem would not arise.

2) One of the many strengths of choice studies is their ability to model the "no purchase" or in this case the "no play" option. Respondents could have been given, say, 100 imaginary dollars to apportion among two game options and the no play option in each choice set. With the utilities resulting from the choice analysis one could model the utility of any game composed of the attributes and levels in the study and of the no play option. If changes in product mix are likely to result in cannibalization (or, as in this case, market expansion) these can be predicted, rather than serendipitous outcomes.

REFERENCE

Louviere, Jordan J. and Gary J. Gaeth (1988), "A Comparison of Rating and Choice Responses in Conjoint Tasks," *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software, 59-73.

USING CONJOINT ANALYSIS FOR PRICE OPTIMIZATION

Richard D. Smallwood
Applied Decision Analysis, Inc.

INTRODUCTION

One of the dramatic lessons of the collapse of the communist economic system in the Soviet Union and Eastern Europe is the importance of the marketplace as the final judge of pricing decisions. In the expanded international economy resulting from these economic and political upheavals, pricing decisions will become even more important. Understanding the sensitivity of the market to different pricing policies is essential to successful participation in the high stakes game of international marketing.

To the manufacturer with multiple products in a complex market, setting the prices of all the products in the portfolio can be an imposing task, particularly if the products compete with one another. Moreover, any adjustments in the price of the product portfolio will undoubtedly cause changes to the competitors' prices. In the face of such complexity, it is not surprising that many companies choose to use *ad hoc* methods and to set each product price as a separate decision.

There are of course analytical aids that can help with these complex decisions. To use them, we need to understand how the sales of a product, or portfolio of products, will depend on the prices of the products. The specification of this demand function, particularly as it depends on the interactions among many products, is often the most difficult part of an analytical approach to deriving an optimal pricing policy.

This paper will demonstrate how conjoint analysis can help in the specification of the multi-product demand function. This leads to the development of techniques for deriving the optimal pricing policy for a portfolio of products within a competitive environment.

FINDING THE DEMAND FUNCTION

Let's imagine a situation in which our client has one or more products competing for sales with competitors' products in the market. Suppose further that we have conducted a conjoint analysis survey over a sample of customers in this market. This will produce for each respondent in the survey the following:

- The utility for price, $u_i(p)$, where the index i refers to the ith respondent
- The utility of all the other attributes for each product in the market; let's call this utility for the ith respondent and jth product u_{ij}

- The number of customers represented by each respondent; w_i will denote this "weight" for the ith respondent.

Notice that the first two quantities are measurements resulting from the conjoint analysis survey, while the third is determined by the sampling plan used to recruit respondents.

With these three quantities it is possible to estimate a demand function over the products in the market. To do so we need an additional model to describe how respondents' ultimate choice of products depends on the measured utilities. The two most common models for this purpose are the logit and probit choice models. If we use the logit choice model, the demand function for the products in the market becomes:

$$D(k) = \sum_i w_i \, Exp[u_i(p_k)+u_{ik}]/\sum_j Exp[u_i(p_j)+u_{ij}] \qquad (1)$$

where D(k) is the total demand for the kth product, p_k is its price, and Exp[.] denotes the exponential function.

The purpose of this equation is to illustrate the intuitive notion that the utilities resulting from a properly administered conjoint analysis can be used to form a demand function over the products within a market.

ESTIMATING SENSITIVITIES TO PRICE

The availability of an analytic demand function is a powerful tool for pricing analysis. For example, with the demand function it is now possible to calculate each product's elasticity with respect to price. Since the elasticity of demand to price is just the percentage change in demand per percentage change in price, the elasticity, E(k), for the kth product is just:

$$E(k) = p_k[dD(k)/dp_k]/D(k)$$

where the quantity in brackets is the derivative of the demand for the kth product with respect to its price. This can be calculated analytically from Eq. 1 above, and so it is possible to provide price elasticities as a direct output of the conjoint analysis.

Price elasticities are an important source of information about how the market is likely to react to changes in price. They can be used to determine which products will have large changes in demand and which will have relatively small changes for a fixed percentage change in price.

This idea can be taken one step further. Suppose that our client has several products in the market and wants to estimate how changing the price of one will affect the demand of another. This cross-pricing effect can be represented as a cross-elasticity, defined as the percentage change in the

demand of one product per percentage change in the price of another. If E(k|j) is the cross-elasticity of the demand of the kth product to the price of the jth product, then we have:

$$E(k|j) = p_j [dD(k)/dp_j]/D(k)$$

where the bracketed term is the derivative of the demand of the kth product with respect to the price of the jth product. These cross-elasticities can be calculated analytically from the demand function in Eq. 1 and so can be reported as a direct output of the conjoint analysis study.

In practice it is often more intuitively appealing to report out the derivatives, $dD(k)/dp_j$, themselves rather than the cross-elasticities. The derivatives represent the change in demand for a change in the price of a single product, and so have the following physical interpretation. Suppose we raise the price of one product and observe that its demand has decreased by some number of units. The derivatives of demand with respect to its price illustrate how the product's losses will be reallocated by the market to the other products.

FINDING THE OPTIMAL PRICE

The availability of a multi-product demand function along with elasticities and cross-elasticities raises the possibility of calculating an optimal price. Let's consider the single product situation first. If the variable cost of the product is c, then the total variable profits for the product are:

$$V = D\,(p - c) \qquad (2)$$

and so we want to know the price that maximizes this quantity. The change in variable profits per change in price is just:

$$dV/dp = [dD/dp](p - c) + D \qquad (3)$$

and this quantity can be calculated analytically from the demand function in Eq.1.

Finding the optimal price is just a question of finding the price that causes the expression in Eq. 3 to equal zero. There are many sophisticated techniques in the arsenal of operations research for solving this type of problem.

The problem becomes more interesting and somewhat more difficult if we seek to optimize the prices of a portfolio of products. In this case the variable profit expression in Eq. 2 must be expanded to include the sum over all the products. Techniques are equally available for solving the multi-product version of Eq. 3 to find the optimal portfolio of prices.

INCLUDING COMPETITIVE EFFECTS

The discussion in the preceding section assumed implicitly that the competition will make no changes to its prices in response to our attempts to optimize prices. This is, of course, not realistic. To include competitive effects in the optimization of prices requires some explicit statement about how the competition will respond. For example, if we raise our prices by 10% will the competition also raise theirs by 10%, or by only 5%, or not at all?

The best approach is to build a simple model of how the competition will respond. This competitive response model can then be incorporated into the optimization of prices. If there is uncertainty about the competitive response, then this can be included in the model so that a probabilistic competition model results.

Once the competitive response model has been constructed, sensitivity analyses can be conducted over the uncertain elements of the model. In most cases, there will be a few key factors about competitive responses that significantly affect the optimal price settings. Once identified, attention can be focused on resolving the uncertainty about these key factors.

LIMITATIONS OF THE APPROACH

The optimization scheme proposed above is based on several simplifying assumptions, each of which can be relaxed at the expense of increased complexity. Some of these assumptions are discussed below:

Logit choice model — The logit choice model in Eq.1 has the obvious advantage of analytical simplicity. Other choice models such as the multivariate probit could be used instead, although the probit model in particular requires considerably more complex numerical computations.

Preference model — Regardless of the form used, the demand function of Eq.1 only describes customer *preferences* for products rather than actual sales. It is quite feasible to add a separate module to the above structure that describes those aspects of the market not contained in simple preferences. Examples of issues that might be included are the distribution network, manufacturing capacity constraints, sales force coverage and effectiveness, and customer reluctance to change.

Linear cost model — The formulation of the optimal pricing problem above assumes in Eqs. 2 and 3 that the incremental cost of each additional product sold is constant. If there are significant economies of scale over the range of demand considered in the demand function, then the cost model can be modified to include these effects.

Homogeneous market — The formulation of the demand function in Eq. 1 assumes implicitly that the size of the market to be distributed among the products is constant. As the prices of products change it is reasonable to expect that customers will enter or leave the market in response to the overall price

level. To include this effect requires a separate model of the total market size and its dependence on overall price levels.

DESIGNING THE SURVEY

If the utilities from a conjoint analysis survey are to be used to estimate optimal prices, care must be taken that the utilities adequately represent customer attitudes toward product prices. There are four issues that require particular attention:

Form of the price utility — The utility for price, $u_i(p)$ in Eq. 1, can in general be either discrete or continuous valued. Continuous valued price utilities are analytically more tractable and allow the analyst to test the appropriateness of different non-linear functional forms. It is even possible to let different respondents within the same study have different functional forms for their price utilities depending on their answers to the tradeoff questions.

Choosing price differentials — Each price tradeoff question requires a difference in the prices used for each side. If this price differential is too small, respondents will always choose the higher cost side thus producing unrealistically low price sensitivities. Similarly, a price differential that is too large will cause respondents to choose the lower cost side. The problem is complicated by the differences among respondents; what is too large for one respondent may be too small for another. If a computer-based interview is used, the computer can alter the price differential based on previous answers by the respondent. For paper/pencil conjoint analysis, pilot test results can be used to adjust price differentials. These alterations of price differentials are much easier to accomplish if the price attribute is continuous-valued.

Adapting the price range to the individual — In some pricing studies individuals will have very different ideas about the range of prices that is appropriate for the product. For example, in choosing a personal computer one person may be thinking about a small computer in the $1000-2000 range while another may want a more powerful unit in the $4000-6000 range. To make the interview credible the respondent must tradeoff prices that are realistic. A computer-based interview can accomplish this easily with a few questions at the beginning of the interview. For paper/pencil studies it is sometimes possible to establish a reference price at the beginning of the interview and then use phrases such as "Reference price + $200" in the tradeoff questions.

Adapting pricing units to the individual — In some situations different respondents may think of price in different units. The most common example of this occurs for expensive items in which some respondents may base their purchase decision on the total price while others may use the equivalent monthly payment. For computer-based interviewing this can be handled easily with a few questions prior to the tradeoffs. It is even possible to include financing attributes such as down payment and loan length, if that is important. For paper/pencil studies this can be handled by asking one or two questions during the recruiting and then sending different questionnaires depending on the responses.

CONCLUSIONS

Conjoint analysis is a data collection and processing technique that can bring some of the simple ideas and techniques of microeconomics to the market planner. Specifically it allows for the estimation of a multi-product demand function in which price is an explicit attribute. When combined with the concepts and tools of modeling and optimization, it can yield new insights into the complex portfolio pricing decisions facing modern corporations.

USING CONJOINT INFORMATION: ORGANIZATIONAL FACTORS

a.k.a

CONFESSIONS OF A MARKET RESEARCH CLIENT: A D&B CONJOINT EXPERIENCE

Sharon W. Salling
The Dun & Bradstreet Corporation

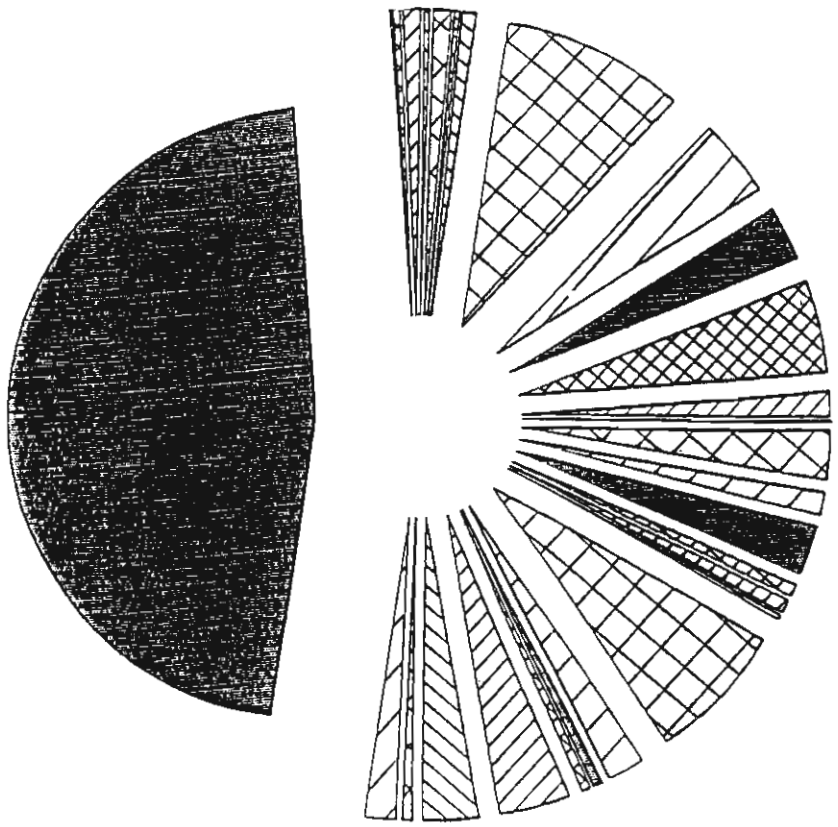
and

John Deegan
Receivable Management Services
a company of The Dun & Bradstreet Corporation

I would like you to imagine that you are the general manager of a company. Your company is in the business of helping its customers manage risk. You are in a business often known as commercial collections. Specifically, you assist your direct customers in reducing the losses that result from their customers incurring overdue payments, partial payments or non-payments.

Now, imagine that the commercial collections marketplace in which your company competes looks like this. It looks like you could use a little help.

AMOUNT TO INDIVIDUAL AGENCIES

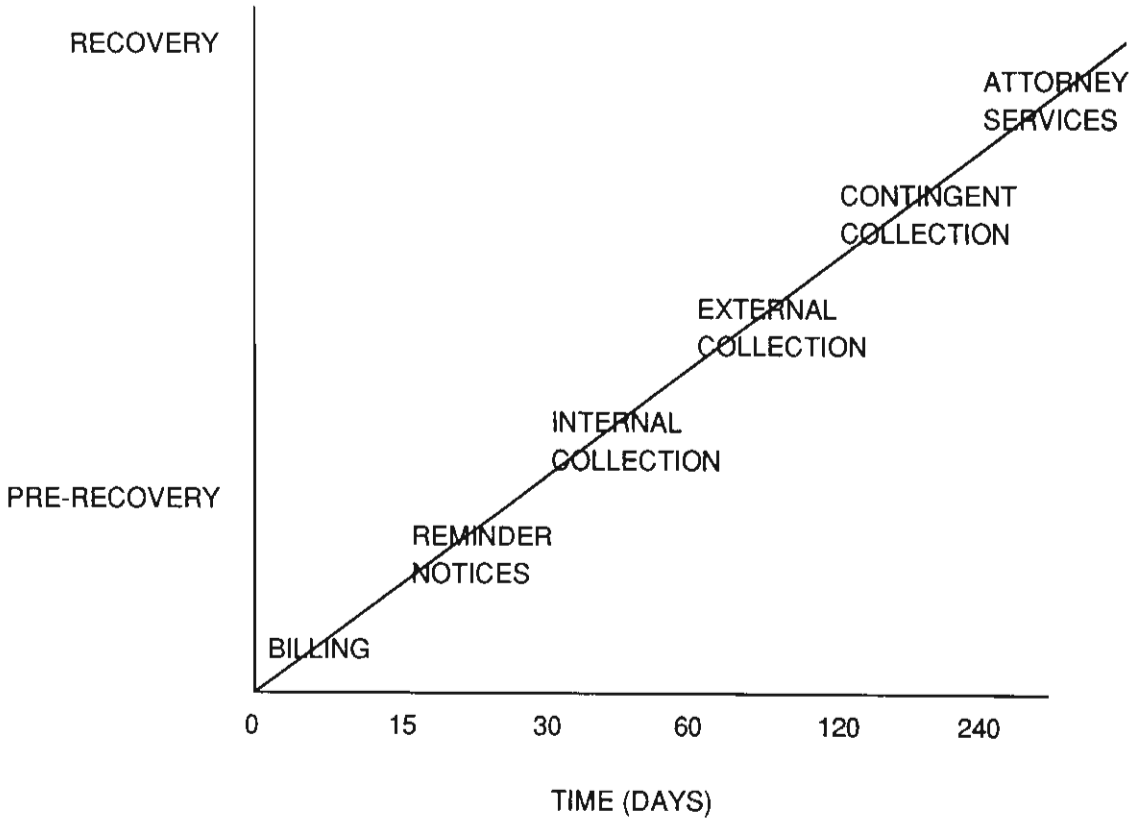


That is the business, and the market situation, that Receivable Management Services faced. We thought we could use a little help. So we turned to our customers, to our sales force and to interactive computer interviewing based on Sawtooth Software, very creatively applied by a company called POPULUS, to provide us with a little help. This assistance arrived in the form of tools and techniques for obtaining and analyzing data that allowed RMS, the Receivable Management division of Dun & Bradstreet, to understand its market in a way it could not have done before.

Before I go further down the path of what we did, where the data for this pie originated and how they are being used, I think a bit of background about Receivable Management Services is in order.

As I mentioned, RMS is in the business of helping its customers manage overdue accounts receivable. This can be viewed as one type of risk management process, illustrated by this process flow diagram.

RECOVERY PROCESS



As you can see, the recovery process can be looked at as having two phases, with the first being referred to as "pre-recovery." This pre-recovery phase typically encompasses about thirty to forty-five days, depending on the dollar amount outstanding, the debtor/creditor relationship and creditor company standards.

During this initial period, the creditor is likely to send out reminder notices and second billings; he will also then begin to "work the paper" through internal efforts, often using a full-time credit and collection staff.

During this process some creditors will turn to an outside agency, such as RMS, to begin a program of recovery attempts — including letters, perhaps telephone calls, and finally attorney referral services — as it appears progressively more likely that payment may not be forthcoming. Some businesses select attorney forwarding quite early in the process, while others do so more or less as a last resort.

However, this deceptively simple diagram and my brief explanation mask both the intensity of the labor required and the human factors involved in collecting overdue accounts. To begin, neither the creditors nor the debtors view involvement in collecting overdue payments as a positive experience.

The debtor may be suffering business reversals, be experiencing cash flow problems, be overextended or unable to meet the payroll. The creditor may also be trying to manage cash flow, to reconcile the books, or may be, in turn, experiencing hard times. Regardless, our direct clients almost always want to retain the debtor as a customer and are eager to work out a reasonable solution. Tact, perseverance, professionalism and accuracy are obviously critical to our success. However, some other factors are not quite as obvious.

RMS had already conducted a comprehensive marketing research project known by those of us who grew up at Procter & Gamble as a "Habits and Practices" Study. This was a conventionally administered piece of telephone research, and it yielded a great deal of descriptive information about the commercial collections marketplace. We learned things like the dollar amounts of typical outstanding accounts, the number of outstanding accounts, the SIC's, how much internal effort is expended by a company to collect, what procedures are followed internally, how many companies use outside collection agencies and a veritable abundance of other usage and habitual information. As with most research, one key outcome was the knowledge that we needed to learn still more, particularly in terms of the "why's" underlying the behaviors we had chronicled.

RMS was able to formulate hypotheses of customer behavior using the data from the habits and practices research. But to truly understand how customers make decisions regarding the selection and retention of an outside agency and to realistically see where RMS stood relative to those characteristics that drive a customer to do business with any agency, RMS required the power offered to us by tradeoff and simulation modeling.

Regardless of the power of this approach, however, some tests had to be passed before the research could begin. Later, we'll give you a glimpse of just how difficult those tests were. They included subjects such as skepticism and the complexity of the marketplace. In addition, we knew the real test would come in the integration of the results into the functional areas of the organization and into the marketing plans for the coming year. We hoped to be able to better determine what our salesforce and our customer service representatives should do to be more in tune with customer needs and to be better able to deal with the highly fragmented nature of the collections marketplace. We had to be sure that D&B could derive full value from the output of this project.

Early in the process we invited the participation of a number of functional areas, particularly sales. As is often the case, the salesforce ranked high on the skeptic charts, but their awareness of the need to have a better understanding of the market and our genuine interest in their participation carried the day. As a result, the study was divided into two parts and the salesforce itself became the respondent base for part one. Equally important, the sales force was able to materially improve the breadth of our customer and prospect sample through its active contribution. Of course, actually sitting at the computer and completing an interview made our sales group most eager to see the project through to completion. Most important, their personal involvement went a long way to ensure that the results would be viewed as credible and would be used.

While the first phase of the research was underway, the customer/prospect phase was being fine-tuned. We mailed nearly 4,000 disks and eagerly awaited that first return. But we really were not worried: we had a great questionnaire. Who could resist answering questions like these?

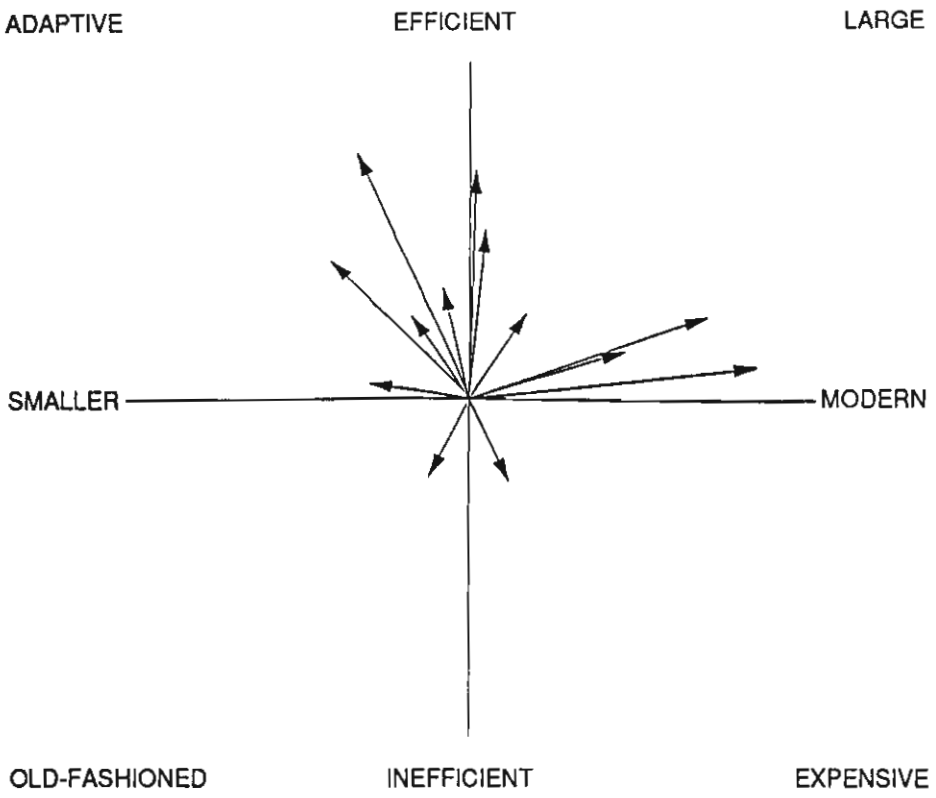
- INTERNAL COLLECTION PRACTICES
- OUTSIDE PARTY USAGE
- OUTSIDE PARTY ATTRIBUTE RATINGS
- DELINQUENT ACCOUNT PROFILE
- RESPONDENT PROFILE
- "CORPOGRAPHICS"

After all, this is what our respondents think about all day, this is what they do, how hard could it be to take a break from the tedium of the day and hit a few keys?

Well, some apparently could resist, and did, but we received excellent returns and were able to begin the really enjoyable work of analyzing the data.

Using a variety of analytic approaches and display techniques, we had some very powerful tools to sort through. We mulled over the considerable amount of information that we had obtained, both from the salesforce and from the marketplace. We used perceptual mapping to look at the key areas distinguishing D&B from its competitors. We also used the simulation model to explore the values and perceptions held by our customers and by our prospects. And of course, not to forget our roots, we poured over the crosstabs for those little clues to the secrets held by those pesky segments and subgroups.

PERCEPTUAL FRAMEWORK



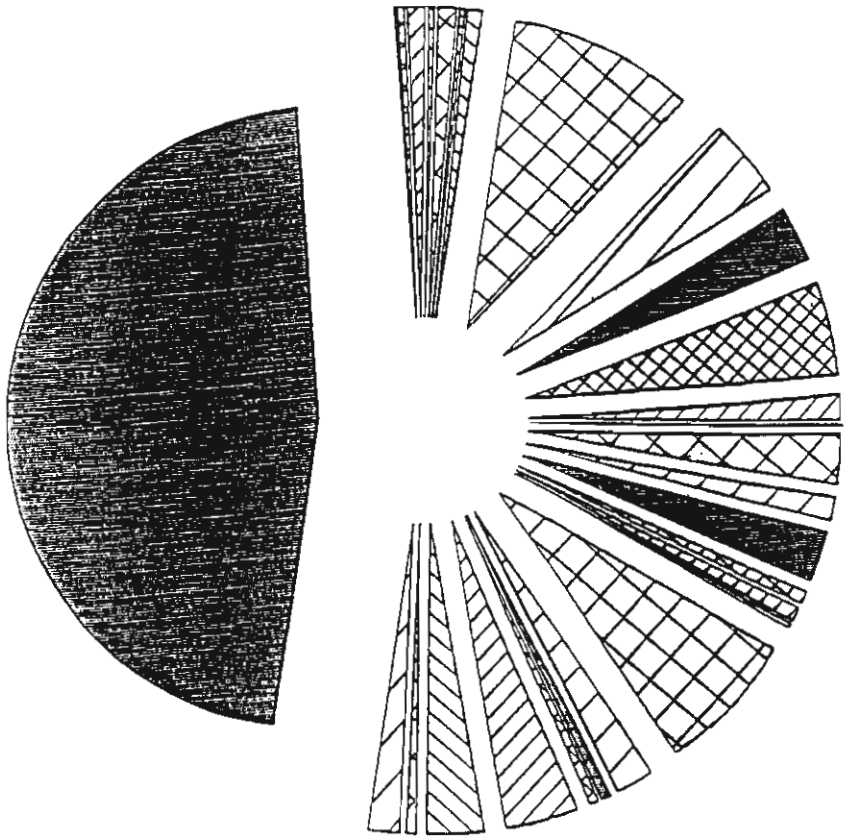
We ran a full range of scenarios, using current situations, likely and not-so-likely alternatives, and were able to develop a set of estimates about changing market conditions, changing customer needs and our own strengths and weaknesses that provided critical input into action plans and strategies. This exercise revealed many of those not-so-obvious factors to us, and really began to explain the "why's" behind the behaviors we had seen in earlier work.

A critical piece of information from this project has focused RMS's efforts on those areas which are clearly capable of delivering the greatest value to customers and which RMS is uniquely able to deliver. In all, this work gave us the help that we needed to better meet customer needs.

But, back to you. Imagine that you are still running a company in the business of helping its customers manage risk. I would suggest that part of the help you need could be found in answers

to questions posed within the framework of a computer interactive conjoint approach. Your market may still look like this:

AMOUNT TO INDIVIDUAL AGENCIES



but armed with information behind the decisions that customers make, you should be on your way to making your slice of the pie a lot bigger.

CONFESSIONS OF A MARKET RESEARCH CLIENT (JOHN DEEGAN)

The use of conjoint analysis in last year's Dun & Bradstreet Receivable Management Services' (D&B RMS) Customer Value Study represented a new and unique approach to market research, both for me personally and professionally, as well as for my unit.

Traditionally, field survey projects had been restricted exclusively to more conventional modes of data collection and processing, namely, telephone contact or mail questionnaires, with responses tabulated and formatted into reports. Our results had been reasonably successful in the past when applying these techniques. Besides, wasn't that what everyone did to secure primary market intelligence and a better understanding of their customers likes and dislikes?

It was with some trepidation, therefore, that we in D&B RMS embarked on this very different venture. As a true pragmatist, however, and knowledgeable of the subtle nuances and complexities of the commercial collection business, I could not help but raise a few doubts as to the efficacy of conjoint in delivering on its promise. Most notably:

- Will the tradeoff portion of our survey, not to mention the computer-assisted interviewing protocol we were electing to implement, intimidate participants into not responding accurately, or even more frightening, into not responding at all?

- How will conjoint possibly help D&B RMS make sense out of a highly competitive and fragmented market map?

- Will the conjoint simulation model be of any true value to developing our strategic direction, or will it create more questions than it actually helps answer?

Assurances to the contrary, when a market research consultant implores you not to worry, experience generally tells me that it's usually your cue to do exactly that.

RESEARCH APPROACH AND CONJOINT METHODOLOGY

Our first concerns centered on the general data gathering approach that we had chosen to employ in our study, namely, the use of computer-assisted interviewing on floppy diskette as a means of recording and collecting the responses of participants.

Securing the sales force perspective on D&B's performance in key agency attributes, then comparing this with the market, was certainly a novel and highly desirable objective. However, the majority of our field people has limited, if any, prior exposure to a personal computer, even though all of our offices are fully equipped with PC technology. For the most part, the availability of PC-trained support staff at each location minimizes their need to utilize this resource.

Similar reservations were expressed about the sophistication of our market sample. More than 60% of our customer base consists of family-owned operations with no more than 10 employees and an average annual sales volume of below $1.0 million. Surely PC availability and competence resided in the credit/collection functions of our larger clients and prime prospects, but what of our typical customer?

The self-administration aspect of the survey via diskette also raised a few doubts among those who were more comfortable with the immediacy of a good old-fashioned telephone interview.

Observation of the pre-test phase of our questionnaire reinforced these fears, as it clearly demonstrated the high level of thought and concentration required of participants to complete our survey. This was particularly true in the tradeoff section, as a total of twenty-nine different attributes were being tested. The jury was out as to whether respondents would find this a stimulating challenge to their professional experience and judgment, or a tedious exercise in frustration that would lead to an early termination of the session.

Likewise, in true client fashion, we at D&B RMS were intent upon leveraging this research opportunity so that as much information as possible could be obtained in support of our Customer Value Study. The end product was a survey which, on average required a minimum of 45 minutes to complete. We found the time element to have been even greater for a few of our pre-test participants.

Much to our delight, the intimidation factor posed by the PC interface, as well as the format and thoroughness of the questionnaire, never materialized, as field reaction to our project was a rousing success. The user-friendliness of the diskette, coupled with the accompanying instructions, helped contribute to a slightly more than 70% response rate among our "PC timid" salespeople. This was higher than that achieved in some of our internal D&B employee surveys in the past, which used a "paper-and-pencil" approach under much more controlled conditions.

Meanwhile, aided by a stimulation telephone call prior to mailing of the diskette, as well as a $25 "thank you" gift, 716 complete and usable questionnaires of the 3,600 mailed were returned within the four-week cut-off date, a 20% response. So much for the notion of minimal PC capability and usage in the average small company office.

MAKING SENSE OF A COMPLEX MARKET MAP

The extremely competitive and highly fragmented nature of the commercial collections industry, of which D&B is a part, posed yet another challenge to the conjoint method of analysis, particularly with respect to the rating of D&B versus other agencies in key attributes. Best available estimates indicate that more than 6,000 companies, ranging from large nationwide organizations to small local enterprises, are engaged in some aspect of debt recovery.

The collections market is further fractionated as a consequence of attorney firm involvement in the business. Some specialize in collections, while others may handle delinquent debtors alongside other legal matters simply as a matter of course. However passive or active this form of competition, attorneys collectively represent a formidable force in the industry.

Meanwhile, a number of companies as a matter of policy do not engage collection agencies, electing instead to devote extensive in-house resources to the recovery of delinquent debt prior to write-off. Our research found that this segment may represent as much as 25% of the total market, as defined as the aggregate volume of past due receivable dollars available for collection.

The study was further complicated by our desire to explore attribute ratings and perceptions of D&B among several groupings of sample, namely, current D&B clients, prime prospects in the commercial marketplace, and former users of our products and services.

An additional goal — to secure similar feedback on a subsidiary operation of our unit, Bruhnke & Silver, a provider of specialized collection support to customers — offered yet another wrinkle to the project. Likewise was true of a secondary objective, which involved the scoping of the size and characteristics of the "overlap" business, defined as collections for companies that maintain both consumer and commercial receivables in their A/R portfolio. A truer test of conjoint neither I nor our friends at POPULUS could have possibly dreamt up in our worst nightmares.

Here again, however, the research design and methodology did not disappoint. Due to the flexibility of the conjoint logic, respondents were able to select their current collection vendors from a long roster of agencies, including Bruhnke & Silver, and even "write-in" names of companies that were not listed. The same was true for the identification of agencies that participants had used at some point in the past five years, but did not presently employ. This allowed us to identify former D&B customers in our sample base.

Likewise, pre-qualifiers in the questionnaire gave us the chance to determine the extent of attorney firm utilization, as well as the incidence of companies that did not outsource their collection activity as a norm.

Meanwhile, in the conjoint portion of the survey, the software approach to the selection of companies for attribute evaluation purposes was constructed so that it gave the appearance of randomness to the survey participant.

Agency names would flash intermittently on the screen, stopping at the name of the respondent's vendor identified earlier in the questionnaire, as well as, very conveniently, that of Dun & Bradstreet.

This clever ability to call up and route earlier responses in the survey for use in other parts of the questionnaire, coupled with the sheer capacity of the diskette as a data gathering tool, were instrumental to our capturing volumes of information in support of each major topic of our study. For example, in addition to the conjoint tradeoff, we were able to investigate the issue of consumer and commercial receivables by profiling respondents' A/R portfolios, including the average age and size of accounts placed with agencies. Also examined were factors driving the decision to place accounts with a third party; the frequency of placement; type of agency services used; and the perceived cost effectiveness of these services and those of attorneys.

Exhaustive in its outlook and approach, the D&B conjoint study yielded a virtual treasure trove of information that remains as our primary market reference. In hindsight, I can think of no other methodology that could possibly have accommodated what we set out to accomplish, short of extensive face-to-face interviews with our customers.

CONJOINT SIMULATOR: PANACEA OR PANDORA'S BOX?

The installation of the conjoint simulation model on my office PC was met with much excitement among executives of D&B RMS. Despite warnings to the contrary, however, too many members of our management team viewed this model as a convenient and reliable means of forecasting the financial performance of our unit, especially the sales/revenue impact of specific investment initiatives.

As a result, there was some disillusionment when the assumptions of perfect information were explained to my colleagues, most notably, the conditions of full communication of changes to the market, as well as competitive reaction as a known and controllable variable. With the cost side of the equation also excluded from the mechanics of the simulator, doubts were raised as to the true value of the model as a facilitator of our internal management efforts.

Having said all that, one might conclude that we chose to dismiss entirely this novel approach to market analysis. To the contrary, once the full scope of the model's capabilities and caveats were known and understood, the true payoff of conjoint came to fruition.

Today, we view the simulator as an important "reality check" for many of our operating decisions, as it provides an instant and measurable reading of what the market reaction might be to certain initiatives. In addition, since the model is based on primary information from a representative sample of current and potential users of our services, we no longer formulate strategies in a vacuum. Rather, we are able to rely on direct input from the very group that our plans are specifically designed to address, namely, our customer base.

When regarded in this context, the value of the conjoint simulator as a point of reference or "barometer" of market reaction cannot be debated. A word of warning, however, to those that extol its virtues: be aware that faulty assumptions at the outset of what the model can and cannot do may lead to unrealistic expectations on behalf of the user and, ultimately, disappointment over the results. This is particularly true in today's business environment, where a premium is placed on efficient technical solutions to many issues, including those involving attitudinal and perceptual factors, as a means of minimizing risk.

THE CONJOINT PAYOFF

Having expressed my concerns over the conjoint approach to business research, I would be less than honest if I did not confess that, in hindsight, my fears were largely unfounded. The proof ultimately was in the payoff for our unit. As mentioned earlier, the study produced a library of market information on a number of diverse topics, ranging from the professional profile of respondents and their departments, in-house receivable management practices, and incidence of third party usage for collections; to the identification of key market drivers and D&B's standing in terms of delivery on those attributes.

The ability to translate practical market intelligence into pro-active strategies and programs is an additional strength realized through this approach. Too many research projects in the past have resulted in findings that were "nice-to-know," but contributed little to the formulation of solutions.

In our case, conjoint helped us to identify what we call the problems of "bigness" associated with our standing as a market leader, as well as some widely held misconceptions regarding our collection performance and product capability. Investments in workstation technology, training and communications programs are already underway that will help us counter these obstacles to our continued leadership and growth.

In this regard, conjoint has truly opened the door for D&B to "do what we couldn't do before" in terms of capturing and replicating through research the dynamics of our business.

Likewise, concerns to the contrary, it also helped expand the horizons of those of us whose exposure to research techniques was limited primarily to the "tried and true" methods of the past.

Comment on Salling and Deegan

Steven M. Struhl
Sophisticated Data Research

The Salling and Deegan paper presents one excellent strategy for use in complex research studies, opens several questions, and raises one serious concern.

The strategy is so strong that, even as a reminder, it makes reading the entire paper worthwhile. Simply, the authors asked the parties most likely to object to their survey (the salesforce) to become involved by answering the questions themselves. They did this with an understanding that responses from the sales force would be compared with responses from customers and prospects. Few things can make a survey come more fully to life, for those not involved in constructing a survey, than this type of exercise. It helps focus the organization's attention on the meaning of the questions, and can make many within the organization eager to see the findings.

In this study, the authors restricted themselves to having the sales force answer questions about perceptions of the corporation. They judged the market to be too fragmented and complex for comparing corporate and marketplace responses to the conjoint potion of the study. Yet, even a highly complex market does not preclude getting people in the organization to answer all the questions. Since conjoint (as a reminder) generates person-by-person utility data, you can have sales people **guess** what their customers and prospects would say. You could then separate these respondents out as a group, and show how their answers compared to the sales peoples' perceptions. This type of approach can strongly boost organizational involvement in a procedure such as conjoint that can seem like a "black box."

I also appreciated the great enthusiasm the authors showed for the conjoint procedure. We observe this in nearly all cases where a company finally decides to go beyond simple crosstabulations. Usually, there comes an initial period characterized by strong anxiety, and many confused-sounding questions. Then, once those involved begin to recognize how much the technique has done, they become fervent about it. We have seen this with all advanced analytical techniques, including conjoint.

This last point brings us to the questions elicited by this paper. The authors understandably feel excited by their new discovery, but they say little in the paper that makes it clear why conjoint was the technique of choice for the problem they faced. Other methods have been used with equal success to understand customers' and prospects' needs and wants. Throughout this paper, I was hoping to see more about what they "traded off," how they traded, and their reasons for these tradeoffs. In particular, I would have welcomed more information concerning the ways in which the authors traded costs for other attributes. In their talk, they mentioned using "relative prices," but they did not give detail on the "how and why" of doing this.

The salient concern this paper raised concerns the response rate to the survey. This was only 20%. In spite of what the authors said, response to a mail survey at this level cannot be considered excellent. I cannot avoid serious concerns about whether their sample was, in fact, representative. Concerning this point, I have occasionally seen similarly low response rates to surveys described as good (or better). I believe some confusion exists between responses to direct mail and responses to mail questionnaires. I am not aware of any published studies showing that response rates below 33% produce representative samples. During the discussion, Steven Cooley of Blue Cross/Blue Shield mentioned that he once did a split-sample study in which one portion of the sample was pre-recruited and the other half not. Response rates were 70% and 20% respectively. Most importantly, he found no significant differences in the characteristics of the two sub-samples. Still more evidence of this type will be needed before we can judge a 20% response rate as adequate.

In summary, this paper's truly strong points lay in its descriptions of how corporate involvement in a survey helped boost acceptance of the results. For this alone, this paper is worth reading.

VALIDATION OF ADAPTIVE CONJOINT ANALYSIS (ACA) VERSUS STANDARD CONCEPT TESTING

James J. Tumbusch
MarketVision Research, Inc.

Published evidence of the validity of conjoint analysis techniques has been sparse, rarely based on data collected for any real marketing purpose, and consisting mainly of unrealistically small simulation experiments (three to six attributes). This paper describes a successful validation of Sawtooth Software's Adaptive Conjoint Analysis (ACA) via independent concept testing. Across four different frequently purchased product categories, strong correlations were obtained between the consumer appeal derived from conjoint studies using ACA and the corresponding appeal obtained from standard concept tests known to be predictive of product success/failure in the real marketplace.

INTRODUCTION

Conjoint analysis is based on a model of consumer choice behavior which underlies nearly all of our more conventional market research techniques (for example, blind and identified product tests, attitude studies, concept testing). The model states that a brand, product, or service can be considered as a "bundle" of attributes or attribute levels, each of which makes some contribution to overall customer/consumer acceptability. Individual customers/consumers may want different mixes of these attribute levels. Each customer/consumer will be more likely to buy an existing product/ service, or try a new product/service, if it comes closer to his/her desired "mix" than other available alternatives.

With more conventional product or concept testing we usually vary one or two elements of the attribute mix at a time to determine the effect of the change(s) on customer/consumer acceptance (or preference). With conjoint analysis, we can estimate the contributions to overall choice behavior of many different attribute-level changes, all in the same study.

With Sawtooth Software's Adaptive Conjoint Analysis (ACA) approach, data are collected via a microcomputer interactive questioning technique, with questions based on the individual respondent's earlier answers and his/her value system relative to the product/service category being investigated. By asking respondent preference between a relatively small number of pairs of hypothetical products (attribute-level bundles), we can estimate the respondent's individual likelihood of choosing <u>any</u> hypothetical product/service over any other such product/service, or group of products/services (encompassed by the conjoint study).

Conjoint analysis can be employed from a point early in the product development process (early concept stage), to just before test marketing. It can be used to: 1) allocate product development resources toward attributes where changes would substantially increase customer/consumer

acceptability, 2) precede conventional concept testing where it is appropriate/necessary to evaluate a large number of attribute-level combinations, and/or 3) determine an optimal positioning for a product/service relative to competition.

VALIDATION OF CONJOINT ANALYSIS RESULTS

Prior to the comparatively recent availability of practical methods for conducting conjoint studies, market researchers depended almost entirely upon standard concept testing to finalize concept design. Beginning in 1985, with the successful implementation of computer-interactive interviewing using ACA, large scale conjoint studies became feasible and economical. But, to date, published information on the validation and/or calibration of conjoint results has been sparse.

Conjoint analysis is, in our judgment, most effectively positioned as a replacement for conventional concept testing, especially when it is desirable to study consumer reactions to many different variations in product attributes all in the same study. It seems entirely reasonable to expect conjoint results to be somewhat predictive of concept test results.

METHODOLOGY

During 1985, conjoint studies were conducted for each of four different frequently purchased product categories. For each study, respondents were screened and recruited by phone; over 300 ACA/Ci2 (Ci2 System for Computer Interviewing) interviews were conducted at central location facilities in several different cities:

Conjoint Study	Number of Attributes	Base Size	Number of Cities
Product Category A	7	706	3
Product Category B	11	342	3
Product Category C	11	319	3
Product Category D	6	365	4

From the results of each of the four conjoint studies, a broad range of concepts (from "most appealing" to "least appealing") were chosen to be evaluated via concept testing. From the specific set of attribute-levels that made up each concept, concept statements were written. For each concept chosen from each of the conjoint studies, questionnaires were mailed to an independent

sample of households large enough to ensure 300 returns for each concept. Each respondent was asked to read the appropriate concept statement and indicate his/her propensity to purchase the product described:

- ☐ Definitely will buy
- ☐ Probably will buy
- ☐ May or may not buy
- ☐ Probably will not buy
- ☐ Definitely will not buy

Data analysis consisted of correlating the total utility across respondents (indexed to 100) derived from the conjoint study versus percent of respondents who answered "Definitely will buy" derived from the concept test.

RESULTS

The table on the next page shows conjoint study appeal (utility index) versus concept test appeal (% Definitely will buy).

Concept	Conjoint Study Utility Index*	Concept Test % Definitely Will Buy	Correlation
Product Category A:			
1	100	51	
2	72	47	
3	68	45	+.72
4	48	54	
5	10	28	
60	39		
Product Category B:			
1	100	19	
2	99	18	
3	98	28	
4	83	18	
5	82	17	+.77
6	82	21	
7	74	18	
8	35	16	
90	9		
Product Category C:			
1	100	21	
2	88	19	
3	77	21	
4	76	25	
5	65	22	+.81
6	60	23	
7	51	10	
8	10	12	
9	0	3	
Product Category D:			
1	100	57	
2	89	39	
3	80	36	
4	76	40	
5	75	42	
6	73	35	+.75
7	58	35	
8	42	40	
9	39	36	
10	34	24	
11	0	26	

* Utility Index - Relative appeal of one concept versus the others, typically 100 for the most appealing concept, 0 for the least appealing concept, encompassed by the conjoint study.

The relationship between concept test appeal and conjoint study appeal is also shown graphically in Figures 1 through 4.

Figure 1. Product Category A

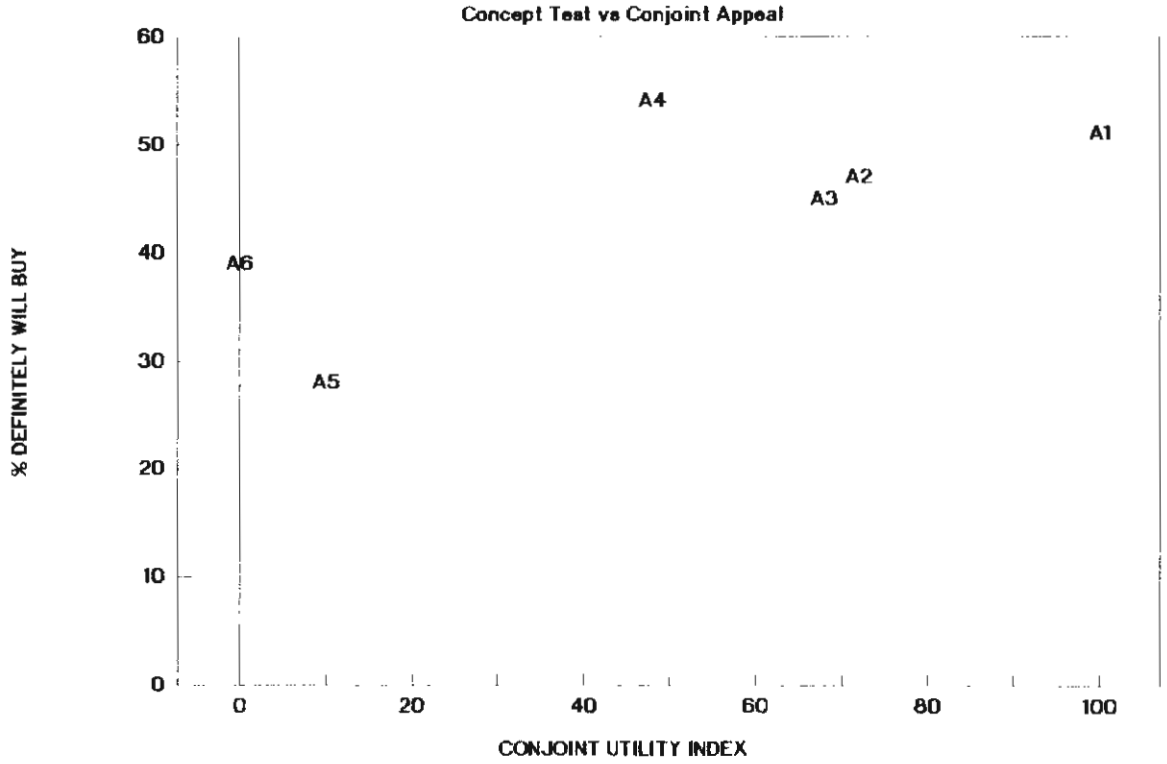


Figure 2. Product Category B

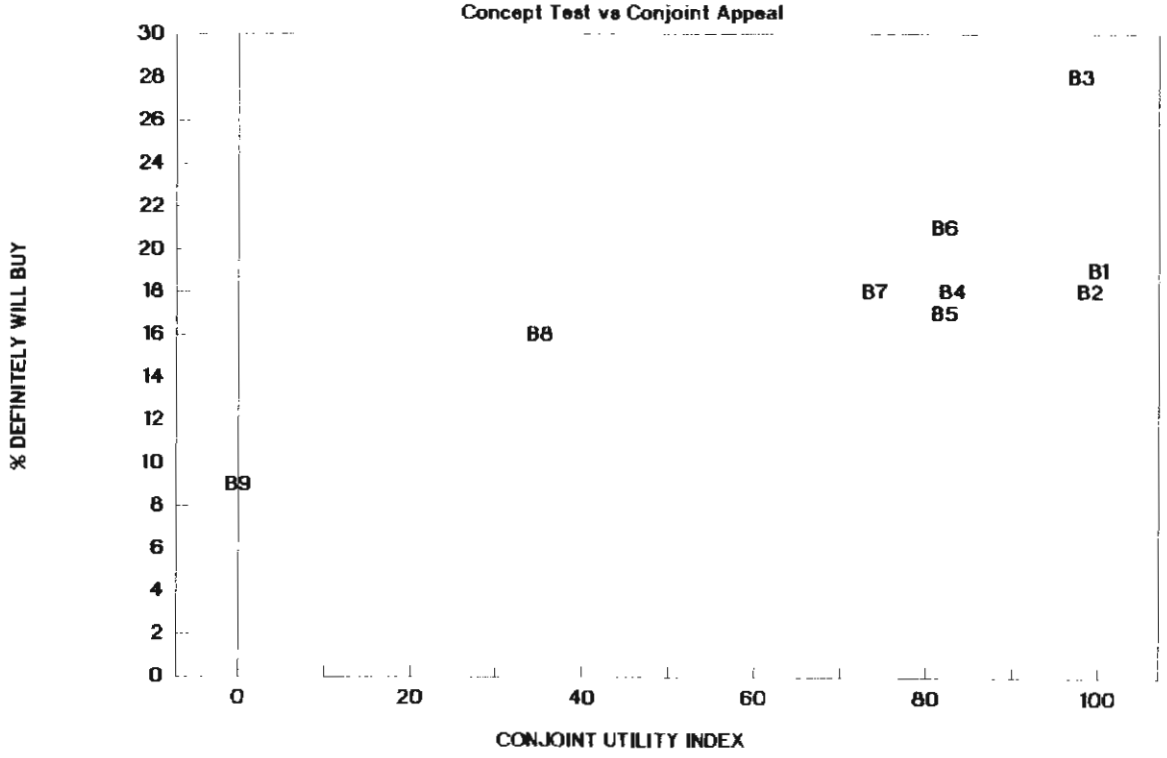


Figure 3. Product Category C

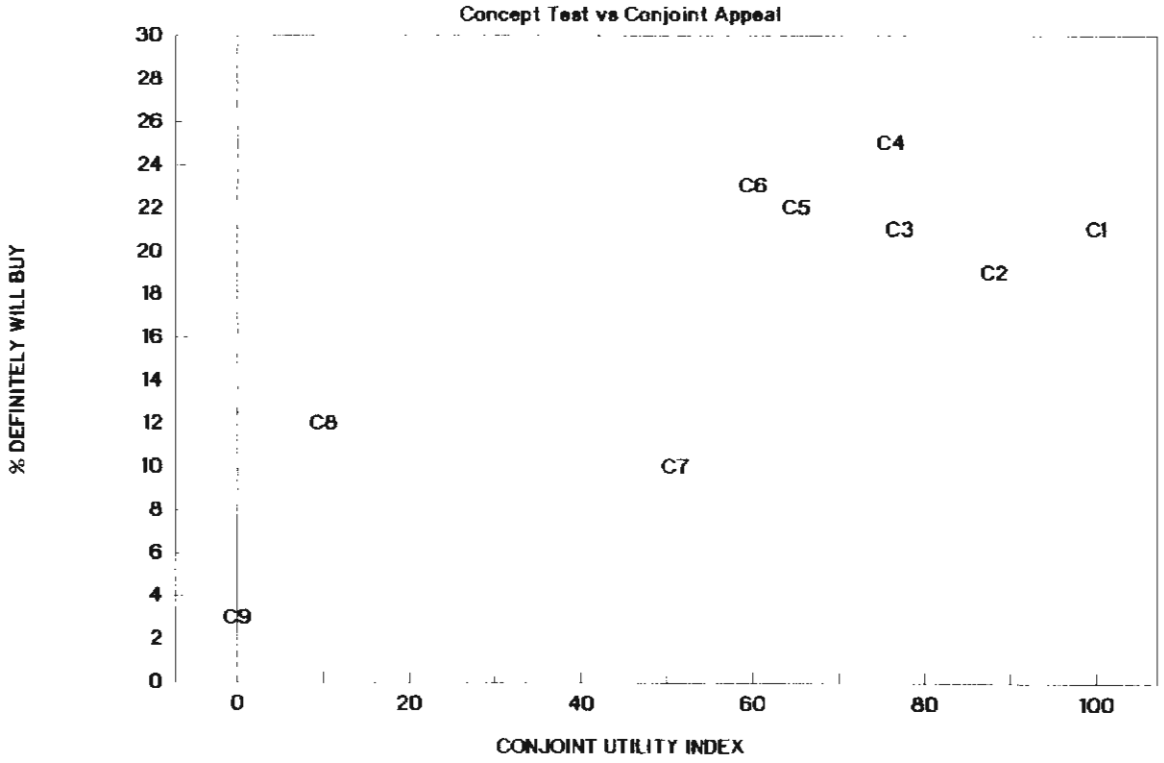
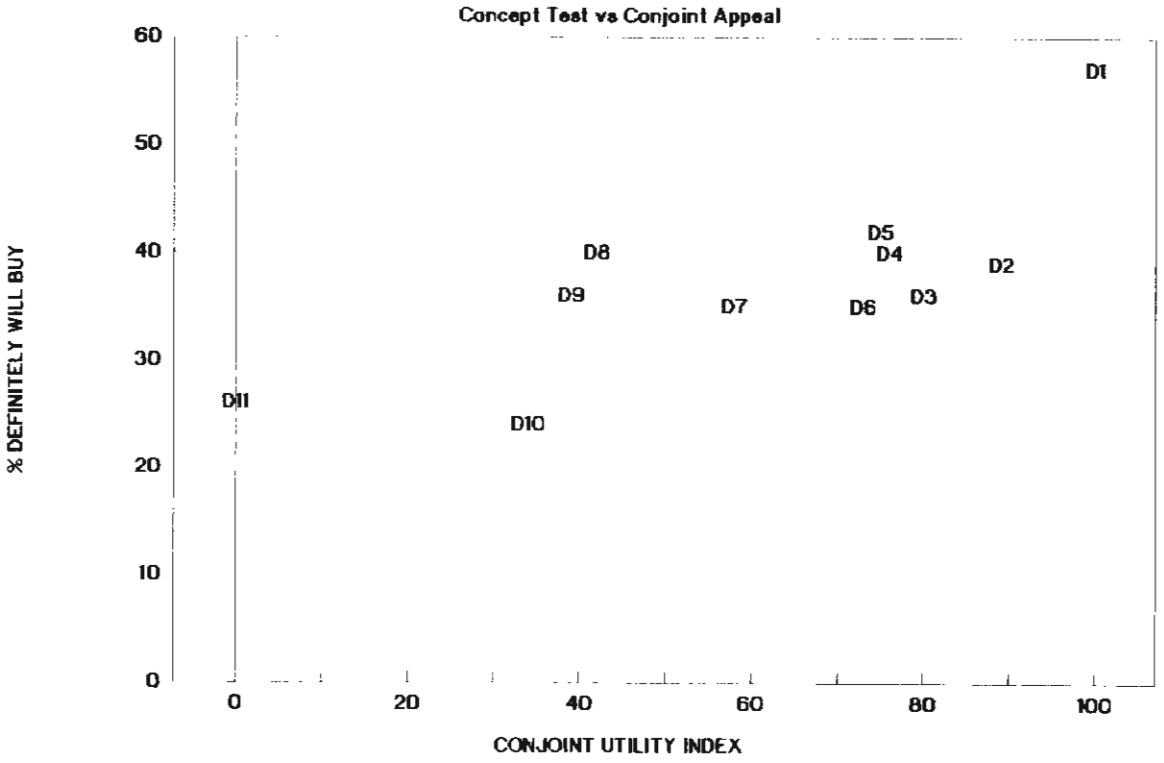


Figure 4. Product Category D



DISCUSSION

Since the concept testing was done from six to twelve months after the completion of the conjoint studies, with completely independent samples of respondents, this attempted validation constitutes the most severe test of ACA results of which we are aware. Sampling alone can account for a variation of plus/minus 5 percent in "% Definitely will buy." In addition, consumer value systems relative to the individual product categories involved could have undergone certain changes over the elapsed time between the conjoint studies and concept testing.

These results represent strong evidence across four different product categories that: 1) the rank order of appeal of concepts derived from conjoint analysis corresponds reasonably well to the rank order of appeal obtained from conventional concept testing, and 2) there is a strong correlation between concept appeal derived from conjoint analysis and appeal resulting from concept testing, which itself has been shown to be predictive of product success/failure in the marketplace.

There is also evidence that conjoint analysis may be more sensitive than conventional concept testing. Conjoint analysis forces the respondent to make choices between pairs of attribute-level bundles whereas we have seen that if we expose consumers (via concept test) to a variety of concepts encompassed by the conjoint study, there may be no appreciable variation in appeal. For example, for the Product Category C study, the conjoint study utility index for the top six concepts ranges from 60 to 100, while concept appeal only varies from 19% to 25% "Definitely will buy."

Comment on Tumbusch

Carl T. Finkbeiner
National Analysts, Division of
Booz•Allen & Hamilton, Inc.

I congratulate Jim on two counts: first, because his paper is a neat and tidy contribution to these proceedings and, second, because I know the extraordinary effort required for him to get permission to present these results. In fact, getting permission may be even more of an accomplishment than the work itself.

When I read Jim's paper, the initial concern I had was over the statistical significance of the four correlations he reports. After all, the number of observations (the concepts) which go into each correlation is small: 6, 9 or 11, depending on the category. I realized that my first impulse — a significance test of the hypothesis of zero correlation in the population using the standard test of a Pearson correlation — was not appropriate for three reasons.

- The standard test requires that the observations be independent. In Jim's data, the observations are dependent because of the fact that the conjoint utility indices are actually repeated-measures on each individual respondent.

- The standard test requires that the two variables be observed jointly on the same observation. However, the two variables in Jim's data are observed in two independent samples for any given concept.

- The standard test assumes normality of the variables. However, percents are known to be generally non-normal in distribution, and one of the two variables here — for the concept test ratings — is percents.

Therefore, the standard test on a Pearson correlation is not appropriate. However, Kendall's tau-coefficient can be used here since its significance test does not require the preceding assumptions. Kendall's tau is an alternative way of measuring the strength of association between two variables.

Applying Kendall's tau to the four product categories yields the following results:

Category	tau
A	.47
B	.59
C	.31
D	.52

Like the Pearson correlation, Kendall's tau varies between - 1 (perfect negative association) and +1 (perfect positive association), with zero indicating no association. The above results, then, indicate a moderate positive association (.3 - .6) between the ACA predictions and the concept tests in all four categories.

Applying a standard significance test to Kendall's tau yields the probabilities below:

Category	p
A	.19
B	.03
C	.25
D	.03

Each of these is the probability of obtaining the observed tau from a population in which the population tau is zero. Thus, two of the four taus are not statistically significant.

However, this result does not actually contradict Jim's conclusion. As a practicing market researcher, I am used to examining data for patterns, even when the sample size is not large enough to allow for statistical significance. Just because a single significance test is not sensitive enough to detect an effect does not mean the effect doesn't exist. A pattern, repeated across the data, becomes compelling even in the absence of statistical significance.

In the present case, we can translate "pattern examination" into formal statistical terms. There are at least the following three ways to do this:

- Because the results for the four categories are from different samples, the four Kendall's tau tests are independent. Consequently, the joint probability of all four results occurring (under the null hypothesis of no association between ACA predictions and concept test percents) is just the product of the above four probabilities: p = .00004 (= .19 X .03 X .25 X .03). This joint probability should be judged against the joint probability of four independent tests, all of which had been accepted, say, by p = .5: in this case, p = .0625. In any event, the joint probability of Jim's data is extremely small.

- If we use a .05 cutoff for assessing the significance of the four Kandall's tau tests, then the binomial probability of obtaining two significant results out of four tests is .01, again a small probability.

- If we had used a looser cutoff of .25 so that all four tests were significant, then the binomial probability is .02, still quite small.

All three of the above procedures indicate a very low probability of obtaining Jim's results if there was in fact no relationship between ACA predictions and concept test results. Consequently, we are justified in concluding that there is a positive association.

I would not be doing my job as a reviewer if I didn't quibble with something in Jim's paper. I would like to reiterate a point Jim made in his presentation about a statement in the written paper: ". . . standard concept tests (are) known to be predictive of product success/failure in the real marketplace." Jim has subsequently acknowledged that this statement should be qualified. It is rarely true, even at Procter & Gamble, that concept tests predict actual market behavior unless the test results are adjusted for appropriate marketing influences (such as advertising or distribution channel effects). Furthermore, concept tests, even when appropriately adjusted, are best regarded as a measure of initial trial and should be further adjusted for the rate of repeat purchase.

A second quibble relates to Jim's statement that ". . . conjoint analysis may be more sensitive than conventional concept testing." Presumably, this is based on the observation that the mean conjoint utilities have a greater range of values than do the concept test percents. But remember that the conjoint utilities were arbitrarily scaled to range from 0 to 100. They could as justifiably have been scaled to range from 0 to 1 and then they would have looked less sensitive. My point is that a sensitivity comparison requires that within-concept variances must be taken into account via some variation of ANOVA or MANOVA. Until that is done, we can't really say anything about the relative sensitivity of ACA versus concept tests.

Neither of the above quibbles affects Jim's overall conclusions. His results do provide evidence that ACA results are at least moderately associated with concept test results. This is valuable evidence for those of us who find concept tests useful in market research.

AN EMPIRICAL COMPARISON OF ACA AND FULL PROFILE JUDGMENTS

Joel C. Huber
Duke University
Dick R. Wittink
Cornell University
John A. Fiedler
POPULUS, Inc.
Richard L. Miller
Consumer Pulse, Inc.

INTRODUCTION

Only a few years ago, Johnson (1987) introduced Adaptive Conjoint Analysis (ACA) to the market research community. This innovative approach to the quantification of consumers' preference structures has become very popular (Green *et al.* 1991). One advantage of ACA, relative to other approaches, is that it combines the design of the conjoint tasks, data collection, data analysis and market simulation in one piece of software. Perhaps the most interesting aspect of ACA is that the use of the personal computer allows ACA to adapt the conjoint task to the individual respondent. The procedure can be considered to consist of two stages. In the first stage, the respondent indicates, among other things, the (relative) importance of each of the attributes defined for the study. In the second stage, answers to conjoint questions are elicited. These questions are defined based on information gathered in the first stage (and based on answers to previous conjoint questions). In this manner, the conjoint tasks can elicit responses to questions designed to reduce the uncertainty in the respondent's estimated parameters. Although ACA can be used for any problem that requires a quantification of preference structures, it should become increasingly attractive (relative to alternative methods) as 1) the number of attributes to be included in the study increases, and 2) respondent heterogeneity in the attributes' relevance increases.

Given the availability of several alternative procedures for preference (and choice) measurement, it is of interest to evaluate and compare their performances. Green and Srinivasan (1990) recommend full profile conjoint analysis if no more than approximately six attributes are involved in a study. For a somewhat larger number of attributes they suggest that tradeoff matrices could be used (although they also favor bridging designs with full profiles). Only when the number of attributes reaches ten or more do they recommend the use of self-explicated data and methods involving a combination of self-explicated and conjoint data (such as ACA).

Currently there is very little information available on the (empirical) performance of alternative methods. MacBride and Johnson (1979) obtained higher first choice predictive validity for a forerunner of ACA relative to the tradeoff matrix approach, although the difference was not statistically significant. Finkbeiner and Platz (1986) compared ACA with the full profile method in a study involving six monotone attributes. They obtained roughly comparable predictive validities.

Agarwal and Green (1989) had respondents go through ACA, a separate self-explicated task and two full profile rating tasks. They also used six monotone attributes. Although Agarwal and Green obtained better predictions from the self-explicated data (than ACA) on the full profile holdout data, Johnson (1991) expresses substantial concerns about confounds.

In this paper, we focus on ACA and full profile as two alternative methods for quantifying preference structures. We mention unique characteristics of each method which can be used to suggest application contexts for which a given method is especially suitable. This is followed by a description of our study to obtain empirical results on the relative performance of the two methods. We show predictive validity results and end with conclusions and suggestions for further research.

ACA Versus Full Profile

Given that ACA and full profile appear to be the most popular approaches for quantifying consumers' preference structures, it is of interest to ask more generally how these two methods compare. One may argue that the two approaches are so different that each has its own application areas. On the other hand, there are many cases where either approach can be used. It is important then to provide both conceptual (theoretical) and empirical comparisons. For a detailed description of ACA see Johnson (1987).

We provide in Table 1 a comparison of some distinct and relevant characteristics of the two approaches. The descriptions in this table are a convenient summary of the essential differences. We believe it is useful to mention that there is potential within ACA to impose constraints on the estimated partworths. For full profiles, Srinivasan *et al.* (1983) show that the predictive validity can be improved if homogeneous *a priori* constraints on the order of individual attributes' partworths are imposed for all respondents. In ACA such constraints can easily be introduced for attributes which are assumed to have a known (typically monotone) preference order for the levels. In addition, constraints can be imposed for attributes for which respondent-specific preference orders for the levels are obtained in the self-explicated data. If these respondent-supplied orders can be assumed to be valid, such constraints give ACA an (additional) edge over full profiles that can only be matched if the full profile method is expanded to include (some) self-explicated data.

Table 1

Characteristics of ACA and Full Profiles

ACA	Full Profile
- computer-interactive	- any desired form of data collection
- combines design, data collection, analysis, and market simulation	- design, collection, analysis, and market simulation are often separated
- can accommodate a large number of attributes	- restricted to about six attributes, unless bridging designs are used
- objects never fully specified (two to five attributes)	- objects specified on all attributes
- combines self-explicated data with paired comparison intensity ratings	- fully decompositional approach
- stimulus design not orthogonal (statistically inefficient)	- stimulus design typically orthogonal
- paired comparisons are "close" in overall utility	- orthogonal arrays produce profiles that may differ greatly from each other in overall utility
- paired comparisons adapted to respondent-specific prior evaluations	- usually same set of full profiles for all respondents
- could impose *a priori* constraints for all respondents, and/or respondent-specific constraints using self-explicated data, on the estimated partworths	- *a priori* constraints could be imposed on the estimated partworths
- expected to be subject to attribute level effects (but to a smaller degree than full profile)	- subject to systematic effects, e.g., results depend on the order of attributes, and the number of attribute levels

It is also of interest to point out that the full profile method is subject to some systematic design effects. For example, Johnson (1989) indicates that the relative importance of attributes (derived from the partworths) is influenced by the order in which the attributes appear in the full profiles. That is, if the same attribute appears first for every respondent, it tends to receive more attention than if it appears later. Such a systematic effect should not exist in ACA, because the paired comparison profiles are not presented based on a constant or systematic order of the attributes. Also, Wittink *et al.* (1989) mention that full profiles (as well as tradeoff matrices) are subject to systematic effects due to the number of levels on which attributes are defined. For ACA attribute level effects should be reduced, because of the strong influence of the self-explicated data on the final partworths (Green *et al.* 1991) and the use of preference intensity judgments.

In ACA the self-explicated data (preference orders for attribute levels and importance ratings for the differences between best and worst levels of attributes) are used to obtain an initial set of partworths. These initial partworths are defined such that the difference between the values for the best and worst levels of an attribute corresponds to the stated importance of the attribute. Intermediate attribute levels obtain values that correspond to the preference order for the levels. For example, if the attribute warranty has three levels and its importance equals 2 on a 4-point scale, the initial partworths are +1, 0 and -1, respectively for the best, intermediate and worst levels.

One may argue that the paired comparison intensity judgments are or should be used 1) to differentially stretch the attributes' importances, and 2) to modify the assumption of equal utility distances between successive attribute levels. However, there is nothing in the estimation procedure that prevents the updated partworths from altering the assumed or specified preference order for the levels. Especially for *a priori* known (e.g., monotone) preference orders for the levels of an attribute, it is very likely that imposing order constraints on the final partworths will improve the results. Green *et al.* (1991) in their discussion of ACA focus primarily on the extent to which the paired comparison intensity values are commensurate with the initial (or previously updated) partworths. They suggest that ACA results can be improved by differentially weighting the self-explicated data and the paired comparison preferences. Johnson (1991) notes that the nature of differential weights for optimal improvement in predictive validity varies between applications. Thus, improvements in ACA performance are possible if the differential weights can be found (optimized) for each application separately.

Experimental Design

The primary objective of this study is to obtain an empirical comparison of the ACA and full profile conjoint methods. In particular, we are interested in determining the predictive validity of the two methods as determined for four choice (validation) tasks.

We chose refrigerators as the product category. (We could imagine that the absolute and relative predictive validities of the two conjoint methods depend on the product category. However, it is not at all clear in what manner or for what reasons the possible superiority of one method over another depends on the product category.) To enhance the generalizability of the results we varied the number of attributes. (As the number of attributes increases the full profile method becomes more difficult to apply.) Specifically, half the respondents saw five, while the other half saw nine. All nine attributes are described in Table 2.

Table 2

Refrigerator Attributes

A. Brand Name - General Electric, Sears/Kenmore, Whirlpool

B. Capacity - cubic feet: 19, 20*, 21*, 22

C. Energy Cost - annual: $70, $80*, $90*, $100

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 two for B and E. In ACA, attributes D and E were each given four levels for half the respondents, and attributes B and C had four levels for the other half.

Each respondent provided both full profile ratings and ACA judgments. However, because the predictive validity of the (holdout) choices for each method may depend on the order in which the methods were applied, we used both orders. Thus, half the respondents completed the full profile task first, while the other half did the ACA task first. Both tasks were administered by computer. This manipulation allows us to determine the extent to which task order effects exist and to adjust the results for their presence.

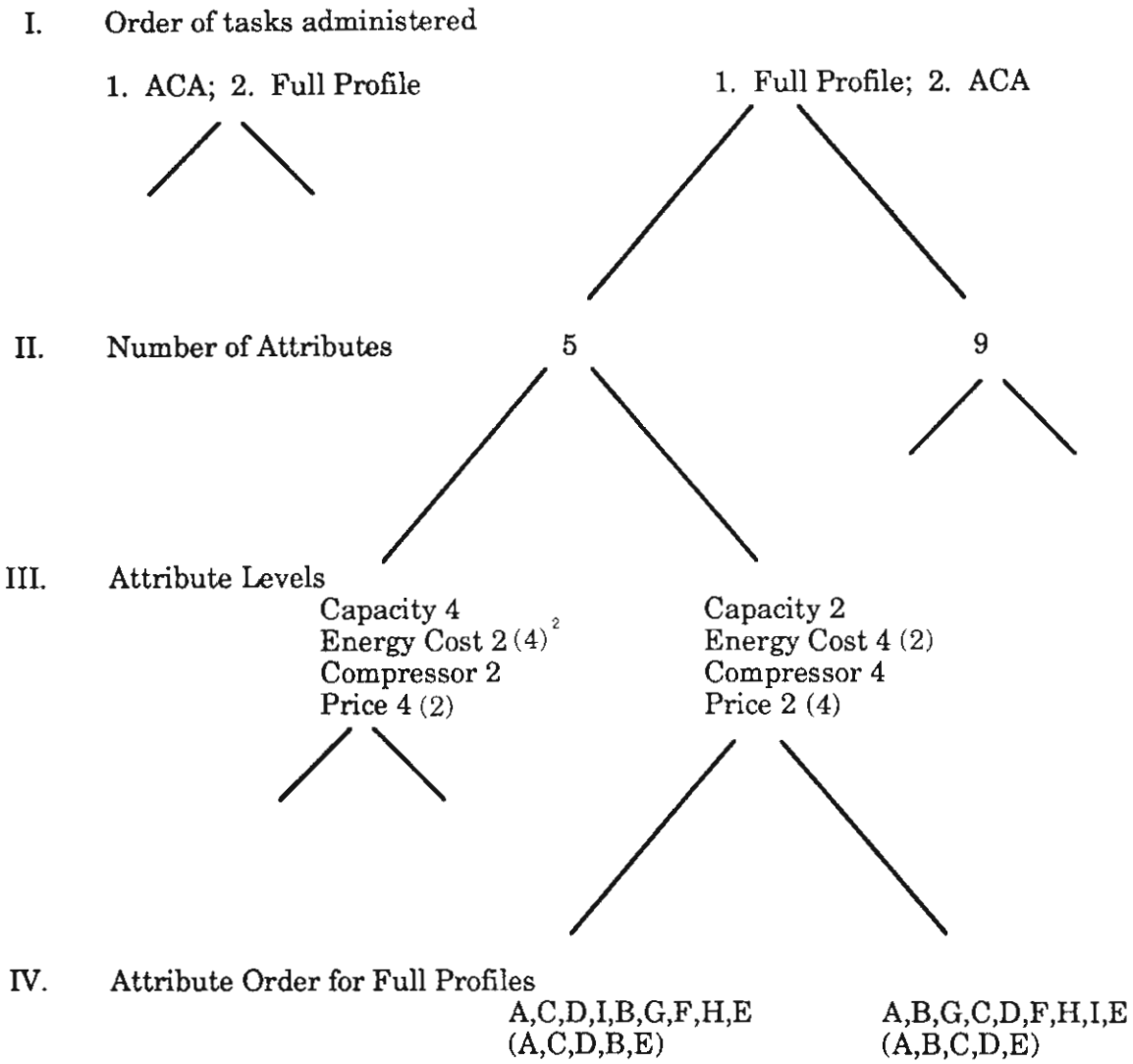
A third manipulation consisted of the number of attribute levels used for four of the attributes. These four attributes — capacity, energy cost, compressor noise and price — were included in all cells of the experimental design. For example, in the full profile design, half the respondents were exposed to four capacity, four price, two energy cost and two compressor levels. The other half saw two capacity, two price, four energy cost and four compressor levels. For all four attributes, respondents' preferences for the levels were assumed to be monotone. Extreme levels were held constant in this manipulation. The primary reason for this manipulation in the study is the expected effect on derived attribute importances (Wittink 1990). We do not address results on the issue of attribute level effects in this paper.

The fourth manipulation involved the order in which the attributes were listed in the full profile method. Attribute order effects have been reported by Johnson (1989). Attributes may receive more attention from respondents if they are listed first (or last) than if they occur closer to the middle of the list. A direct way to investigate this would be to shuffle the attribute order. However, some attributes may have a "natural" place in the order. We decided that all conjoint tasks should have brand name as the first and price as the last attribute. Thus, we manipulated only the order of the remaining attributes. We note that this constraint reduces the opportunity for systematic order effects to occur. (Our reason for this constraint is that in this study we were not interested in learning the maximum magnitude of such an attribute order effect in the full profile method. Rather, we wanted to maintain realism in the task characteristics while allowing for an effect to exist. Of course a validation task that has exactly the same attribute order as the conjoint task would show inflated validities if 1) there is an attribute order effect, and 2) the order in which attributes are attended to in the marketplace differs from the order used in the conjoint method.) No systematic effects of this manipulation on predictive validity were detected. An overview of the experimental design is shown in Figure 1.

No attribute order manipulation is indicated for the ACA method. The reason for this is that the order (and identity) of attributes varies between the pairs of profiles (and between respondents). That is, there is no fixed attribute order for the paired profile intensity ratings. The order in which attributes are evaluated in the self-explicated part of ACA is the same as it is in the full profiles for half the respondents. However, the order of the attributes for the objects in the validation choice tasks differs from either order used for the full profiles (and hence for the self-explicated part of ACA).

Figure 1

Experimental Design[1]



[1] The design is a 2^4 design with 16 cells

[2] The number in parentheses is the number of levels used in ACA.

Respondent Selection. To obtain broad geographic representation, 400 respondents were selected from a total of 11 cities: Baltimore, Charlotte, Cincinnati, Cleveland, Colorado Springs, Denver, Detroit, Los Angeles, Milwaukee, Philadelphia and Washington, D.C. Each city contributed 36 or 37 respondents. Respondents were selected by intercepting shoppers in super-regional malls. Given four manipulations with two options each, every one of the sixteen cells in the experimental design has 25 respondents. The task characteristics faced by respondents were varied systematically according to the experimental design. This minimized the chance of having systematic differences in respondent characteristics between experimental cells.

To participate in the study, respondents had to have a refrigerator in their home. They were recruited to central locations and were paid $5 each as compensation for their time. All questions, including those for the full profile task, were asked via computer. Prior to the conjoint tasks, respondents provided information about the brand of refrigerator they owned as well as its age. They were then told to imagine they were shopping for a new refrigerator.

For the full profiles, we used a buying likelihood question: "How likely would you be to buy this refrigerator if you were to buy a new refrigerator today?" A nine-point scale was used (1 = 10% or less (not at all likely); 2 = 20% ; ... ; 9 = 90% or more (extremely likely)).

The ACA task followed the standard format of 1) eliciting preferences for the levels of each attribute, if necessary, 2) obtaining importances for the difference between best and worst levels for the attributes, and 3) preference intensity ratings for one of two refrigerators in 12 pairs for designs with 9 attributes, and 6 pairs for designs with 5 attributes. The preference ratings were obtained by asking: "Which refrigerator do you prefer?" A nine-point scale was used varying from 1 = strongly prefer left, to 9 = strongly prefer right.

Validation Tasks. The validation tasks consisted of the following: For each of two pairs of refrigerators, the respondent was asked to indicate which one he or she would be most likely to buy. And, for each of two triples of refrigerators, the respondent indicated which refrigerator he or she would be most likely and (of the remaining two) least likely to buy. Each respondent provided validation data twice (on the identical set of pairs and triples). The validation tasks were separated by the second conjoint task. One important reason for collecting two sets of validation data is that it allows us to assess the reliability or consistency of the choices provided. The predictive validity of either conjoint task cannot exceed the reliability of the choices (Johnson 1989).

In the validation tasks, all refrigerators were defined in terms of the five common attributes. These attributes were listed in the order A, B, E, C, D (see Figure 1). For the attributes with level manipulations, only the two extreme levels (which were included in all cells of the experimental design) were used to define the refrigerators. The refrigerators within each choice set were defined at approximately equal expected average overall utilities, based on our prior judgment.

At the end of the interview session, respondents were asked to answer nine questions about the two conjoint tasks. The questions asked whether the task "was enjoyable," "was easy," "was realistic," "allowed me to express my opinions," "took too long to do," "was frustrating to do," "asked about too

many refrigerators," "made me feel like just pushing the numbers to get done," and "had too many features to consider at once." In addition, respondents provided age and family income information.

RESULTS

For each of 393 respondents (7 respondents provided incomplete data) we estimated two partworth models: one based on the full profile evaluations and another based on the ACA responses. Partworths were obtained based on ordinary least squares. We show results for the holdout choice tasks in Table 3. To provide an overall indication of the predictive validities we converted each triple to three pairs. Thus, with two pairs and two triples twice evaluated we have sixteen pairs per respondent. In this overall sense, full profile has a 68.5 percent and ACA a 72.6 percent hit rate. This difference is statistically significant ($p < .01$).

Table 3

Predictive Validities for Each Method

		Method		
		Full Profile	ACA	p-value[1]
Overall[2]		.69	.73	<.01
Task Order				
First		.64	.73	<.01
Second		.74	.73	-
Number of Attributes				
Nine		.66	.71	<.05
Five		.71	.75	<.05
Consistency	**Sample Percent**			
≥4	13	.58	.59	-
5 or 6	27	.60	.69	<.01
7	29	.70	.74	<.10
8	31	.79	.80	-
	100			
Choice Task				
Pairs from Triples		.68	.73	<.01
Pairs		.70	.72	-
Triples				
Most likely		.54	.63	<.01
Least likely		.59	.62	-

[1] The p-value is the level of statistical significance for the null hypothesis of no difference between the methods.

[2] For all calculations, except for Triples, the triples were converted into three pairs each. Thus, with choices on two pairs and two triples being provided twice, we have sixteen pairs per respondent.

We have also decomposed the hit rates in several ways. Given that there may be a systematic task order effect, we show the hit rates separately for the estimated part-worths obtained from the first and the second conjoint task completed by the respondent. Apparently, the task order has no effect on the ACA hit rates. On the other hand, the full profile hit rates are 64 percent if ACA is completed first and 74 percent if it is the second task. One interpretation of this difference is that the ACA task provides a training opportunity that enhances the consistency of subsequent full profile evaluations. Of course, in practice one would use either full profiles or ACA, and in that context only the results for the first task are relevant. Thus, ACA is substantially better if only one conjoint task is administered to each respondent. It is conceivable, however, that the full profile method improves if a practice or warm-up task is administered first.

The next item in Table 3 is a breakdown by number of attributes. Half the respondents did both conjoint tasks for five attributes, and the other half did both for nine attributes. All choice tasks involved the five common attributes only. The results indicate that the predictive validity is reduced when the conjoint tasks involved four additional attributes. This attenuation is similar for the two conjoint methods. For both five and nine attributes ACA has higher predictive validities. We note that with nine attributes and sixteen profiles the partworth model is highly saturated. Nevertheless, the increase from five to nine attributes apparently does not reduce the predictive validity more for full profiles than for ACA.

We have also grouped respondents by the degree of consistency in choices. We have choices for each of eight pairs twice, for each respondent. In Table 3 we show predictive validities separately for respondents with low (four or fewer), moderate (five or six), high (seven) and perfect (eight) consistencies. If the consistency is low (unreliable choices), the predictive validities are equally low for the two conjoint methods. Similarly, for respondents with perfect consistency (31 percent of the sample) the predictive validities are equally high. The differences in predictive validity occur for respondents with moderate and high consistencies in choices. One interpretation of this result is that for respondents with moderate to high consistency in their choices (56 percent of the sample) ACA better captures their preferences because of the way in which it simplifies and breaks down the preference elicitation task.

We also note that as the consistency in choices increases, the predictive validities increase for both conjoint tasks. This is in part because 50 percent is the best that can be done by a respondent who is entirely inconsistent. One way to adjust the predictive validities for choice inconsistencies is the following:

For all choices and respondents, the average number of consistent choices equals 6.44 out of 8 (or 80.5 percent). We want to find the hit rates for the two methods when the choices are perfectly consistent. This means adjusting the observed hit rates of .685 for full profile and .726 for ACA by the inconsistent choices for which the hit rates are necessarily .5.

Let (p con) (adj hit_i) + (1-p con) (.5) = hit_i (1)

where: p con is the proportion of consistent choices (80.5);
adj hit_i is the hit rate adjusted for inconsistent choices for method i;
hit_i is the observed hit rate for method i;
i = 1, 2.

Solving equation (1) for the adjusted hit rates gives .730 for full profiles and .781 for ACA. This adjustment for inconsistent choices increases the predictive validities for both methods and increases the difference slightly.

So far we have made no distinction between the pairs and triples. There are several reasons why it is useful to examine the triples separately. One, real world choices are not restricted to pairs. Thus, it is of interest to determine the predictive validity of a method when respondents choose, say, the best out of three options. Two, in ACA the paired comparison intensity ratings may be similar to the pairs in the choice tasks.

The separation of the percent hits for all sixteen pairs into twelve pairs constructed from the triples and four pure pairs shows that ACA has higher hit rates for both the pairs derived from the triples (.73) and the pure pairs (.72). However, the difference is more dramatic for the triple-based pairs. Also, the full profile data provide higher hit rates for the pure (.70) than for the triple-based pairs (.68), while the opposite is true for ACA. Overall, these results suggest that ACA has a greater edge over full profiles when the choice task consists of many alternatives compared with a choice task involving two alternatives.

Instead of creating three pairs from each triple, we can also examine the performance of the two methods in a) predicting the most likely to purchase, and b) predicting the least likely to purchase. Here the difference between full profiles and ACA is especially dramatic for the most likely to purchase item (.63 for ACA and .54 for full profiles). For the least likely to purchase item ACA also achieves a higher hit rate, but the difference is not as large (.62 for ACA versus .59 for full profiles). Of course predicting marketplace behavior involves picking the best item, and hence, the results for the most likely item are the best indicator of relative performance for marketplace predictions.

Finally, we show the average respondent perceptions of the two tasks on nine scales in Table 4. On several scales the average difference between the methods is very close to zero. For example, the responses leaned, on average, toward agreement on the question of task realism (both at 6.29) and toward disagreement on the question whether the task asked about too many refrigerators (3.57 and 3.55) for full profiles and ACA. However, ACA achieved a higher average score on "being enjoyable" ($p < .05$) and a lower average score on "taking too long to do" ($p < .05$). The only other difference approaching statistical significance is that full profile achieved a higher score on "being easy" ($p < .15$).

Table 4

Respondent Perceptions of the Conjoint Methods

	Method			
	Full Profile	ACA	Difference	t
The task . . . [1]				
- was enjoyable	5.55	5.81	-0.26	-2.51
- was easy	6.78	6.57	0.21	1.59
- was realistic	6.29	6.29	0.00	0.02
- allowed me to express my opinions	6.43	6.44	-0.01	-0.07
- took too long to do	4.13	3.86	0.27	2.29
- was frustrating to do	3.25	3.31	-0.06	-0.60
- asked about too many refrigerators	3.57	3.55	0.02	0.19
- made me feel just like pushing the numbers to get done	3.47	3.36	0.11	0.95
- had too many features to consider at once	3.77	3.74	0.03	0.20

[1] Responses to each question were obtained on a nine-point scale (1 being not at all agreeing and 9 being very much agreeing).

CONCLUSIONS

We provide evidence on the predictive validity of full profiles and ACA. In our study half the respondents considered questions involving five attributes (four monotone), while the other half had information on nine attributes (seven monotone). The advantage of varying the number of attributes is that it allows us to see if the relative performance of the methods changes with changes in the number of attributes. In either case, however, the number of attributes falls short of the number for which Green and Srinivasan (1990) suggest procedures such as ACA may have an advantage over full profiles.

For both five (when full profile is recommended) and nine attributes we find that ACA outperforms the full profile method. We have broken down the overall performance results in a number of ways, and observe that ACA maintains an edge each time. The difference is greatest when 1) we use the partworths estimated from the conjoint task performed first by the respondents, 2) we focus on the respondents for whom the holdout choices show moderate to high consistency, 3) we use the triples in the holdout choice tasks, and 4) we consider the predictive validity for the most likely option in the triples. We also have a higher average score for ACA than for full profiles on the task "being enjoyable" and a lower average score on the task "taking too long to do."

We intend to do a more detailed analysis of the results, for example to document the reliability (statistical significance) of all observed differences. There are also additional experimental manipulations and other effects that still need to be investigated. So far, however, the results quite clearly show that ACA outperforms the full profile method. Of course, the results of this study cannot be assumed to generalize to all contexts for which compositional or decompositional preference measurement is considered. The selection of the product category and the respondents may have a bearing on the results. Also, we administered the full profile method by computer. Seeing and evaluating one profile at a time may make it more difficult for respondents to be consistent in the full profile evaluations. And the finding that the hit rate for full profile is much higher when it is used after ACA suggests that a warmup card sort or attribute importance task may improve the predictive validity of full profile.

REFERENCES

Agarwal, Manoj K. and Paul E. Green (1989), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results," Working paper, State University of New York at Binghamton, December.

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, October.

Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing Research: New Developments and Directions," *Journal of Marketing*, 54 (October), 3-19.

Green, Paul E., Abba M. Krieger and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research* (in press).

Johnson, Richard M. (1987), "Adaptive Conjoint Analysis," in: *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, July, 253-65.

Johnson, Richard M. (1989), "Assessing the Validity of Conjoint Analysis," *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, June, 273-80.

Johnson, Richard M. (1991), "Comment on Green *et al.*," *Journal of Marketing Research* (in press).

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

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.

Wittink, Dick R. (1990), "Attribute Level Effects in Conjoint Results: The Problem and Possible Solutions," *First Annual Advanced Research Techniques Forum Proceedings,* Chicago: American Marketing Association, forthcoming.

Wittink, Dick R., Lakshman Krishnamurthi and David J. Reibstein (1989), "The Effect of Differences in the Number of Attribute Levels on Conjoint Results," *Marketing Letters,* 1, 2, 133-23.

Comment on Huber, Wittink, Fiedler, and Miller

Richard D. Smallwood
Applied Decision Analysis, Inc.

This is a very good paper indeed. The authors are to be congratulated on a thorough piece of research that will prove very useful for the practitioner. The market research field in general is full of woodsy lore and rules of thumb. Papers such as this one provide an empirical basis for developing well-thought-out and logical rules for the design of market research programs.

In the discussion that follows, I want to raise several questions that occurred to me as I read the paper. These are not criticisms, but rather questions that the authors might consider in further work, or perhaps in modifying the paper.

My first concern deals with the amount of data that is applied to each of the two techniques, ACA and Full Profile. Let us conjecture that the accuracy of both ACA and Full Profile will improve as the amount of data used to estimate the parameters increases. Imagine a graph of accuracy versus amount of data with the accuracy gradually increasing to some asymptotic value. The question addressed by the paper is whether the asymptotic accuracy of ACA or Full Profile is higher than the other. My concern is that we may be operating at a different point on the x axis for the two techniques. For example, it appears that there are 15 parameters for the nine-attribute situation in the paper. For the Full Profile technique only 16 comparative questions were asked, while for the ACA technique there are 12 paired comparisons plus approximately 18 level questions and 9 attribute importance questions. Thus, there are considerably more data for the ACA than for the Full Profile technique.

Secondly, each of the Full Profile questions was asked in comparison to a "not buy" alternative. In effect this is adding another parameter to the Full Profile technique thus contributing further to the possibility of unequal amounts of data for the two techniques.

In addition, a Full Profile advocate might complain that this is not a real test of Full Profile since there are no direct comparisons of product profiles, only comparisons against the "not buy" alternative. It would be interesting to see whether a Full Profile card sort can be simulated on a computer.

I found it surprising that the performance of both techniques improved as the number of attributes decreased from nine to five. One possible explanation for this result is the small ratio of data to free parameters in both techniques. By keeping the amount of data constant and reducing the number of parameters, the accuracy of the parameter estimates should increase. Thus, it is possible that the improved accuracy of the smaller number of parameters more than compensates for the greater explanatory power of the four additional attributes.

The observation by the authors that Full Profile seems to perform better when it is the second technique rather than the first technique has one possible explanation. It has been conjectured (and I have observed in separate observations) that respondents in a Full Profile study tend to focus on just a few attributes to keep the comparison tasks relatively easy. However, a typical respondent might require several questions before settling on the two or three attributes that are most important. This "settling in" process would naturally cause inconsistencies in the initial tradeoff questions, which would cause less reliable parameter estimates. On the other hand, if the respondent did the ACA process first, it would be much easier to identify the most important attributes and hence produce more consistent answers to the Full Profile questions.

A final minor question: One of the arguments often used to support use of the Full Profile technique is its ability to identify interactions among the attributes. I wonder if the authors have had a chance to test the data for these interactions.

Finally, let me reiterate my introductory comments. I found this to be a very interesting and useful paper and applaud the authors for their dedication and perseverance in a very difficult subject area.

UNRELIABLE RESPONDENTS IN CONJOINT ANALYSIS: THEIR IMPACT AND IDENTIFICATION

Keith Chrzan
Walker: Research and Analysis

INTRODUCTION

As defined by Green and Srinivasan (1990) "Conjoint analysis is any decompositional method that estimates the structure of a consumer's preferences given his/her overall evaluations of a set of alternatives that are prespecified in terms of levels of different attributes." For example, in studying consumers' preferences for home stereo systems, a researcher might expose consumers to stimuli involving various combinations of levels of price (e.g., $150, $500, $1,000), audio power (5 watts, 50 watts and 125 watts) and feature package (presence/absence of a compact disc player). From a respondent's evaluation of the various stimuli the researcher can estimate the utility, or preference, that the respondent attaches to each level of price, audio power and features.

Conjoint estimation procedures can present these attribute level utilities as either discrete utilities for each level of a attribute (a qualitative or partworth model) or as a utility function (a quantitative model) that fits the derived utilities to a continuous linear (vector) or quadratic (ideal point) equation (Green and Srinivasan 1978, 1990). The following discussion focuses upon the partworth model but it applies to studies employing utility function models as well.

Partworths are the basic unit of analysis in a conjoint study. Directly computable from them are attribute utility (usually the partworth for the most preferred level of an attribute minus partworth for the least preferred level of that attribute) and attribute importance (utility of given attribute divided by the sum of all the attribute utilities). In addition, partworths can be combined to estimate the total utilities for real or imaginary objects. Objects are constructed by specifying a level for each attribute and then summing the partworths for the levels. A major strength of conjoint analysis is this ability to estimate respondents' preferences for object configurations not even included in the experimental design. Commercial conjoint packages go one step further, allowing the researcher to run simulations that estimate the market (or preference) shares that would result if a specified set of objects were in competition (Green and Srinivasan 1990).

Another common use for partworths is market segmentation (Wittink and Cattin 1989). Partworths (or attribute utilities or importances) become the variables for cluster analysis and segments of subjects with similar preference structures result (Currim 1981, Green, Wind and Jain 1972).

All of these further analyses depend upon conjoint analysis as a psychometric technique.
If individuals' partworth estimates contain error, they can pollute these subsequent analyses.
An "unreliable" respondent in a conjoint study is one whose conjoint estimated evaluative model does not adequately fit his actual evaluative framework: his estimated level utilities contain too much

error. This may occur simply because the conjoint task bores or confuses the respondent or because the respondent perversely provides inconsistent data; either way he has not responded accurately to the stimuli of the conjoint study. It may also occur because the respondent's actual evaluative framework contains complexities that the usual additive main-effects-only conjoint model does not take into account. For instance, a respondent's actual composition rule for building level utilities into full profile utilities may be multiplicative or may contain significant interactions.

As a working definition, call a respondent unreliable if the conjoint model fails to fit his evaluations of the stimuli significantly better than it would have fit random data. Intuitively, respondents not performing better than chance likely are not well measured.

Section I provides techniques for identifying unreliable respondents. Section II compares reliable and unreliable respondents from 14 recent conjoint studies to test the hypothesis that they produce different results, both in terms of the fundamental partworth utilities and of the subsequent simulation and segmentation analyses. Finally, Section III uses 2,948 respondents from 12 full-profile conjoint studies to explore the parts that study design features and estimation program selection play in respondent reliability.

I. Identifying Unreliable Respondents

How one discovers unreliable respondents depends upon the conjoint estimating algorithm one uses. Probably the three most frequently used conjoint algorithms are ordinary least squares regression (OLS), linear programming (LINMAP) and adaptive conjoint analysis (ACA).

OLS

Multiple regression theory provides the tools with which one can identify a critical R^2 value for any desired level of significance (Cohen and Cohen 1983). The standard formula for determining significance of a multiple regression with n observations, k independent variables and observed multiple coefficient of determination R^2 is:

$$F_{a,k,n-k-1} = \frac{R^2(n-k-1)}{(1-R^2)k}$$

Solving for R^2 yields:

$$R^2 = \frac{F_{a,k,n-k-1}(k)}{(n-k-1)+F_{a,k,n-k-1}(k)}$$

Variables n and k come from the specific experimental design: n is the number stimuli and k is the sum across attributes of (the number of attribute levels minus 1). Specifying the significance level (α) fixes the value of F and thus of the critical R^2. The critical values for R^2 at α=0.25, α=0.10 and α=0.05 for each entry in Addelman's (1962) catalog of orthogonal main effects experimental designs appear in Exhibit I. If a respondent's R^2 exceeds the given critical value, the respondent qualifies as reliable; if not, the respondent is unreliable.

Three common OLS programs are Bretton-Clark's Conjoint Analyzer, PC-MDS's Conjoint and SPSS's Conjoint. The latter two provide the coefficient of multiple correlation, R, as printed output for each respondent, so that identifying unreliable respondents is fairly straightforward. Unfortunately, none of the three OLS programs save the respondents' Rs into the respondent utility file, leaving a laborious task for the researcher.

A benefit of OLS conjoint estimation is that it allows one to compute the standard error around each partworth for each respondent, as in SPSS (Green and Srinivasan 1978). The standard errors for the partworths of respondents unreliable at α=0.05 usually exceed in magnitude the partworths themselves, a good reason in itself for treating unreliable respondents warily. Both for this reason and because it is a *de facto* standard of type I error protection, the analyses below employ the α=0.05 criterion in the identification of unreliable respondents. Performing the same analyses at the more generous α=0.25 criterion did not alter any of the below conclusions, although it did reduce the number of respondents classified as unreliable by about half. Conjoint respondents are expensive, so in practice the researcher may want to be less conservative than α=0.05 and allow a greater percentage of respondents to qualify for subsequent analysis as reliable.

SPSS also displays a misleading measure of "significance": it shows the significance of the bivariate correlation (r, not R) between the respondent's vector of evaluations for each stimulus and the vector of evaluations predicted by the model for the given respondent. This measure of significance utilizes a t-test with n-2 degrees of freedom, however, instead of the proper k and n-k-1 (Cohen and Cohen 1983). By incorrectly computing the significance of a simple linear correlation instead of that for a multiple correlation, SPSS's "significance" is usually dramatically, sometimes maximally, overstated; it is best ignored.

LINMAP

Inversely analogous to regression's R^2 are LINMAP's two indices of poorness-of-fit. One, B, is simply the percent of paired comparisons that are predicted differently by the conjoint model than are actually made by the respondent (Srinivasan and Shocker 1973b). It is B that is produced by Bretton-Clark's LINMAP program (Bretton-Clark 1989) and conveniently stored in each respondent's record in the utility file. The other measure, C*, is related to B as follows (Srinivasan and Shocker 1973b):

$$C^* = \frac{B}{1+B}$$

Originally designed for ordinal scale stimuli evaluations, LINMAP almost automatically counts ties as incorrect predictions. Therefore, indices of fit B and C* will differ for ranking and rating measures of respondent evaluations of stimuli (Bretton-Clark 1989), for rating measures with different numbers of points on the scale, and for rating scales displaying different distributions.

No formula exists for computing the statistical significance of LINMAP's index of fit. Fortunately, Monte Carlo studies are easy to perform to find the appropriate cutoff percentile (e.g., 10th or 5th) for either B or C*. Note the use of the left tail percentiles (e.g., 5th instead of 95th); B and C* are poorness of fit statistics: the higher their values, the worse the fit.

By running LINMAP with a given experimental design and an appropriate set of random data in place of actual responses and examining the "Distribution of Violations" screen in Bretton-Clark's LINMAP, one can readily discover the value for B above which (e.g.) 90 or 95% of sets of random responses reside. These become the cutoffs.

Gary Mullet and Marvin Karson (1986) have done the Monte Carlo work for some of Addelman's (1962) orthogonal main effect designs for rank order data. They tabulate the values of C* for selected percentiles of the distributions that derive from sets of randomly ranked inputs. As noted, however, Bretton-Clark's LINMAP displays B. The above equality of B and C* can be restated to solve for B as follows:

$$B = \frac{C^*}{1-C^*}$$

In this way, users of Bretton-Clark's LINMAP program can transform the appropriate critical value from Mullet and Karson's tables to use as the cutoff for the B statistic from their LINMAP program. Exhibit II tables the α=0.50 α=0.10 and α=0.05 critical values for B for the designs covered by Mullet and Karson.

For studies using rank order data but with designs other than those in Mullet and Karson (1986) or using rating scale measures of stimuli evaluations, the researcher must conduct his or her own Monte Carlo study, as described above.

ACA

Unlike OLS and LINMAP, ACA does not provide an index of fit statistic for the experimental data. Instead, it provides information on the bivariate correlation between actual and predicted respondent evaluations of a holdout sample of stimuli (in the "Calibration" section), and appends this information to each respondent's utility file (Sawtooth Software 1988).

OLS programs from SPSS and Bretton-Clark also allow one to add holdout profiles to the experimental design and to examine the correlation between actual and predicted evaluations of the holdout profiles for each respondent; Bretton-Clark's LINMAP program does this as well.

SPSS displays the correlation between actual and predicted respondent evaluations of holdout profiles and another flawed measure of significance of the bivariate r: this time SPSS correctly uses the critical t values for a one-tailed t-test but ignores the sign of the correlation, thus creating a mixed directional and non-directional t-test. The correlation seems correct, however, so one can determine the significance of the respondents' correlations.

The critical values for r shown in Exhibit III use the standard formula for a t-test of correlation

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

(where n is the number of holdout profiles), but solves for r:

$$r = \frac{t}{\sqrt{t^2+n-2}}$$

Exhibit III tables critical values for one-tail t-tests at $\alpha = 0.25$ $\alpha = 0.10$ and $\alpha = 0.05$.

The Exhibit can also be used for the Spearman rank order correlation coefficient appended to respondent's utility files in LINMAP, if profiles were ranked. If they were rated, Spearman's r might differ from the Pearson r used in the table.

One of the 12 OLS studies examined included three holdout profiles and provided the data for a crosstabulation of reliabilities, first as defined by the multivariate R^2 from the OLS estimation equation and then as defined by the r for the holdout profiles. The result was the 2 X 2 contingency table in Exhibit IV which included a non-significant and weak (φ=0.13) relationship between the two measures. The little available data suggests that the two measures of respondent reliability are not highly correlated, though including a larger number of holdout profiles may have defined respondent reliability more similarly to the OLS R^2 measure.

As noted, ACA also stores the Pearson r for actual *versus* predicted evaluations of holdout profiles. Unfortunately for this purpose, ACA presents the holdout profiles in increasing order of predicted utility, tells the respondent so, and will not allow the respondent to give a lower rating to one holdout profile than to a previous one (Sawtooth Software 1988). Thus respondents are constrained to be consistent and actual and predicted preference for holdout profiles will be monotonic transformations of one another. The critical values for r shown in Exhibit III, then, understate the critical r necessary for an ACA respondent to qualify as reliable. For use with ACA the column headings $\alpha = 0.25$ $\alpha = 0.10$ and $\alpha = 0.05$ should read instead $\alpha > 0.25$, $\alpha > 0.10$ and $\alpha > 0.05$.

These profiles are not included in the interview to provide a measure of reliability, but rather to provide information for the scaling of utilities. Sawtooth Software believes that the calibration is better if the respondent is coached about the ascending desirabilities of the profiles, which, of course, makes the data less appropriate as a measure of reliability. They recommend that holdout concepts be included elsewhere in the interview to assess the quality of each respondent's data.

Finding exact cutoffs for ACA's r would involve manual entry of random responses to ACA questions, a process that is almost certainly too laborious to justify. The extent that using Exhibit IV with ACA tends to overestimate respondent reliability may be ameliorated by increased respondent interest in the interactive conjoint task: it is almost as easy for a respondent to provide reasoned responses to the ACA survey as random ones, whereas it is much easier for a card sort respondent to shuffle the deck of profile cards than to sort it carefully. One can also correct for ACA's biased estimate of r by choosing a conservative cutoff (e.g., corresponding to α=0.01 instead of 0.05).

II. The Impact of Unreliable Respondents

Differences in Partworths

Unreliable respondents produce different mean partworths than do reliable respondents. Exhibit V shows the results of a MANOVA of mean partworths by reliable/unreliable for a recent OLS rank order conjoint study on a new generation of durable medical instruments. Nine of the 16 partworths have significant ($p<0.01$) main effects for respondent reliability. The multivariate test was also significant at $p<0.01$. Significant main effects for respondent reliability occurred in such MANOVAs of OLS partworth estimates in 11 of the 12 studies.

Note that for each partworth in Exhibit V with significant main effects, the absolute value of the partworth for reliable respondents exceeds the absolute value for unreliable respondents. The same occurs in the other two studies examined with sample size in excess of 200. Partworths follow this pattern less strongly in studies with smaller samples. What may be happening is this: some of the respondents are providing near-random responses to the stimuli and thereby producing partworths with means of 0 and large standard errors. In a large sample these errors will cancel and the random data will drag the mean partworths toward 0. With smaller samples, however, the errors will not always cancel and the absolute values of mean partworths for unreliable respondents will not always be less than for those of reliable respondents. Unreliable respondents tend to dilute the true group mean partworths but add error in any case. The same patterns occurred in the LINMAP and ACA analyses shown in Exhibits VI and VII.

LINMAP produces similar results. Only one of the rating LINMAP conjoints examined produced the thirty or so reliable respondents to expect the ANOVAs to have enough power to detect differences. Two of its 25 partworths had significant ($p<0.01$) differences and the multivariate test showed significance at $p<0.01$. Of the rank order LINMAPs, three out of the six had enough respondents in both reliable and unreliable categories to detect significant differences of reasonable effect size.

Two of the three produced significant main effects of reliability in MANOVAs. Exhibit VI shows the MANOVA of mean partworth by respondent reliability; eight of 16 partworths exhibited significant main effects for reliability (at $p<0.05$). The multivariate test was significant at $p<0.01$.

The same occurs for ACA, two data sets for which were available. Each included five holdout or "Calibration" profiles so the cutoff for $a>0.05$ is 0.8053. At this level 88% of respondents qualified as reliable in one study and 81% in the other. Significant differences due to respondent reliability occurred in 18/37 partworths in the first study and 22/50 in the second. Recalling the artificial positive bias in r imposed by ACA, the MANOVAs were rerun at $a>0.01$. This time 57% and 64% of respondents qualified as reliable and 18/37 and 19/50 partworths had significant main effects of respondent reliability. The MANOVA for the latter study appears in Exhibit VII.

These differences in partworths between reliable and unreliable respondents are reflected in subsequent analyses. That differences in partworths can lead to differences in attribute utilities and importances is obvious, an exercise left to the interested reader. Examples of differences that arise in segment membership, simulation shares, estimating program and data collection method appear below.

Differences in Segment Membership

Two of the studies involved segmentation on the basis of partworths. Crosstabulating segment membership by respondent reliability, where segments were defined by cluster analysis of the partworths generated by the SPSS-PC conjoint program, one of the studies exhibited a χ^2 statistic of 34.88, significant at $p<0.01$ (see Exhibit VIII). Thus one can reject the hypothesis that segment membership and respondent reliability were independent. Rerunning the cluster analysis with just reliable respondents produced the same three clusters, though clearly this need not have been the case. While significant, the differences between cluster membership for reliable and unreliable respondents were not dramatic.

Differences in Simulated Shares

With respect to market simulations, one of the studies featured a brand-anchored conjoint (Louviere and Johnson, forthcoming). Respondents were medical professionals who recommend the conjoint's product category. Unfortunately, this was the one study wherein no significant differences in mean partworths between reliable and unreliable respondents emerged. The relative predictive validity of the two groups' models thus are not compared.

All 14 data sets, however, can support simulations. As noted, with some of the unreliable respondents providing near random responses, the means of the unreliable respondents partworths tend to be closer to zero than those of reliable respondents. Using $\alpha=0.05$ to define reliability, Z tests of proportions yielded significant differences in simulated shares between reliable and unreliable respondents sometimes, but infrequently. As one becomes less strict in defining unreliability (by using a larger α, say 0.25) fewer respondents qualify as unreliable and truly random responders

make up a greater proportion of the unreliable group than at smaller values of α. In such cases, differences between simulated shares of unreliable respondents and those of reliable respondents seem to become more common, though not many of the 14 data sets provided large numbers of respondents unreliable at α=0.25.

Differences in Estimation Programs

In 10 of the 12 full-profile studies LINMAP produced fewer reliable respondents than did OLS. Due to the positive bias in ACA's estimate of r, counts of reliable respondents are not directly comparable to those for OLS and LINMAP. OLS produced about four times as many reliable respondents as LINMAP for studies with rated dependent variables and about one and a half times for studies with rank-ordered dependent variables. In all four of the studies so analyzed, LINMAP and OLS assignments of reliability were highly and significantly correlated. Exhibit IX, for instance, shows a single set of respondents and crosstabulates their reliability as LINMAP respondents by their reliability as OLS respondents. With 85% of respondents classified equivalently by the two programs, the correlation is 0.71 and the hypothesis of independence can be rejected at $P<0.01$ (χ^2=368.46).

Differences in Data Collection Technique

Finally, unreliable respondents cause some of the differences between different methods of data collection. An earlier study (Chrzan 1990) showed that most of the significant differences between mean standardized partworths of card sort and Likert rating scale versions of two conjoint studies owed to the effects of respondent reliability and not to the method of data collection (card sort versus survey).

III. Sources of Respondent Reliability

For exploratory purposes stepwise logistic regressions used respondents' reliability in the 12 conjoint studies as DVs and various design features as IVs to discover the variables that affect respondent reliability (best possible subset regressions produced similar results). Independent variables included:

Ratio: the ratio of the number of stimuli to the degrees of freedom used in the model = n/T; this variable is suggested by several researchers as influencing the accuracy of the OLS estimation model (Green and Srinivasan 1989, Hintze 1990, Mullet 1989).

Difference: residual degrees of freedom = n - T.

Convention:	"1" if the conjoint task occurred at a convention, "0" if not
Focus:	"1" if the conjoint task occurred at a focus group, "0" if not
OLS:	"1" if OLS defined unreliability, "0" if LINMAP
Card sort:	"1" if the conjoint task involved an ordinal card sort, "0" if not
DV Range:	number of points on the dependent variable scale

Splitting the sample and examining OLS and LINMAP separately shows that variables impact respondent reliability differently for the two estimating programs. The predictive equation for OLS was:

-2.24 + 0.86 (Focus)
+ 0.73 (Ratio)
- 0.29 (Convention)
+ 0.11 (Difference)
+ 0.04 (DV Range)

whereas that for LINMAP was

3.99 + 3.06 (Card sort)
- 2.82 (Ratio)
- 2.71 (Convention)
+ 2.55 (Focus)
- 0.11 (DV Range)

With 1,200 cases, the binomial logit analyses were both significant at $p < 0.01$.

Convention earns a negative coefficient for both techniques. Conjoints performed as parts of larger studies at conventions suffered from lack of room for the respondents to work: more often than not, respondents stood up to sort decks of 16 - 18 cards without anywhere to set them down. Not surprisingly, conjoints performed at conventions had lower proportions of reliable respondents.

The reverse occurs for the conjoints performed at focus groups. At focus groups, time is specifically allotted for the conjoint task, and help is provided for respondents who fail to comprehend the task. Also, it is as difficult for a subject to feign compliance as to comply, and the result is a greater proportion of reliable respondents.

An ordinal preference measure obtained via card sorting greatly increased the probability of LINMAP subjects' reliability but affected that of the OLS subjects only slightly. LINMAP performs better on the data for which it was designed (ordinal) than for interval data. OLS performs only slightly better on interval data than on ordinal.

The most interesting result concerns the effect of the n/T ratio, which enhances reliability for OLS but decreases it for LINMAP. Although this finding occurred every way the data were analyzed, it is counter-intuitive. Indeed, at a ratio of 1.0, no LINMAP respondents can be reliable.

SUMMARY

Not all respondents in conjoint studies provide reliable data. One can assess respondent reliability using any of the major conjoint algorithms OLS, LINMAP or ACA. The partworths from unreliable respondents differ from those of reliable respondents and this leads differences in the subsequent analyses founded upon the partworths.

REFERENCES

Addelman, Sidney (1962) "Orthogonal Main Effect Plans for Asymmetrical Factorial Experiments," *Technometrics*, 4 (February), 21-46.

Bretton-Clark (1989) *Conjoint LINMAP* User's Manual.

Chrzan, Keith (1991) "A Survey Version of Full-Profile Conjoint Analysis," *The Proceedings of the Society for Consumer Psychology 1990 Annual Convention*, Curtis P. Haugtvedt, editor, (in press).

Cohen, Jacob and Patricia Cohen (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* 2nd ed. Hillsdale: Lawrence Erlbaum.

Currim, Imran S. (1981) "Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models," *Journal of Marketing Research*, 18 (August), 301-9.

Green, Paul E. and V. Srinivasan (1978) "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103-23.

Green, Paul E. and V. Srinivasan (1990) "Conjoint Analysis in Marketing Research: New Developments and Directions," *Journal of Marketing*, 54 (October), 3-19.

Green, Paul E., Yoram Wind and Arun K. Jain (1972) "Preference Measurement of Item Collections," *Journal of Marketing Research*, 9 (November), 371-7.

Hintze, Wayne J. (1990) "The Degree of Freedom Problem in Conjoint Analysis," *Proceedings of the Advanced Research Topics Forum*, forthcoming from the American Marketing Association.

Johnson, Richard M. (1987) "Adaptive Conjoint Analysis," *Sawtooth Software Conference Proceedings*, Ketchum: Sawtooth Software, 253-65.

Louviere, Jordan J. and Richard D. Johnson, (Forthcoming) "Reliability and Validity of the Brand-Anchored Conjoint Approach to Measuring Retailer Images," *Journal of Retailing*.

Mullet, Gary M. (1989) "On Conjoint Studies with Scarce Degrees of Freedom: Is there Enough Utility to Go Around?" *Marketing Research*, 1 (June), 24-30.

Mullet, Gary M. and Marvin J. Karson (1986) "Percentiles of LINMAP Conjoint Indices of Fit for Various Orthogonal Arrays: A Simulation Study," *Journal of Marketing Research*, 23 (August), 286-90.

Sawtooth Software (1988) *ACA System* User's Manual Version 2.0 Ketchum: Sawtooth Software.

Srinivasan, V. and Allan D. Shocker (1973a) "Linear Programming Techniques for Multidimensional Analysis of Preferences" *Psychometrika*, 38 (September) 337-69.

Srinivasan, V. and Allan D. Shocker (1973b) "Estimates of Weights for Multiple Attributes in a Composite Criterion Using Pairwise Judgments" *Psychometrika*, 38 (December) 473-93.

Wittink, Dick R. and Phillip Cattin (1989) "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53 (July), 91-6.

Exhibit I

Critical R^2 Values for OLS Conjoint Models

n	k	α=0.25	α=0.10	α=0.05
8	3	.6059	.7586	.8317
8	4	.7611	.8768	.9240
8	5	.8913	.9587	.9797
8	6	.9818	.9971	.9993
9	4	.6732	.8043	.8647
9	5	.8007	.8985	.9376
9	6	.9085	.9655	.9830
9	7	.9845	.9976	.9994
16	4	.3634	.4802	.5499
16	5	.4429	.5575	.6248
16	6	.5177	.6296	.6920
16	7	.5893	.6963	.7538
16	8	.6602	.7586	.8100
16	9	.7264	.8162	.8601
16	10	.7908	.8684	.9046
16	11	.8512	.9149	.9422
16	12	.9074	.9543	.9722
16	13	.9567	.9839	.9921
16	14	.9925	.9988	.9997
18	5	.3909	.4990	.5644
18	6	.4581	.5659	.6276
18	7	.5236	.6278	.6873
18	8	.5872	.6871	.7417
18	9	.6471	.7423	.7923
18	10	.7071	.7941	.8383
18	11	.7644	.8426	.8808
18	12	.8194	.8870	.9182
18	13	.8711	.9267	.9504
18	14	.9199	.9605	.9760
18	15	.9624	.9860	.9932
18	16	.9935	.9990	.9997

Exhibit I, continued

Critical R^2 Values for OLS Conjoint Models

n	k	α=0.25	α=0.10	α=0.05
25	5	.2776	.3645	.4190
25	6	.3258	.4152	.4700
25	7	.3738	.4637	.5190
25	8	.4203	.5110	.5643
25	9	.4670	.5563	.6085
25	10	.5105	.6000	.6500
25	11	.5543	.6421	.6900
25	12	.5984	.6825	.7290
25	13	.6409	.7240	.7660
25	14	.6817	.7598	.8002
25	15	.7235	.7959	.8338
25	16	.7642	.8299	.8649
25	17	.8031	.8637	.8942
25	18	.8409	.8953	.9213
25	19	.8772	.9244	.9457
25	20	.9123	.9505	.9667
25	21	.9451	.9732	.9838
25	22	.9742	.9905	.9953
25	23	.9955	.9993	.9998
27	5	.2553	.3375	.3895
27	6	.3017	.3854	.4382
27	7	.3451	.4315	.4844
27	8	.3886	.4755	.5273
27	9	.4309	.5180	.5696
27	10	.4737	.5592	.6088
27	11	.5136	.5994	.6480
27	12	.5541	.6373	.6744
27	13	.5951	.6753	.7207
27	14	.6333	.7121	.7549
27	15	.6716	.7474	.7876
27	16	.7100	.7811	.8186
27	17	.7478	.8142	.8487
27	18	.7837	.8459	.8774
27	19	.8193	.8759	.9035
27	20	.8544	.9045	.9281
27	21	.8876	.9309	.9503
27	22	.9196	.9547	.9696
27	23	.9496	.9754	.9851
27	24	.9763	.9913	.9957
27	25	.9959	.9994	.9998

Exhibit I, continued

Critical R^2 Values for OLS Conjoint Models

n	k	α=0.25	α=0.10	α=0.05
32	5	.2145	.2857	.3325
32	6	.2528	.3265	.3741
32	7	.2899	.3661	.4148
32	8	.3275	.4041	.4529
32	9	.3625	.4412	.4902
32	10	.3983	.4776	.5249
32	11	.4333	.5123	.5596
32	12	.4693	.5468	.5933
32	13	.5028	.5810	.6262
32	14	.5355	.6138	.6574
32	15	.5693	.6452	.6878
32	16	.6006	.6764	.7183
32	17	.6362	.7073	.7484
32	18	.6675	.7376	.7752
32	19	.6995	.7662	.8021
32	20	.7304	.7940	.8281
32	21	.7615	.8221	.8533
32	22	.7922	.8484	.8771
32	23	.8214	.8739	.9006
32	24	.8513	.8984	.9212
32	25	.8794	.9216	.9412
32	26	.9072	.9430	.9592
32	27	.9335	.9628	.9749
32	28	.9584	.9797	.9877
32	29	.9803	.9928	.9965
32	30	.9966	.9995	.9999

Exhibit II

Critical B Values for LINMAP Conjoint Models

n	Design	α=0.50	α=0.10	α=0.05
8	3×2^2	.1226	.0075	.0038
8	3×2^3	.0150	.0000	.0000
8	3×2^4	.0075	.0000	.0000
8	4×2^2	.015	.0056	.0000
8	4×2^3	.0000	.0000	.0000
9	3^2	.1723	.0146	.0102
9	3^3	.0146	.0000	.0000
9	3^4	.0000	.0000	.0000
16	2^4	.4721	.1680	.1280
16	2^5	.3836	.1435	.1023
16	2^{10}	.0856	.00149	.0076
16	2^{15}	.0000	.0000	.0000
16	4^2	.2842	.1038	.0712
16	4^3	.1225	.0310	.0168
16	4^4	.0162	.0000	.0000
16	4^5	.0000	.0000	.0000

Exhibit II continued

Critical B Values for LINMAP Conjoint Models

n	Design	α=0.50	α=0.10	α=0.05
18	2^5	.4446	.1898	.1499
18	2^6	.3499	.1552	1233
18	2^7	.2735	.1205	.0917
18	3^3	.3580	.1545	.1140
18	3^4	.2231	.0863	.0641
18	3^5	.1300	.0394	.0263
18	3^6	.0661	.0128	.0065
18	3^7	.0131	.0014	.0000
18	3×2^3	.5535	.2384	.1777
18	3×2^4	.4441	.1935	.1466
18	3×2^5	.3499	.1486	.1178
18	3×2^6	.2810	.1198	.0911
18	$3^2 \times 2$	.5525	.2299	.1754
18	$3^3 \times 2$	.3580	.1545	.1140
18	$3^4 \times 2$	.2227	.0861	.0641
18	$3^5 \times 2$	.1300	.0394	.0263
18	$3^6 \times 2$	.0661	.0128	.0065
18	$3^2 \times 2^2$	.5525	.2299	.1754
18	$3^3 \times 2^2$	.3580	.1545	.1140
18	$3^3 \times 2^3$	.3580	.1545	.1140
18	$3^4 \times 2^3$	.2227	.0861	.0641
25	5^2	.3593	.1690	.1332
25	5^3	.1930	.0874	.0692
25	5^4	.0854	.0312	.0205
25	5^5	.0173	.0025	.0001
25	5^6	.0000	.0000	.0000

Exhibit III

Critical r Values for Studies with Holdout Samples of Size n

n	α=0.10	α=0.05	α=0.01
3	.9511	.9877	.9995
4	.8001	.9000	.9800
5	.6871	.8053	.9158
6	.6083	.7293	.8822
7	.5509	.6694	.8329
8	.5068	.6215	.7888
9	.4716	.5823	.7498

Exhibit IV

Crosstabulation of Reliability Derived from R^2 and from Holdout Profiles

	From R^2	
	Reliable	Unreliable
From Holdouts:		
Reliable	13	35
Unreliable	25	139

$\chi^2 = 3.54$, ns

r = .1292

Exhibit V

MANOVA of mean part-worth by respondent reliability

Part-worth #	Reliable	Unreliable	p
1	229	157	<0.01
2	44	-2	<0.01
3	-273	-155	<0.01
4	-15	-11	ns
5	-2	9	ns
6	17	2	ns
7	-13	8	ns
8	-31	-12	ns
9	44	3	<0.01
10	-136	-65	<0.01
11	35	2	<0.01
12	102	63	<0.01
13	-128	-128	ns
14	128	128	ns
15	-96	-64	<0.01
16	96	64	<0.01
n	472	257	

Exhibit VI

MANOVA of mean part-worth by respondent reliability

Part-worth #	Reliable	Unreliable	p
1	115	89	<0.01
2	16	5	0.04
3	-131	-94	<0.01
4	-9	-10	ns
5	12	7	ns
6	-3	3	ns
7	-4	-1	ns
8	-11	-12	ns
9	14	12	ns
10	-68	-33	<0.01
11	14	0	<0.01
12	54	33	<0.01
13	-66	-65	ns
14	66	65	ns
15	-46	-32	<0.01
16	46	32	<0.01
n	472	257	

Exhibit VII

MANOVA of mean part-worth by respondent reliability*

Part-worth #	Reliable	Unreliable	p
8	-19	-77	0.04
12	17	9	<0.01
15	-12	-6	<0.01
16	24	15	<0.01
17	8	4	0.02
18	17	-6	<0.01
20	-19	-10	<0.01
21	14	7	<0.01
23	-11	-5	0.02
24	-16	-9	<0.01
25	15	8	<0.01
27	-11	-6	0.03
29	-15	-9	<0.01
41	16	9	<0.01
43	-15	-8	<0.01
46	-10	-4	<0.01
47	12	5	<0.01
49	-7	-2	0.04
50	-8	-3	0.04
n	472	257	

*Not shown for reasons of space are 31 non-significantly different pairs of means

Exhibit VIII

Crosstabulation of Reliability by Segment Membership

	Reliability	
	Reliable	Unreliable
Segment:		
A	29%	16%
B	24	30
C	39	54
D	7	1
Total	100%	100%
n	472	257

$\chi^2 = 34.88$, $p<0.01$

Exhibit IX

Crosstabulation of Reliability Derived from LINMAP and from OLS

	LINMAP	
	Reliable	Unreliable
From OLS		
Reliable	379	93
Unreliable	15	242

$\chi^2 = 368.46$, $p<0.01$

$r = .710$

Comment on Chrzan

William G. McLauchlan
McLauchlan & Associates, Inc.

Keith Chrzan is to be commended for his undertaking. As the title of his paper indicates, he has provided us with a comprehensive understanding of how unreliable respondents can be identified in a variety of conjoint or tradeoff techniques, and the impact that those respondents can have on partworth estimates.

My specific comments fall into two areas. First, the author has correctly stated that the identification of unreliable respondents in Sawtooth Software's Adaptive Conjoint Analysis (ACA) may not be as statistically straightforward as in other partworth estimation techniques (OLS regression and linear programming). In ACA, a Pearson r is computed between actual and predicted responses to a set of "calibration concepts." The nature of ACA is such that the responses to the calibration questions are necessarily constrained to be monotonically increasing. As such, "reliability" assessed by this measure will be artificially high.

It is my recommendation that the ACA calibration correlation not be used to assess respondent reliability or unreliability, even with the extremely conservative alpha levels proposed by the author of the paper. Rather, the use of repeated holdout tasks (pre- and post-ACA) represents a better measure of respondent reliability. The use of holdout tasks is recommended not just for ACA, but for other partworth estimating algorithms as well.

My second comment is really a proposal for discussion, the topic of which is not specific to the paper, itself, but is raised in response to the issue which motivated the paper in the first place. The fundamental problem that I believe all of us need to consider is the appropriateness of "throwing out" unreliable respondents.

I believe that one can reasonably argue that just as there are unreliable respondents in conjoint data sets, there are unreliable consumers in the marketplace. As such, respondents should not be discarded perfunctorily. It is probably true that this premise makes the most sense when exploring preferences in categories which tend to be heavily image driven and advertised (such as automobiles, tobacco, liquor, and perfume).

Further, as Joe Curry from Sawtooth Software has suggested, the premise may make the most sense when conducting share of preference simulations on existing products, as opposed to conducting new product optimization analyses. In any event, the inclusion of unreliable respondents may add some necessary "noise" to the data set which, in turn, may lead to better predictive validity. The premise is clearly one which lends itself to researchable hypotheses.

MODELING PREFERENCE IN CONJOINT MEASUREMENT

Paul F. Hase
Hase/Schannen Research Associates

OVERVIEW

One of the most important roles marketing research can play is to provide management with tools to evaluate their many alternative courses of action; to go beyond traditional analyses of data bases. One class of tools that meet this management need is models that relate marketing activities to changes in consumers' perceptions and values and then to unit demand and financial payoffs. The ability to answer a wide variety of "what-if" questions with these models is one of the more important developments in market research over the past several years.

In recent years both behavioral and preference models have become more widely available and popular. The behavioral models (for packaged goods) have often utilized purchase data from consumer panels and aggregate scanner compilations obtained under varying market conditions, while the preference models have been developed using several types survey information. In these surveys, the information gathered has focused on consumer needs and perceptions of current (and potential future) competitors' products and services. Because they use survey data, models based on this type of information can and have been developed for all types of markets — consumer, commercial, and industrial products and services. Behavioral models, on the other hand, have been restricted to those situations for which appropriate data are available.

Historically, the data collection techniques that have been used to measure "needs" are many. However, we are interested in that class of measurements that provides end users' (consumers') "utility" values for each of several options (or "levels") for each product or service attribute under investigation. These utility values, which contain detailed information about consumers' needs, can be obtained in a variety of ways: "derived" via conjoint measurement; directly measured through "self-explicative" approaches; or obtained by a combination of the two ("hybrid" methods). Most popular are the data collection techniques that provide utilities at the individual respondent level, but also in use are models that use data collected at the "segment level."

The purpose of this paper is to describe the current state of one such utility-data based preference modeling effort. Much of what we will discuss is of a practical nature, having evolved over 15 years of development in which we have encountered a wide range of management needs. As we have encountered these needs, we have operationalized as many as realistically possible into our models. While in some cases the management issues were so explicit that a custom model was required, in most we were able to incorporate the feature in the more general model. It is these more general features on which we will be concentrating.

MANAGEMENT ACTIONS IMPACT CONSUMERS, WHICH IN TURN AFFECTS THEIR PURCHASING PATTERNS

MANAGEMENT ACTIONS	CONSUMER IMPACT
Actual change in product or service attribute	Change in brand perception/image
Change in emphasis in marketing (repositioning)	Change in brand perception/image
Change in marketing level of activity	Change in awareness/degree of familiarity
Change in distribution/channels	Change in access/awareness
Line simplification	Fewer items to choose from
Line expansion/new product introduction	More items to choose from
Change in availability of options for a product or service	More/less variety to choose from
Pricing changes	Change in perceived value of brand

For the purposes of this paper, we assume utility data or their equivalent are available, whether derived, self-explicated or from a hybrid data collection procedure. We will not be discussing the type of stimuli used to obtain the utilities (full-profiles, linked partial profiles, paired comparisons, etc.); how the price and brand "attributes" are handled (Allison, N. and Hase, P. (1989) "Decision Processes and Choice Behavior: Conjoint Issues in Strategy Development," AMA Conference in Technology in Marketing, June 1989); the advantages and disadvantages of alternative methods of interviewing; how one goes about defining attributes and how many to include; the inclusion of interactions in the measurement design; and so on. While none of these issues is a topic here, they are mentioned because the choices made in each of these areas can have a profound effect on the utilities obtained and thus on the results obtained from the analytical models that are implemented.

OBJECTIVES OF PREFERENCE MODELING

In general, preference models are used for two types of analyses, with the second, which includes marketing mix and other environmental factors, being an extension of the first:

1. Feature Evaluations

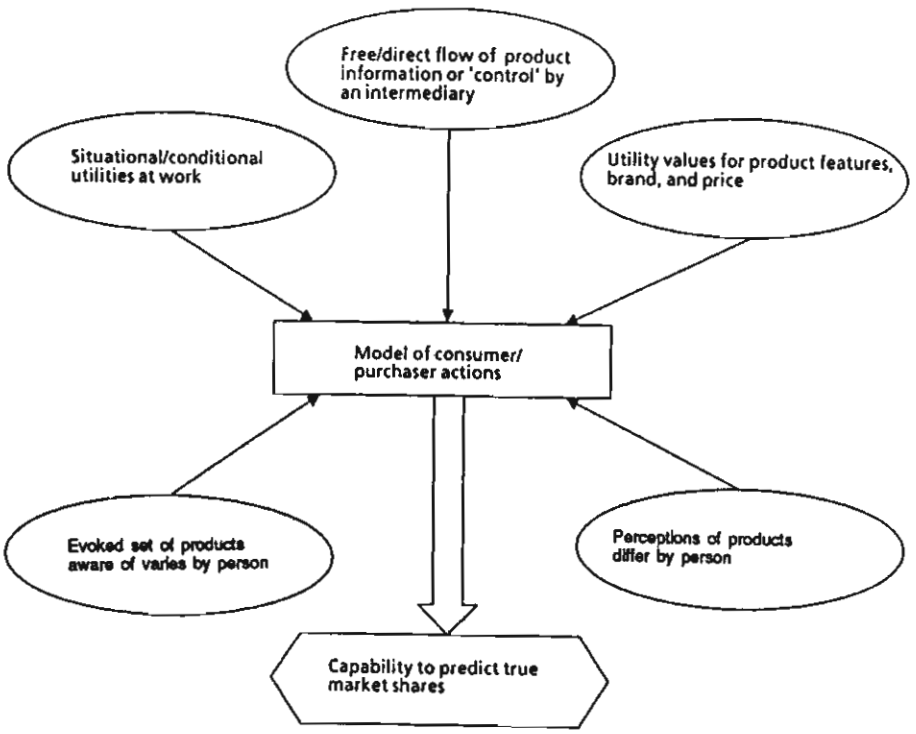
   In this type of analysis, one is interested in determining which of several alternative feature packages (aka "bundles") yield the greatest "preference share" or purchase likelihood. Because it is obviously the case that people would like to have the best of everything, these analyses usually concentrate on feature (including price) tradeoff evaluations in which one or more "better" options are added as others are taken away, to estimate the net effect on the criterion measure of interest. From the set of alternatives examined, those felt to have the

most potential based on preference shares, costs, and pricing are recommended for further development. These preference models use only the survey data, and the results are not interpreted as anything more than shares of preference or purchase likelihoods.

2. Predictive Models

The objective of this class of models is to predict actual market shares. Just as for the Feature Evaluation analyses, these models also have as their most important purposes the evaluation of changes to current products/services, alternative pricing strategies, and the introduction of new items or the dropping of old ones. However, what is different is that the criterion measures — whether they be share or unit/dollar demand — are actual marketplace values, obtained by converting (by various means) preference into reality. To accomplish this goal several ingredients (as fit the particular case) are added to the model's specification. These ingredients include such factors as changes in marketing activities (for example, advertising, promotion), changes in channels of distribution and the roles of intermediaries (for example, sales people, distribution).

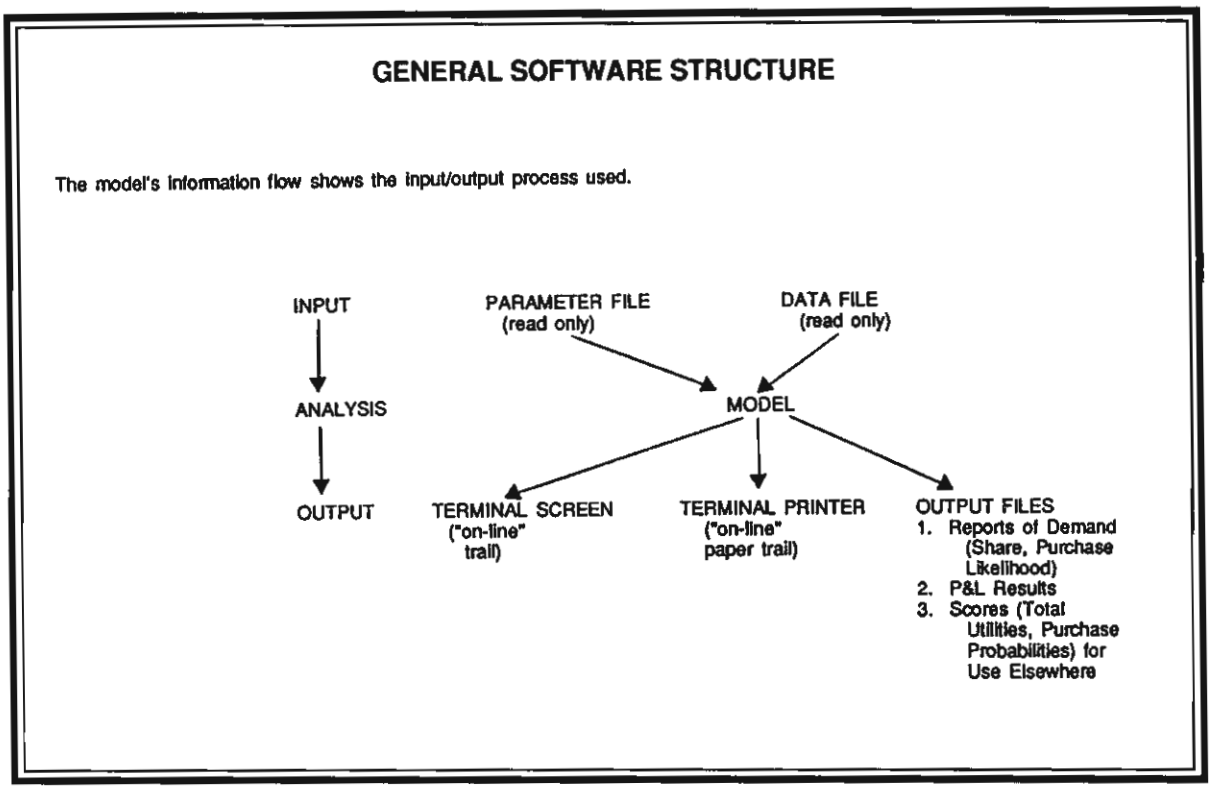
FORCES REQUIRED TO MODEL TRUE MARKET SHARES



The focus of much of this paper will be on the predictive class of models, because, in our experience, these have been of greatest interest to managers.

GENERAL MODEL STRUCTURE

Before discussing how we have addressed selected modeling topics we have confronted, it is useful to present the general flow of the software structure we use. By so doing, it will help in explaining how certain objectives are met.



In our setup, there are three sources of information to the model: a parameter file, a data file and instructions (commands) from the terminal. The parameter file describes the nature of the data file, defines the attributes and the segments to be analyzed, the types of models one wants to use, product line structures, the existence of submarkets, and so on.

The data file includes consumers' "values," segment designations, and, if included in the study design, perceptions of competitors' products on the attributes studied.

Instructions from the terminal define things such as the specific analyses to be conducted, what models to run and what information the user wants to review. The user responds to queries, each of which lists the required values. Several modes of run-time checking are provided, as are protections against invalid inputs.

The output is of several types, fully controlled by the user. Requested reports (the results of one or more models — as discussed below — the user wants to run) go to a disk file, broken out by all the segments one wants to examine. Data output (such as predicted shares, purchase probabilities, dollar volumes, and profits) can also be sent to a data file for later use in a spreadsheet program or other application. And lastly, a "paper trail" of all work done at the terminal can be generated, so the user has a backup document with which to check his or her inputs.

MODELING ISSUES AND ALTERNATIVE APPROACHES

Having described the basic preference modeling objectives and structure, several issues that confront the modeler are presented and discussed below. While those discussed are certainly not all inclusive, they reflect many of the practical, client needs — related ones we have confronted over the years. While some are not complex and can be incorporated into models quite easily, others are more complex, and still others have no "right" answer — for which we must be prepared to offer alternative solutions.

Even though each issue is presented separately, we do recognize from the outset that many are interrelated and, as appropriate, these interrelationships will be pointed out.

1. Examination of Results by Market Segments

   Analysis of market segments — whether they be demographic, behavioral, "value segments" defined by cluster analyses, or others — to see if some react differently than others is always of strategic interest. In this process there is often an initial zeal for "maximum understanding," as reflected by the desire to look at an extremely large number of subgroups. For our model, this desire requires that we provide the capability to provide model estimates for all segments of possible interest, no matter how many (the maximum we have had requested is over 100!). Then, as analysis progresses and we begin to home in on the segments that matter, we need a procedure to retain just those.

   Further, as this progression unfolds, whether we are analyzing many or just a few segments, we want each simulation run of the model to provide the estimates for all of them, eliminating the need for multiple passes.

   To provide maximum flexibility, we have configured our model such that every segment that has been defined by one or more variables in the database has the potential to be analyzed. We then provide two avenues for the users to filter the complete set (if they wish), one in the parameter file and then again in "real-time" mode as the model is being run. In this way the users need not redefine segments or make multiple runs to examine the subsets of interest.

   We have provided the same capability for report generation. The users can define which models they want to run (see below), so that in any one simulation estimates from each are provided for all those selected. Thus, all models' results can be obtained for all segments of interest in one simulation run.

2. Types of Models Used to Estimate Preference Share

There are several commonly used models for estimating preference share (prior to modification into calibrated, "real" shares), all of which are discussed widely in the literature. Our objective in this paper is not to discuss each model's merits or demerits, but to present a two-stage, hierarchical model as an extension of the popular logit, logit adjusted for similarities and BTL models. Its major application is for those cases when we are analyzing product lines with more than one entry, in which cases total line demand and cross-elasticities of demand (cannibalization) are of critical interest. In addition to eliminating the problem cited, we believe that this model provides a reasonable representation of how purchase decisions are often made. (It also helps to resolve much of the so-called "red bus/green bus" problem characteristic of the logit and BTL models.)

All the most common preference models start with each respondent's total utility scores for each competitor's bundle of attributes. Let us designate them as:

$$\mathbf{T}(i),\ i=1,2,\ldots,p$$

The First Choice model assumes that a respondent has a 100% likelihood of choice for the product with the highest score, no matter how close the "win." The number of wins across respondents provides the preference share estimate for each competitor.

The logit, logit adjusted for similarities, and BTL models, on the other hand, do take into consideration the differences in the T(i), creating for each respondent a probability of purchase for each competitor. To obtain a significant win, a product's total utility must exceed all the others by a goodly margin. (It is important to note that these models — and thus the probabilities they generate — are strongly affected by the scaling of the utility values used to generate the total scores. This issue will be touched upon in the Model Calibration section that follows below.) The average of each competitor's probabilities across respondents provides the preference share estimates.

Logit Model Probabilities:

$$\frac{e^{T(i)}}{\text{SUM}[\ e^{T(i)}\ ]}$$

BTL Model Probabilities:

$$\frac{T(i)}{\text{SUM}[T(i)]}$$

A problem with these latter two models is that as similar and even identical entries are added to the market for a given competitor, the competitor's total share goes up.

There is some logic to this result — as a product line is lengthened, adding variety and market exposure — we would expect overall share to rise. However, these models, as presently defined, tend to overstate this effect.

Other models, known as adjusted logit models (Johnson, R.M., "Adaptive Conjoint Analysis," Sawtooth Software, Inc., March 26, 1987, pp 22, 23), attempt to overcome this problem by modifying the logit probabilities using the similarities of the profiles. The effect of this adjustment is that the addition of profiles similar or identical to those already in existence take more from those they are like; this serves to reduce the overstatement of the total share obtained by similar entries. (Occasionally Adjusted Logit models can lead to strange outcomes, in which an improvement to a profile leads to a loss in share. This result comes about when a relatively unique, (dissimilar) profile gains less in total utility score by an improvement than it loses by becoming more "me-too" (that is more like several other profiles).

Across brands we generally find it desirable to retain the similarities adjustments, so that similar products compete more among themselves for customers; it is among entries within a brand's product line that we want to eliminate the "red bus/green bus" effect, so that just by adding more and more items we do not overstate the brand's total share. The two-stage model we use is meant to accomplish this goal.

Conceptually it is very simple. It assumes that the consumer shops within each competitor's set of offerings (product line) and then compares each competitor's best offering when making a final selection. An analogy would be a car shopper first determining his or her preferred model within each manufacturer's showroom and then making the final purchase decision among manufacturers from just these preferred models. Other analogies would include products such as personal computers, package goods, services in the financial area, and car rentals — any situation which involves multiple entries being offered by one or more competitor.

Because respondents' utilities vary, their preferred models within and across competitors will differ. As a competitor adds more and more entries to its line, its share goes up because of the added "variety," but in a way that seems to be more consistent with *a priori* expectations.

Two-Stage Preference Model

Base Case Simulation

Product	Product Line	Model: First Choice	Logit	Adjusted Logit	Two-Stage: 1st Choice/ Adj. Logit
Brand A1	1	17.5%	17.6%	17.3%	17.3%
Brand A2	1	NA	NA	NA	NA
Brand A Total	1	17.5	17.6	17.3	17.3
Brand B	2	4.0	8.4	8.2	8.2
Brand C	3	78.5	74.0	74.5	74.5

Brand A Adds Entry to Line

Product	Product Line	Model: First Choice	Logit	Adjusted Logit	Two-Stage: 1st Choice/ Adj. Logit
Brand A1	1	14.2%	15.1%	15.7%	12.0%
Brand A2	1	6.6	12.8	12.5	11.3
Brand A Total	1	20.8	27.9	28.2	23.3
Brand B	2	4.0	7.2	7.6	8.1
Brand C	3	75.2	64.9	64.2	68.6

As the model results indicate, we are able analyze both total line demand and cannibalization for our brand and all other brands in the market. We believe that this structure makes sense, in that utilization of the first choice rule in stage-one correctly captures the fact that within a brand's line even a small win in total utility — possibly for a seemingly inconsequential difference — would lead a consumer to state that "if I buy that brand I will choose model X." And utilization of a probabilistic model in stage-two among the winners also makes sense, in that it considers the greater uncertainty that most likely exists in these final decisions.

The procedure for implementing the two-stage model is quite straightforward. In the parameter file we define not only each profile's features, but also the product line to which it belongs. When the model is executed, and whatever profile modifications that are to be made for that particular simulation are defined, we simply scan (for each respondent) each product line and choose the best one using the first choice rule. Having made these determinations, we then use a logit model and/or a BTL model (all are available to the user) to estimate shares among the competitors using these "winners." Shares for each individual profile are reported, as well as product line summaries.

3. Concurrent Modeling of Sub-Markets

In most preference modeling work, the user changes features and/or the price of one or more profiles to determine the effect on each competitor's share. When doing so the changes are implemented for all respondents, thus affecting everyone's predicted choices. Realistically, however, the strategies that decision-makers often wish to evaluate do not involve ubiquitous changes, but rather differential or targeted changes that affect only one or more segments. Some examples would be:

- Offering special discounts to high volume customers while at the same time imposing surcharges on those with low volumes.

- Providing special services to better customers of a financial institution/credit card while at the same time charging higher fees/ interest to less attractive customers.

- Offering alternative versions of products in different regions and/or through different channels of distribution.

In the above examples, consumers in one segment are being offered (are choosing among) different sets of products/services, which would reflected in the model as follows:

	High Volume Customers		Low Volume Customers	
	Brand A	Brand B	Brand A	Brand B
Special Discounts	Yes, Level 1a	Yes, Level 1c	No, Level 1e	No, Level 1e
Extra Services	Yes Level 2b	Yes, Level 2a	No, Level 2e	No, Level 2e
Incentives to Increase Volume	No, Level 3d	No, Level 3d	Yes, Level 3b	Yes, Level 3c
Price	Level 4b	Level 4a	Level 4d	Level 4c

In the above market structure, high volume customers would be reacting to a different set of offerings than would low volume customers, with the competitors differing from each other in different ways (on different attributes) within each of these segments. While this example is an extremely simple one, the actual types of sub-market strategy analyses with which we deal are usually much more complex.

We have constructed our model's parameterization and command language in so that these analyses can be easily performed. In the parameter file we can indicate how many sub-markets (if any) we wish to create, and which segment variables define them (see the discussion above concerning segment analysis). The only requirement is that they be mutually exclusive and totally exhaustive. In the model command language, we can direct any changes in features, pricing, awareness, and so on, to any sub-market we wish, so we can target these modifications to just those to whom they relate. (We could also use the command language to create the differential competitive offerings in real-time mode, but this would be quite cumbersome.)

With these two capabilities we have the tools to evaluate some rather complex business strategies in a single simulation run, both by the sub-markets themselves and for the market in total (because the sub-markets are combined in their proper proportions as respondents are aggregated in the model). Further, any other subgroup (other than the sub-markets themselves) can be evaluated — again because they will have the correct mix of respondents from each sub-market as the data are aggregated.

Of course, one could conduct sub-market and total market analyses without the model capabilities outlined above. In order to do so one could run separate simulations for each sub-market and then accumulate each sub-market's results in some way after the fact to obtain the total market. However, this would be a very inefficient approach, and would also suffer from not allowing the user to evaluate segments other than the sub-markets.

4. Defining Profiles

Even though the topic of how the competitive set is determined does not receive much attention in the literature, it is as important to preference modeling as the derivation of consumers' value systems (utilities). How the competitive set is determined, what levels of the attributes are used to describe each profile, and from what source they are obtained, requires careful thought and planning. The importance of careful profile definition cannot be overstated, in that once the set is put into place in the "base case" simulation it forms the foundation for all that follows.

The way profiles are defined for the model usually reflects some critical assumptions (often not recognized) about how a marketplace works (that is, how consumers make choices and with what information). The discussion that follows attempts to outline various approaches and their implications. However, in most cases there is no one "correct" way; rather, what is important is to be aware of the implications of the procedures and to choose one.

Method 1: Product/Service "Facts"

In this method — which is one of the more prevalent — each competitor is described by what its "actual" levels of performance are, whether it be the acceleration of a car, the checking account fees charged by a bank, the published life of a tire, the copies per minute for a copier, or the number of ounces of beef in take-out hamburger. In some cases — such as copies per minute — "true" facts are known. In other cases, only "average" facts may known because they are dependent on other factors (for example, the type of car and the type of driver on the life of a tire). For those attributes in which indisputable facts are not available, management judgment/consensus is often required.

This procedure is used whenever no other information about a product/service (such as consumers' perceptions) is available. If used, however, one must recognize two critical assumptions that are being made:

- All consumers have exactly the same perception of all the competitors.

- All consumers have perfect knowledge of all the competitors.

While neither of these assumptions is really true in any market, this fact does not negate the value of preference share modeling under these circumstances. In these analyses we can learn a great deal about what features have the greatest leverage, what their "value" is to people, and what bundles (profiles) have the greatest potential.

What is more difficult to accomplish when profiles are defined in this way is to calibrate and model actual market shares. This follows from the fact that to get to real shares, we must in some way modify the results obtained from this model, since it makes the two highly unlikely assumptions cited above.

Method 2: Management Judgment of Consumer Perceptions

This procedure is very similar in nature to Method 1, but rather than using "facts" management uses beliefs about consumers' perceptions of the various competitors. The source may be prior research (for example, as to perceived accelerations, checking account fees, and so on), which we would use to define the profiles even though consumers' beliefs might, in some cases, be quite far from the "truth."

By using profiles defined in this way we are able to eliminate one assumption of Method 1; the one concerning perfect knowledge. However, we are still assuming that all consumers have the same perception of all the competitors.

A sub-issue in this procedure is: whose perceptions about a competitor do we use? Do we use those given by users of each brand, or do we use a weighted average of users and aware/non-users (whose ratings are normally less favorable than those of users)?

There is no easy or obvious answer to this question, but one solution is to implement the sub-market model described above. To do this we would develop two profiles for each competitor, one based on the perceptions of users and one based on the perceptions of aware/non-users. We would then create sub-markets based on brand usage (or usage of combinations of brands), with the sub-markets being assigned brand user and non-user attribute profiles as appropriate.

Method 3: Measure Consumer Perceptions Directly

In this approach consumers indicate which level of each attribute best describes each competitor. In the questionnaire, individual respondents evaluate just those competitors (usually no more than four, depending on the number of attributes), with which they are most familiar or use most.

By obtaining these perceptions directly we are not required to assign the same "average profiles" to all respondents as is the case in the two methods above. We are able to use each respondent's unique set of perceptions. By so doing we need not make any assumptions about perfect knowledge or that all consumers have the same perceptions of all products.

However, this method is not without shortcomings. While we obtain individuals' perceptions, we can only obtain them for the competitors about which they have knowledge. For example, if a market has 10 brands, and each consumer rates up to four (fewer if he or she is not aware of this many), how do we model the unrated six? Do we make it impossible for the consumer to ever become aware of and choose any of them; or do we assign them some set of levels based on other (non-user) consumers' perceptions; or do we not model using individuals' ratings, but rather revert to Method 2, using the perceptions data collected as the source for the "average perceived profiles?"

As pointed out previously, there is no "correct" answer to this problem, although we do reject any modeling approach that does not make it possible for a brand to be chosen by a non-aware/non-rating consumer. To do so would severely detract from the model's analytical value.

While we would prefer to use individuals' perceptions, problems with too few of the brands in a market being rated by each individual have led us in most cases to use the data as input to a Method 2 type of profile definition. However, when we can — usually in markets with very few well-known competitors — we do use individual respondents' ratings.

A mixed approach, in which individual ratings are used for brands used or aware of and averages of other non-users' ratings for those not rated, we believe is inappropriate. The ratings obtained from an individual express that person's idiosyncratic view of these brands' relationships to each other; to "compare" (in the model) these brands to other brands based on ratings given by others — ratings which may bear no relationship to the brands rated by the individual — can very likely lead to erroneous conclusions.

5. Calibrating Preference Models to Actual Market Shares

To model actual market shares, procedures are needed to translate the preference share estimates generated by the basic model. Of course, one must have known market shares on which to calibrate, which, in some markets (particularly industrial, commercial, and professional markets) may not be readily available.

Three procedures we have used to calibrate preference models are listed below, in reverse order of preference and general usefulness:

1. Aggregate, post-simulation adjustments

2. Utility rescaling

3. Direct use of marketing factors (such as awareness and availability)

Method 1: Aggregate, Post-Simulation Adjustments

The first method is the most straightforward and easily implemented. A base case simulation is run, and the preference shares obtained are ratioed to each profile's actual share. The shares obtained in ensuing simulations are multiplied by the respective factors and the adjusted results re-percentaged. These re-percentaged shares are then used as estimates of actual shares.

Determination and Use of Aggregate Calibration Factors

	Base Case Simulation		
	Uncalibrated Share	Actual Share	Calibration Factor
Brand A	15.7%	20%	1.27
Brand B	7.6	10	1.32
Brand C	12.5	20	1.60
Brand D	64.2	50	0.78

	Feature Change Simulation				
	Uncalibrated Share	x	Calibration Factor	= Revised Totals	Calibrated Share
Brand A	19.1%		1.27	24.3	22.1%
Brand B	11.1		1.32	14.7	13.3
Brand C	20.3		1.60	32.5	29.5
Brand D	49.5		0.78	38.6	35.1
				110.0	

The rationale for this procedure (in addition to its simplicity), is that it is supposed to adjust for all those factors not captured by a survey, for example, levels of marketing activity (such as those reflected in awareness), and breadth and depth of availability. However, as will be discussed below, there are better procedures for capturing these effects — procedures which can be more easily generalized to more types of market analyses.

We say there are better ways because of several practical problems with this post-adjustment procedure. The first is how to perform simulations involving the addition of new entries. There is no satisfactory procedure to determine what a new entry's calibration factor should be, and use of some arbitrary value (such as 1.00 or, if a line extension, the same value of the brand's other entries) will lead to erroneous results.

Second, by use of these factors, identical changes in a feature to two different profiles will lead to quite different effects on their estimated preference shares — depending on their calibration factors. That is, if a feature is improved for a profile with a factor greater than 1.00 it will (almost always) gain more (show more leverage for the feature) than will a profile with a factor less than 1.00. It is not at all clear that this is what we would expect to happen in the market, that a profile will show greater gains (or losses) just because the preference model understated its actual share in the base case.

And lastly, as the number of profiles in the model increases, calibration factors obtained this way become quite unstable, again putting into question the validity of the results obtained. In fact, this procedure often cannot be used with the First Choice model, in that profiles with no "wins" (a situation that becomes more likely as the number of profiles increases) cannot be calibrated (their adjustment factor is indeterminant).

Method 2: Utility Re-Scaling

In this method we don't adjust the base case preference share estimates (that are obtained using the original utilities in the calculations), but rather re-scale the utilities themselves. This single (not individual respondent) utility re-scaling factor is estimated so that the total sample's preference shares and the actual shares are a "best fit."

This utility adjustment factor is incorporated as follows into the logit model formulation:

$$P(i) = \frac{e^{\theta Ti}}{SUM[e^{\theta Ti}]}$$

What has been added to this formulation is the parameter "theta," which is determined through maximum likelihood estimation. In this estimation the dependent variable is the actual shares for the competitors, and the independent variable is the total utility for each competitor based on the original utilities. The "theta" derived is a utility multiplier, which, while preserving the relative differences of the utilities among the attributes, does affect the logit model's estimated shares as illustrated in the table below:

Logit Model Shares for Different Values of Theta

Product	Total Utility[1]	Theta 0.5	1.0[1]	2.0	5.0
Brand A	-0.6	11.7%	4.6%	0.5%	0%
Brand B	0.3	18.4	11.3	3.1	0
Brand C	1.2	28.9	27.9	19.1	2.9
Brand D	1.9	41.0	56.2	77.3	97.1

(1) Based on original utilities

As the table demonstrates, as theta increases the logit model approaches the first choice model. (It should be noted that theta will have no effect on the BTL model; however, an additive factor would.)

The advantages of this calibration method compared to the first are that a) the addition and/or deletion of profiles can be accommodated and that b) since we are modifying the utilities themselves the leverage of a feature change is not a function of the calibration factor.

The major disadvantage of the procedure is that we are still not incorporating marketing variables into the model; rather, we are using an algorithm to "force" the preference shares to look like real shares, without the explicit inclusion of these marketing issues. (Just as in Method 1, we assume that the factors we use to "force" the shares are "accounting for" these marketing variables, but this is a less satisfactory treatment than including them explicitly.)

Method 3: Directly Using Marketing Variables

The third method of calibrating is to directly incorporate marketing factors. The attractiveness of this approach is that it generalizes preference modeling to market modeling, allowing us to evaluate these marketing issues as well as feature and pricing changes.

A requirement for this method is that we know not only the actual market shares but also each competitor's awareness/availability — so we can employ this variable in our calibration.

Assuming these awareness/availability values are known, they are incorporated as part of each profile's description in the model's parameter file, and are taken into account in the estimation of shares.

We include these variables by using Monte Carlo techniques, in which each respondent is evaluated multiple times — each time based on a different set of profiles he or she can choose among (are aware of). Based on each profile's awareness level, the model determines whether a respondent is "aware" or "not aware" of it; this determination is repeated for each profile. The model then calculates the respondent's share just among "aware set." This procedure is repeated several (user specified) times. In any one given "pass," a respondent can be aware of just a few or of many profiles — depending on the Monte Carlo determinations.

Just by incorporating these factors we are not, of course, assured of a good fit between the model and actual shares. While we come much closer than when awareness, and other levels are not incorporated, we normally have to either use slightly different awareness levels than those we used initially and/or simultaneously use the procedure outlined as Method 2 above.

We have also on occasion customized the model to include other market variables in the calibration process. For example, when we have known that brands have different usage rates, we have applied them to respondents who "choose" each brand; this differential usage rate thus affects the estimated shares. However, we have not yet accommodated this or other similar variables into our general model's structure and command language.

By using Method 3, along with Method 2 if required, the end result is a model that we can now use to evaluate many more types of management scenarios using realistic assumptions, directly incorporating the most information about the actual marketplace.

6. Modeling Changes to Consumer Values

Most preference modeling focuses on changes in features and prices and their effects on share/demand. However, in many markets managers need to also understand the impact should consumers' values change as well. For example, in a market involving a "new technology," the value of an attribute might not be fully appreciated at the outset — and the

manager might want to know what would happen over time if its value (utility) increases as consumers become more familiar with it. Or, in a market dominated by one or competitors — and thus not particularly price sensitive — the manager might want to evaluate what would happen if it became "commoditized" (which would be the case, say, when a brand lost its patent and "generics" were allowed to enter).

To allow users to conduct evaluations of this type, we have built into our command language the capability to change not just product profiles but respondent's utilities as well. Managers can thus evaluate not only the effects of changes in competitors' performance but also in the underlying values that drive the market itself.

FUTURE DEVELOPMENT

The purpose of this paper has been to outline some of the practical preference modeling issues that we have encountered, and, whenever possible, to provide methods to address those issues. While we are now in a position to provide more sophisticated analyses than we were several years ago, there is much more that we can do in the future. Some of the possibilities would be to add the following capabilities:

1. Add market growth as a possible outcome of product/service feature changes (in addition to share changes)

2. Include differential usage rates when calculating shares

3. Add functional relationships between marketing activities (for example advertising, direct sales, and breadth and depth of availability) and market factors such as awareness so that the user inputs the marketing plan directly.

4. Develop optimization procedures to search for the best plan. Given the infinite number of business alternatives available to the user, this goal will be difficult to achieve in the general case; but, if formulated to search within a specified set of constraints, the possibilities for success will be greatly enhanced.

5. Work more vigorously to develop a data bank of validation experiences (successful or otherwise), so that the ultimate goal of proven predictive power can be assessed and established.

As we go forward, efforts to accomplish these and other objectives will allow us to provide even more valuable analyses and direction for management, enhancing the value of modeling research's input.

Comment on Hase

Catherine M. Schaffer
University of Denver

The paper by Paul Hase is commendable in that it addresses some real world issues in the modeling of preference in conjoint measurement. Mr. Hase addresses some of the trade-offs in tradeoff modeling.

The paper addresses six issues including:

(1) examination of results by market segments
(2) types of methods used to estimate preference share
(3) concurrent modeling of submarkets
(4) defining profiles
(5) calibrating preference models to market shares
(6) modeling changes to consumer values

Given that the concept of segmentation is so important in the field of marketing, the examination of results at the segment level is essential. The model developed by Hase and associates provides the user with a great deal of flexibility in that results can be provided for all possible segments.

With respect to the types of models used to estimate preference share, Hase recommends the use of a two-stage hierarchical model. It is possible that this type of model is appropriate under a certain set of assumptions. For example, one necessary consumer behavior assumption may involve how consumers classify stimuli using categories of meaning stored in memory. For example, do consumers evaluate alternatives first by product form, or by company offering brand?

The model developed by Hase and his colleagues also enables the user to evaluate changes that affect only one or more segments. This dramatically increases the flexibility of conjoint models.

With respect to the definition of profiles, Hase discusses the use of consumer perceptions to achieve this. However, this issue raises a set of sub-issues, such as what is to be done with missing data. In other words, if a respondent has no basis on which to evaluate a specific brand, what value do you use for this particular person? The issues involved in using actual perceptions as input to conjoint models are quite thoroughly addressed in the paper.

The calibration of preference models to actual market shares is often difficult. Three procedures for doing this are described in the paper. A problem sometimes encountered is a case where the simulator gives results that are not even in the same rank order as actual market shares. Under such a condition, it is unclear that the use of a calibration factor would be appropriate. With respect to method 2 — utility rescaling — one suggestion might be to rescale at the segment level.

The final issue, that of modeling changes to consumer values, is probably the most controversial. Users have the capability to change respondents' utilities. Hase's example of the value of this feature for a new technology is a good one, but the possibilities for misuse of this feature are many. It is my opinion that the ability to change utilities is dangerous and could easily be misused.

With respect to the future developments highlighted in the paper, many of these ideas have already been implemented. For example, Green, Carroll, and Goldberg's (*Journal of Marketing*, 1981) POSSE system dealt with the issue of optimization. In addition, much work (although most of it of a proprietary nature) has been done in assessing the validity of conjoint models.

One final suggestion would be the inclusion of the "fuzzy product" in these models. Defining profiles probabilistically (rather than deterministically) can capture the uncertainty of exactly what product the competitor will introduce.

In summary, this paper makes a valuable contribution to the field of conjoint analysis in that it reviews for us some of the major issues faced by modelers.

SCALING PRIOR UTILITIES IN SAWTOOTH SOFTWARE'S ADAPTIVE CONJOINT ANALYSIS

William G. McLauchlan
McLauchlan & Associates, Inc.

INTRODUCTION

Background

A variety of research methodologies and data collection techniques are currently employed in commercial and academic applications of conjoint analysis. Included are full-profile, self-explicated, "hybrid," and graded paired-comparison designs. Regardless of the methodology, the ultimate value of any conjoint design resides in its ability to correctly predict choice behavior from the resultant utility data.

Because marketplace choice behavior is always impacted by variables extraneous to the environment in which predictor data are collected, the validity of conjoint techniques is often assessed using internal criteria, such as holdout concepts. The better able we are to predict internal holdout behavior, the more comfortable we feel about a given technique as a basis for making external decisions.

Since its introduction, Sawtooth Software's Adaptive Conjoint Analysis (ACA) has enjoyed a large and loyal following. As such, it is not surprising that ACA has begun to receive attention in the literature and at conferences in terms of its ability to correctly predict choice behavior (Finkbeiner and Platz 1986, Huber and Hansen 1986, Finkbeiner 1988, Herman 1988, Agarwal and Green 1989).

Most recently, a paper in press by Green, Krieger and Agarwal (1990) challenges ACA's utility estimation procedures in a number of areas and offers several suggestions for improving those procedures. One specific criticism of ACA is that attribute importance ratings are "too coarse; only four response values are permitted." (In ACA Version 3.0, prior utilities are calculated by weighting reflected attribute level preference ranks by the importance ratings. The priors are updated in real-time by OLS regression in the graded-pairs section of the interview.) The hypothesized implication is that scale incompatibility between the preference section (ranks), the importance section (4-point ratings), and the graded pairs section (9-point ratings) leads to more extreme response patterns than would be predicted given the objective of ACA to present pairs which are nearly equal in utility. Green *et al.* encourage the use of "finer-grained" importance scales (e.g., 1-10 or even analog measurement) as a way of increasing the predictive accuracy of ACA.

The purpose of the study reported here was to empirically evaluate the impact of alternative scaling on the ability of ACA to predict holdout choice behavior. Specifically, with the cooperation of Sawtooth Software, three versions of ACA were obtained and evaluated:

- Current ACA Version 3.0

  In this version of ACA, attribute-level preferences are obtained through ranking. Attribute importance is collected using a 4-point rating scale. In the paired tradeoffs section, evaluations are made using a 9-point scale (1= Strongly Prefer Left, 5=No Preference, 9=Strongly Prefer Right). (See Johnson 1988 for a complete description of utility calculations).

- ACA Version 3.0 with a 9-point importance scale

  This version is identical to current ACA with the exception that the importance rating is obtained using a 9-point scale (each importance rating is multiplied by 4/9 before the priors are calculated).

- Analog ACA

  Prior utilities are elicited directly in this version of ACA. The preferred level within each attribute is assigned a value of 100. Using the cursor keys, the remaining levels are moved on the screen and placed by the respondent along a preference continuum which ranges from 100 to 0 points (the priors are then rescaled to range from -50 to +50). As such, importance ratings, *per se*, are not evoked.

The subject of the study was consumer preference for grocery store features. A total of 10 attributes were investigated. The attributes and levels were the same regardless of the ACA version (see Table 1).

TABLE 1

Grocery Store Attributes and Levels

BAGGING

You bag your own groceries
Store bags groceries for you

STORE HOURS

Open 24 hours everyday
Open 7AM to Midnight everyday
Open 7AM to 9PM everyday

DRIVE TIME

Less than 5 minutes driving time
5 to 9 minutes driving time
10 to 14 minutes driving time
15 to 29 minutes driving time
30 minutes or more driving time

DOUBLE COUPONS

Double coupons everyday
Double coupons once a week
Does not offer double coupons

FLORAL DEPARTMENT

Floral department with delivery
Floral department; no delivery
Does not have floral department

BRAND NAMES

Brand name products only
Brand name & store brands only
Brand name, store brands, generics

PHARMACY

Has a pharmacy
Does not have a pharmacy

BANK

Full-service bank inside store
Store has Automatic Teller Mach.
No bank or ATM inside store

LOCATION

Located in an enclosed mall
Located in a shopping center
"Free-standing" location

CANDY IN CHECKOUT LANES

All checkout lanes have candy
Some checkout lanes have candy
No checkout lanes have candy

Methodology

As described below, the sample was divided into three cells. In- person interviewing was used to collect the data. Respondents were intercepted and screened in six geographically dispersed shopping centers (Detroit, Houston, Oklahoma City, Orlando, Santa Fe, and Staten Island). Those individuals who qualified and agreed to participate were brought back to an enclosed room where the interview was administered on PC. Although an interviewer was present at all times, the respondents keyed-in all answers to the computer-presented questions.

Sample Composition/Size

To participate in the study, respondents had to meet the following qualifications:

- Primary Grocery Shopper
- Between 18 and 65 Years of Age
- Not Competitively Employed
- No Past 3 Months Research Participation

A total of 604 interviews were completed with respondents meeting these qualifications. Interviewing was conducted from October 16 through October 30, 1990.

Design

A three-cell design was used where the cells were differentiated based on the version of ACA to be administered. In each cell, both preceding and subsequent to ACA, each respondent also provided likelihood-to-shop ratings for five grocery store holdout concepts. As indicated in the questionnaire section below, the set of holdout concepts Pre- and Post-ACA were identical. Further, the holdout concepts were the same regardless of the respondent's cell assignment.

The key differences between the ACA versions tested, as well as the number of completed interviews by cell, are as follows:

	Cell 1	Cell 2	Cell 3
Importance Scale	4-Point	9-Point	
Preference Scale	Ranks	Ranks	Analog
Paired-Comparisons	9-Point	9-Point	Analog
Completed Interviews	206	201	197

To facilitate communication of results, Cell 1 will be subsequently referred to as the 4-Point Cell, Cell 2 as the 9- Point Cell, and Cell 3 as the Analog Cell.

Questionnaire

Regardless of cell assignment, the questionnaire sequence and content were the same for all respondents:

- Likelihood to shop ratings (0-100 points) for each of five grocery store holdout profiles (Table 2). All respondents in the study rated the same holdout concepts. The holdouts were initially constructed from a random combination of the attributes and levels included in ACA.

- ACA
  Preference: Ranks or Analog ratings

  Importance: 4-point, 9-point, or Analog ratings

  Paired Tradeoffs: 9 point or Analog ratings (11 pairs consisting of 2 attributes each were evaluated by each respondent)

- Likelihood to shop ratings (0-100 points) for the same five grocery store holdout profiles evaluated prior to ACA

- Profile of grocery store shopped most often

- Demographics

TABLE 2

Holdout Profiles

Profile #1

* You bag your own groceries
* Open 7AM to 9PM everyday
* 10 to 14 minutes driving time
* Double coupons once a week
* Has a floral department but no delivery
* Carries brand name, store brands, and generics
* Does not have a pharmacy
* Store has an Automatic Teller Machine
* "Free-standing" location
* None of the checkout lanes have candy

Profile #2

* You bag your own groceries
* Open 24 hours everyday
* Less than 5 minutes driving time
* Double coupons everyday
* Does not have a floral department
* Carries brand name, store brands, and generics
* Does not have a pharmacy
* No bank or Automatic Teller Machine inside store
* Located in an enclosed mall
* All of the checkout lanes have candy

Profile #3

* You bag your own groceries
* Open 7AM to Midnight everyday
* 5 to 9 minutes driving time
* Does not offer double coupons
* Has a floral department that will deliver
* Carries brand name and store brands but not generics
* Has a pharmacy
* Store has an Automatic Teller Machine
* Located in a shopping center
* Some of the checkout lanes have candy, some do not

(continued)

TABLE 2 *(continued)*

Profile #4

* Store bags groceries for you
* Open 7AM to Midnight everyday
* 15 to 29 minutes driving time
* Double coupons everyday
* Has a floral department but no delivery
* Carries brand name and store brands but not generics
* Has a pharmacy
* Full-service bank inside the store
* "Free-standing" location
* None of the checkout lanes have candy

Profile #5

* Store bags groceries for you
* Open 24 hours everyday
* 10 to 14 minutes driving time
* Does not offer double coupons
* Has a floral department but no delivery
* Carries brand name products but no store brands or generics
* Has a pharmacy
* No bank or Automatic Teller Machine inside store
* Located in an enclosed mall
* All of the checkout lanes have candy

RESULTS

The results of this study are presented in three sections: Pre-Post Holdout Evaluations, ACA First Choice Results, and Interview Length.

Pre-Post Holdout Evaluations

The validity of any conjoint technique, as reflected by the successful prediction of holdout choice behavior, is mathematically constrained by the reliability of the holdout task. As such, the discussion of the impact of priors scaling on holdout prediction is ultimately based on results among respondents for whom test-retest reliability was high. In the present context, reliability was assessed in three interrelated ways: consistency of preference in the Pre-Post holdout task, Pre-Post product-moment correlations, and an analysis of implied paired-comparisons. This section presents the results of these reliability analyses.

Pre-Post Consistency. For each respondent, the Pre-ACA likelihood-to-shop ratings on the holdout concepts were compared to the Post-ACA ratings on the same set of concepts. Shown below is the proportion of respondents in each cell that gave the highest rating to the same concept both Pre- and Post-ACA. (In all cells, respondents who had tied first-choice ratings had their first-choice "vote" divided evenly among the tied concepts.) As can be seen, the proportion observed among respondents in the Analog Cell is significantly lower than the proportions of consistent respondents in the 4-Point and 9-Point Cells.

Cell	Pre-Post First Choice Hits (%)
4-Point	65.5
9-Point	60.7
Analog	49.7*

* Significantly lower than the 4-Point and 9-Point results ($p<.05$)

The reliability of the holdout task and the criterion-based validity of ACA will ultimately be impacted by the degree of variation in appeal for the holdout concepts. Better predictive validity should be expected of ACA when the distribution of preference for the holdout concepts is skewed compared to when that distribution is uniform. To provide context for the subsequent discussion of validity, shown below, by cell, are the first-choice distributions for the holdout concepts, both Pre- and Post-ACA.

First Choice Shares of Preference

	4-Point		9-Point		Analog	
	Pre-ACA (%)	Post-ACA (%)	Pre-ACA (%)	Post-ACA (%)	Pre-ACA (%)	Post-ACA (%)
Concept						
1	5.3	6.8	7.0	9.0	7.1	9.1
2	24.8	25.7	20.9	26.9	23.9	20.8
3	14.6	14.6	17.9	13.9	10.2	19.3
4	38.8	31.6	31.3	29.4	35.5	36.5
5	16.5	21.4	22.9	20.9	23.4	14.2

Pre-Post Product-Moment Correlations. Pearson product-moment correlations were computed for each respondent using the Pre- and Post-ACA likelihood-to-shop ratings. Mean correlations were then calculated by converting the individual Pearson r's to Fisher's Z' scores, averaging over all respondents in a given cell and then converting back to a mean Pearson r. Shown below are the mean Pre-Post correlations by cell. Like the Pre-Post First Choice results discussed above, respondents in the 4-Point and 9-Point Cells were significantly more reliable in their holdout concept ratings than were their counterparts in the Analog Cell.

Cell	Average Pre-Post Correlation
4-Point	.52
9-Point	.52
Analog	.36*

* Significantly lower than the 4-Point and 9-Point results ($p<.05$)

Implied Paired-Comparisons. As a final measure of holdout task reliability, implied paired-comparisons were constructed and tested using a variation on a technique described by Huber and Hansen (1986). In this analysis, each holdout task is viewed as implying 10 paired-comparisons (the first-choice concept is preferred over the remaining four, the second-choice is preferred over the remaining three, and so forth). To assess reliability, 5x5 one-zero (preferred-not preferred) paired-comparison matrices were composed for each holdout task for each respondent. The matrices were then multiplied using Boolean logic to produce individual respondent paired-comparison "hits" matrices. For each respondent, then, the proportion of correct Pre-Post implied paired-comparisons is calculated. As an example of the calculations, if a respondent's holdout ratings were:

	Concept				
	1	2	3	4	5
Pre-ACA	20	15	50	100	90
Post-ACA	25	30	65	90	95

then the implied paired-comparison matrices would be:

	Pre-ACA Concept				
	1	**2**	**3**	**4**	**5**
1	-	0	1	1	1
2	1	-	1	1	1
3	0	0	-	1	1
4	0	0	0	-	0
5	0	0	0	1	-

	Post-ACA Concept				
	1	**2**	**3**	**4**	**5**
1	-	1	1	1	1
2	0	-	1	1	1
3	0	0	-	1	1
4	0	0	0	-	1
5	0	0	0	0	-

and the "hits" matrix would be:

	Concept				
	1	**2**	**3**	**4**	**5**
1	-	0	1	1	1
2	0	-	1	1	1
3	1	1	-	1	1
4	1	1	1	-	0
5	1	1	1	0	-

The proportion of implied paired-comparison "hits" would be 16/20 = 80%.

The observed average proportions by cell are shown below and are not significantly different at the p=.05 level.

Cell	Average Correct Implied Paired-Comparisons (%)
4-Point	65.8
9-Point	64.9
Analog	60.5

ACA First Choice Results

The ability of ACA to correctly predict holdout choice behavior was assessed in three ways: first choice hits predicted by ACA, holdout task variance explained by ACA, and an estimation of the proportion of error variance in ACA due to the unreliability of the holdout task.

First Choice Hits. For each respondent, the ACA-predicted first- choice profile was compared with the highest rated profile in each of the two holdout tasks. Shown below are the proportions of respondents in each cell for whom ACA correctly predicted holdout first choices (ties are counted as "hits" throughout). There are no significant differences among cells in the first choice accuracy of ACA for either holdout task.

	Proportion of First Choice ACA Hits	
	Pre-ACA Holdout	Post-ACA Holdout
Cell	(%)	(%)
4-Point (n=206)	34.5	36.9
9-Point (n=201)	35.3	37.8
Analog (n=197)	33.5	32.5

As previously noted, respondents in the 4-Point and 9-Point Cells were significantly more reliable in the Pre-Post ACA holdout tasks than were the respondents in the Analog Cell. However, when the data for consistent respondents are isolated and the holdout first choices are compared with ACA-predicted first choices, the resultant proportions for each cell are identical.

Cell	Proportion of First Choice Hits Among Consistent Respondents (%)
4-Point (n=135)	45.9
9-Point (n=122)	45.9
Analog (n=98)	45.9

Variance Explained. Mean r-squared values were computed between the holdout task likelihood-to-shop ratings and the corresponding sums of the ACA partworths. The analysis was conducted among the total sample and among those respondents consistent in their Pre-Post choices. In all instances, r-squares are based on product-moment correlations corrected for attenuation. Shown in Table 3 are the results of these analyses. None of the between, or within, cell differences are significant at p=.05 level.

TABLE 3

Variance Explained in the Holdout Tasks by ACA

	Total Sample Average R-Squared			
Cell	ACA and Pre-ACA Holdout Task	ACA and Post-ACA Holdout Task	Post-Pre Difference	Mean
4-Point (n=206)	.64	.71	+.07	.68
9-Point (n=201)	.64	.64	.00	.64
Analog (n=197)	.67	.69	+.02	.68

	Pre-Post Holdout Consistent Average R-Squared			
Cell	ACA and Pre-ACA Holdout Task	ACA and Post-ACA Holdout Task	Post-Pre Difference	Mean
4-Point (n=126)	.63	.69	+.06	.66
9-Point (n=112)	.56	.59	+.03	.58
Analog (n=88)	.65	.68	+.03	.66

To provide insight into how great the impact of holdout task unreliability can be on the estimates of the holdout variance explained by ACA, the observed mean R-squared values uncorrected for attenuation are reported below.

Cell	Total Sample	Holdout Consistent
4-point	.18	.23
9-point	.08	.16
Analog	.16	.27

The predictive improvement in holdout behavior, over chance, that can be attributed to ACA can be inferred from $1-\sqrt{(1-r^2)}$. This calculation reveals the extent to which error in criterion prediction is reduced by ACA. When $1-\sqrt{(1-r^2)}$ equals 0, the error in ACA prediction will equal the error observed if holdout preferences were simply guessed. Shown below are the results of this calculation for the total sample and for respondents who were consistent in their first choice holdout behavior.

	Total Sample	Holdout Consistents
	$1-\sqrt{(1-r^2)}$	$1-\sqrt{(1-r^2)}$
4-Point	.43	.42
9-Point	.40	.35
Analog	.43	.42

Estimating ACA Error Variance. As an extension to the above analyses, the relationships between ACA predictions and holdout task behavior were evaluated using an approach described by Johnson (1989). This analysis consisted of the following calculations:

1. Estimation of the correlation between the ACA predictions and a perfectly reliable choice measure.

2. Estimation of the error variance in ACA prediction that would be expected if the holdout tasks were perfectly reliable.

3. Calculation of the observed ACA error variance (uncorrected for attenuation).

4. Calculation of the proportion of observed ACA error variance due to unreliability of the holdout tasks.

Results of these calculations are shown in Table 4 for both the total sample and for those respondents who were first-choice consistent at the holdout task. As can be seen, for the total sample, and also for those respondents who were holdout consistent, more than half of the error observed in ACA's predictions is apparently due to unreliability in ratings of the holdout concepts.

Table 4

Estimation of Error Variance in ACA

Total Sample

	Cell		
	4-Point (206)	9-Point (201)	Analog (197)
Estimated r Between ACA and Perfectly Reliable Choice Measure	.82	.80	.82
Relative Error Variance in ACA if Holdout were Perfectly Reliable	.32	.36	.32
Observed Error Variance in ACA	.82	.92*	.84
Proportion of ACA Error Variance due to Unreliability in Holdout Task	.61	.61	.62

* Significantly higher than the 4-Point and Analog results ($p<.05$)

Pre-Post Holdout Consistent Respondents

	Cell		
	4-Point	9-Point	Analog
	(126)	(112)	(88)
Estimated r Between ACA and Perfectly Reliable Choice Measure	.81	.76	.82
Relative Error Variance in ACA if Holdout were Perfectly Reliable	.34	.42	.34
Observed Error Variance in ACA	.77	.84	.73
Proportion of ACA Error Variance due to Unreliability in Holdout Task	.56	.50	.54

Interview Length

To determine if interview length would vary as a function of the type of scaling, the time spent completing the ACA portion of the interview was measured for each respondent using the PC's internal clock. Shown below, by cell, are the mean times, in minutes, that the ACA portion of the interview took to complete.

Cell	Average Length of ACA Interview
4-Point	7.2 minutes
9-Point	7.4 minutes
Analog	11.5 minutes*

* Significantly longer than the 4-Point and 9-Point cells ($p<.05$)

DISCUSSION

Validity

The results of this study refute the contention that scaling incompatibilities lead to degradation in the ability of ACA to correctly predict holdout behavior. This conclusion is based on the following key findings:

- There are no significant differences among cells in the proportions of first-choice hits by ACA for either the Pre-ACA or Post-ACA holdout tasks.
- The proportions of first-choice hits by ACA among respondents who were consistent in their holdout task preferences were identical for the three scaling alternatives.
- The variance in the holdout ratings explained by ACA is constant across the three cells.
- The prediction error attributable to ACA does not vary as a function of scaling differences.
- The estimated error variances in ACA, given a perfectly reliable holdout task, are equal regardless of the type of scaling.

In spite of the results of this study, it might still be argued that constancy in scaling, if nothing else, would be less disconcerting to a respondent as he/she shifts from question type to question type as the interview proceeds. If the psychological comfort of the respondent is of concern, the point would be well taken. However, there is nothing in the data reported here to suggest that ACA suffers because of variations in scaling.

As noted above, shifting to the "finest" level of scaling significantly increases the length of the ACA portion of the interview. In the absence of evidence to support the hypothesis that finer scaling would improve the validity of ACA, the longer interview length is, by itself, sufficient reason to avoid changing ACA scales to analog measurement.

Reliability

As shown by the following findings, the reliability of the holdout task degrades as ACA scaling becomes "finer."

- The proportion of Pre-Post holdout hits in the Analog Cell was significantly lower than the proportions observed in the 4-Point and 9-Point Cells.
- The Analog Cell Pre-Post holdout ratings were significantly less correlated than were the ratings in both the 4-Point and 9-Point Cells.

It could be hypothesized, in the same spirit that motivated this research, that the significantly lower holdout reliability in the Analog cell results from the psychological, if not mathematical, impact of variations in scaling between the holdout tasks and the ACA interview. The reliability may have been higher if the holdout ratings had been collected using analog measurement as well.

Because predictive validity is constrained by the reliability of the criterion variables, efforts to improve ACA (or any conjoint technique) are best assessed in an environment where the criterion is measured as reliably as possible. Anastasi (1976) has stated that the criteria most often used in validating tests are, in themselves, dynamic rather than static. As a consequence, criterion-based validity will always be transient. Furthermore, Johnson (1990) reported that the ACA experience can actually help respondents clarify their value systems.

The ACA interview should be expected, then, to have a bearing on the test-retest reliability of the holdout task. As such, the Post-ACA holdout tasks should be better predicted by ACA than are Pre-ACA tasks. In fact, the variance explained in the Post-ACA holdout ratings by the ACA predictions is higher, though not significantly so, in all three cells among consistent respondents.

Finally, ACA is too often held up against paper-and-pencil conjoint techniques (and vice versa) in validity "showdowns" (Herman, 1990). 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? Probably not. Are value systems impacted by the ACA experience? Probably so. Definitive answers to these questions will evolve only from further empirical research and not from hyperbole.

REFERENCES

Agarwal, M. K. & Green, P. E. "Adaptive Conjoint Analysis versus self-explicated models: Some empirical results," Working paper, State University of New York at Binghamton, 1989.

Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.

Finkbeiner, C. "Comparison of conjoint choice simulators," In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1988.

Finkbeiner, C. & Platz, P. "Computerized versus paper and pencil methods: A comparison study," Paper presented at the Association for Consumer Research Conference, Toronto, 1986.

Green, P. E., Krieger, A. M., & Agarwal, M. K. "Adaptive Conjoint Analysis: Some caveats and suggestions," *Journal of Marketing Research* (in press).

Herman, S. "Software for full-profile conjoint analysis." In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1988.

Herman, S. "Notes on conjoint analysis," Issue 2. New York: Bretton-Clark, 1990.

Huber, J. & Hansen, D. "Testing the impact of dimensional complexity and affective differences on paired concepts in Adaptive Conjoint Analysis." In M. Wallendorf & P. Anderson, P. (Eds.), *Advances in Consumer Research.* Provo, UT: Association for Consumer Research, 1986.

Johnson, R. M. Comment on Green *et al. Journal of Marketing Research* (in press).

Johnson, R. M. "Assessing the Validity of Conjoint Analysis." In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1989.

Johnson, R. M. "Adaptive Conjoint Analysis," In *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, 1987.

Comment on McLauchlan

Paul F. Hase
Hase-Schannen Research Associates

The topic of this paper is a critical one, given the importance of the prior utilities' estimates in the derivation of respondents' value systems.

Since this paper presents the results of a single investigation, the author's recommendation for further empirical work (and the publishing of the results!) is strongly endorsed.

My comments fall into two categories — those specific to the paper's methodology and those relating to the general issue of scaling priors.

COMMENTS SPECIFIC TO THE PAPER

Regarding the low pre-post reliability in the "analog" cell, I assume the samples were balanced, and the differences observed are not due to bias. The paper does not explicitly mention this consideration.

Assuming that the samples are matched, it is troubling that the analog exercise could "cause" the much lower pre-post first choice hit rates and correlations. Since I do not know how the analog method was operationalized (I have not seen the actual execution), perhaps the self-explicated priors could be collected using "traditional" keyboard procedures.

It should also be noted that since the pairs were collected via analog means in that cell and by current ACA (Adaptive Conjoint Analysis) methods in the other cells, the issue of the effect of analog data collection on the scaling of priors is confounded and not "purely" measured in this experiment.

Were eleven pairs enough? While it is understandable to keep the interview short and costs down, it would be greatly preferred to have had more observations. With 30 total levels and 10 attributes, Sawtooth would recommend 27 pairs. With just 11 pairs, 30 of the 41 total observations in the final estimating regression would be from the priors, which thus have too much influence on the final utilities obtained. It would thus be expected that the ability to predict the holdout choices would be negatively affected, especially since the concepts were fairly close in "total value" (generated at random, as opposed to being developed systematically to ensure the presence of good and bad ones). I think we all would have preferred to see more observations and thus greater influence of the tradeoff questions in the utility estimation.

Regarding the random generation of the holdout concepts, I would expect them to be "close" in total value, thus increasing the difficulty of proving or disproving the predictive ability of each method. I would have preferred to have included one "good" concept and one "not so good" one (but not the

best and worst — that would be too obvious) and three inbetween. Then I would test both "first choice" and "last choice" predictions (for if we can't do this, then there is a problem with our methods), as well as predictions of those falling between the extremes (which is pretty much what happens with random concept generation).

We did not do well in our first choice predictions in this experiment (30% range for the total sample, 46% among "consistents"). We would have gotten mid-30's if we assigned everyone to Concept 4, the highest scoring one. However, given the proximity of the concepts, this relatively low hit rate is not surprising.

I also believe that holdout analysis using the first choice criterion is incomplete, in that good fit across the full spectrum of high to low is relevant for many situations in which we model. Being able to estimate how close one concept is to another does matter for estimating logit preference shares and what it would take to become the first choice product (that is, the distances among the concepts).

GENERAL ISSUE (SCALING OF PRIORS) COMMENTS

Even though this paper would lead one to conclude that the method of scaling priors does not matter in the ability to predict holdouts, and whether or not additional research on this issue is ever published, from a "face validity" point of view I would still rather see us move to a self-explication of values for all levels within each attribute (possibly using keyboard, non-analog response entry).

My concerns with the current ACA approach for scaling priors are twofold:

- Even if the attribute levels themselves are equidistant (for ordered attributes), the current assumption that peoples' utilities are equal-intervaled might very well not be the case (and probably is not). For example, a person's delta for 30 mpg vs. 35 mpg might be much smaller than for 20 mpg vs. 15 mpg.

- On the other hand, if ordered attribute levels are not equal-intervaled (which they often are not) they are even more likely not to have equal-intervaled utilities.

Self-explication overcomes both these issues, in that people can be shown attributes with equal or unequal intervals and give utilities with equal or unequal intervals. Whether or not this task is less or more difficult than the separate ranking (unordered attributes) and importance rating tasks is still unresolved.

INCLUDING INTERACTIONS IN CONJOINT MODELS

Carl T. Finkbeiner and *Pilar C. Lim*
National Analysts Division of Booz•Allen & Hamilton, Inc.

INTRODUCTION

Our intention in this paper is to describe interaction effects in the conjoint analysis context and why you should want to use them. Think of this as a motivational paper; it is not a "how-to" paper, nor, except for the appendices, is it especially technical. We assume only that you have some passing familiarity with conjoint analysis, its main outputs (partworths or utilities, and importances), and the important supplementary use of conjoint results in simulation of preference shares.

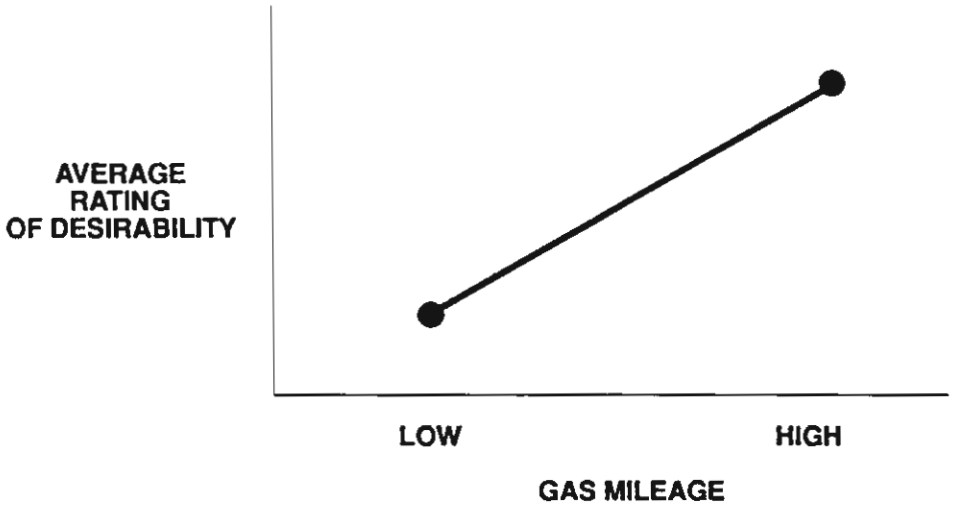
Most applications of conjoint analysis in marketing involve "main effects" only. In estimating the partworths and importance effects for each respondent, the typical conjoint assumes that the effect of an attribute on the respondent's preference judgment depends only on its level, its so-called main effect, and not on the levels of other attributes. This assumption holds for Sawtooth's ACA (ACA System for Adaptive Conjoint Analysis) as well as most other commonly used approaches.

To illustrate main effects, suppose that a researcher is interested in determining how price and gas mileage affect the desirability of an automobile. If a higher price tends to detract from the desirability, we would observe the rating pattern shown in Figure 1.

Figure 1
AVERAGE RATING FOR AUTOMOBILE AT TWO LEVELS OF PRICE

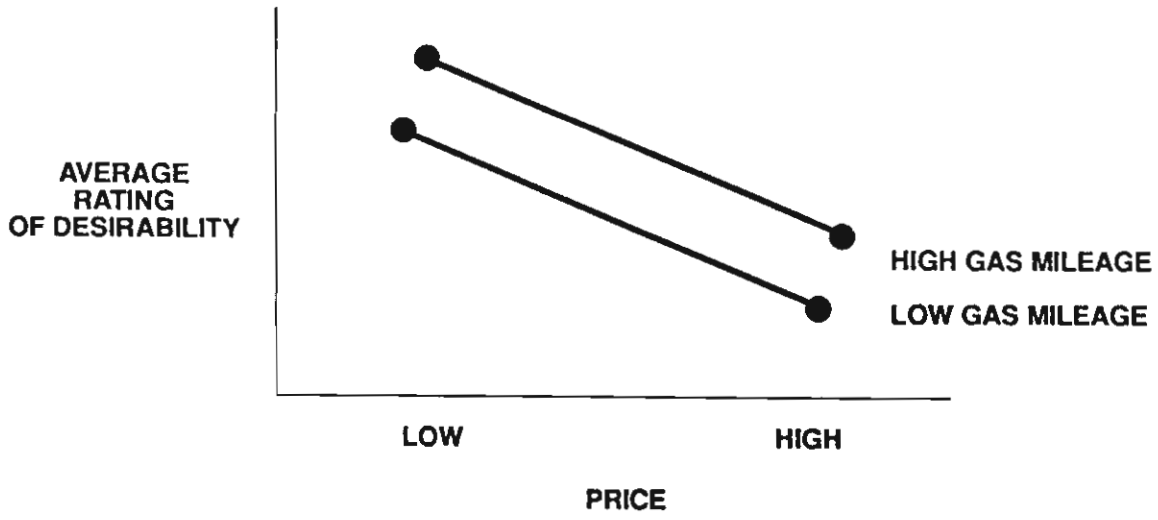
If higher gas mileage enhances the desirability of the automobile, we would observe the rating pattern shown in Figure 2.

Figure 2
AVERAGE RATING FOR AUTOMOBILE AT TWO LEVELS OF GAS MILEAGE



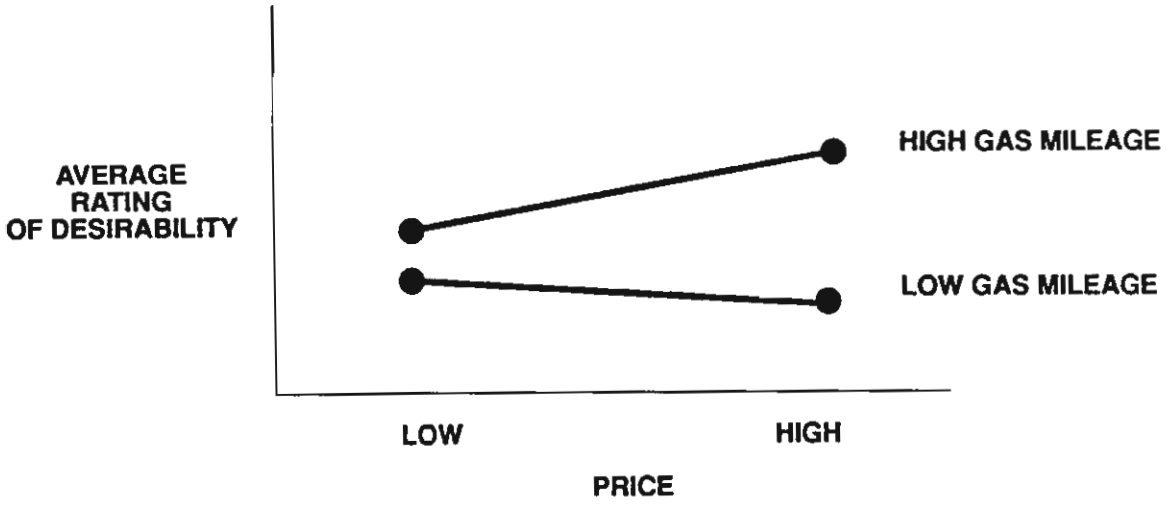
In a main effect situation (that is, when the desirability of price and gas mileage do not depend on one another), we would expect to observe the same pattern for low gas mileage at both levels of price, as Figure 3 shows.

Figure 3
AVERAGE RATING IN A MAIN-EFFECT SITUATION



However, in many situations the effect of one attribute upon the rating of an automobile depends upon the levels of some other attributes; the attributes are said to interact. A two-way or first order interaction is demonstrated in Figure 4.

Figure 4
AVERAGE RATING IF ATTRIBUTES INTERACT



Notice that the effect of changing price from low to high depends on the level of gas mileage. The attributes now interact in such a way that higher gas mileage actually enhances the desirability of the automobile with a higher price.

We believe there are many conjoint research situations in which, to fully explain judgments, it is necessary to estimate first order interactions among a group of attributes, producing a set of partworths (also called "utilities") for both main effects and interactions. To accommodate the interactions, the analysis requires more data from respondents than we need for main effects only. The estimation of partworths and importances is also slightly more involved, as the next section shows.

What advantage is gained by including interactions in conjoint? Recall that the ultimate objective of conjoint analysis is to produce a regression equation using partworths to predict the "total utility" (i.e., desirability) for any possible product as accurately as possible. Failure to include conjoint interactions, in cases where the market is interaction-ridden, clearly detracts from the purpose of the conjoint study by creating inaccuracies.

Treating conjoint analysis as the prediction of ratings using the main effects and interactions as independent variables in a regression framework, the problem of ignoring interactions becomes akin to the problem of omitted variables (specification error) in econometrics. Since only main effects are included in standard approaches to conjoint, the respondent's use of the missing interactions

in making judgments introduces noise and bias. An indication of the magnitude of the resulting noise and bias in the estimates is shown formally in Appendix B in terms of the partworths from which total utilities are computed. This translates into less accurate predicted total utilities.

Since predicted total utilities are then converted into preference shares using some choice model, omitted interactions further bias determination of the preference shares. This means that we will be less likely to produce externally valid market forecasts. Hence, it is doubly important to include interaction terms in a conjoint model when such interactions actually exist.

FULL-PROFILE CONJOINT WITH INTERACTION

The focus of this paper is on the form of conjoint analysis called full-profile conjoint analysis which is characterized in two respects: (1) it is full-profile in that responses are evaluative ratings made by each respondent of a set of product profiles, each comprised of all attributes, and (2) the method of analysis presumes a separate regression model for each respondent. We typically create a set of cards, each of which has a description of a hypothetical product in terms of all attributes.

SAMPLE CONJOINT CARDS

Automobile A	
Manufacturer:	Foreign
Number of Doors:	Two
Performance:	High
Gas Mileage:	High
Price:	Low

Automobile B	
Manufacturer:	American
Number of Doors:	Four
Performance:	Low
Gas Mileage:	High
Price:	Medium

(Note - this is a constructed example: in general we do not recommend using vague level labels such as "high" gas mileage or "medium" price.)

The respondent judges a set of such cards by physically sorting them onto a rating board such as the following:

Not at all likely to buy							Extremely likely to buy
1	2	3	4	5	6	7	8

We then treat the ratings for a single individual as the dependent variable in a least squares regression in which the attributes are the independent variables.

The typical full-profile conjoint determines partworths for main effects only. Our purpose here is to extend this technique to determine partworths for interaction effects as well. Main-effect partworths are associated with various levels of each attribute while interaction partworths correspond to combinations of levels of the attributes thought to interact.

We do have reasons for focusing on the approach described above. There are a number of alternative conjoint techniques which involve rankings instead of ratings, or non-metric regression models instead of least squares regression, or pairwise attribute comparisons instead of full-profile ratings, or, of course, adaptive partial-profile paired comparisons (as in ACA). The technical arguments surrounding these techniques are legion; however, full-profile conjoint using ratings and least-squares regression is one of the most (if not the most) commonly used techniques in practical applications. Furthermore, in our experience, it performs at least as well as other approaches and is better than some when it is appropriate. The arguments concerning the legitimacy of including interaction effects apply to all approaches of conjoint; the details of how to accomplish it are specific and we describe them only for the full-profile approach we are focusing on.

We should point out that we are not introducing any technology that hasn't already been invented elsewhere. The mechanics of designing experiments and estimating interactions has appeared long since in the experimental design literature. However, it has never really been applied in full-profile conjoint for practical market research studies supporting real-world marketing decisions. The basic reason is the obscurity of the source material: it is technically very difficult and hard to find. We are not so much introducing something new as urging practitioners to use an existing technology which we believe will be of significant benefit in understanding respondent preferences.

We should note one conjoint approach which has been used in practical market research to estimate interactions: the discrete choice experiments approach of Louviere & Woodworth (1983). However, this approach does not entirely meet the needs that the more standard full-profile approach meets. For one thing, the approach is a model of choices and not of preferences and if the user's objective is the estimation of preferences (as in customer satisfaction studies), then the choice experiments approach is not applicable. For another thing, the choice experiments approach typically provides only aggregate partworths and importances and does not provide estimates for each respondent separately. This is both a conceptual and a practical difficulty if you either: (1) believe that different people value different attributes, or (2) wish to do exploratory comparisons of sub-samples of respondents on their partworths/importances/preference shares. This is not to say that the choice experiments approach is never useful, but simply that it does not solve exactly the same problems being addressed in this paper for the standard full-profile approach.

The next part of this section describes the construction of interaction designs (that is, determining the layout of the cards to use in the study) while the last part deals with the intricacies of analyzing conjoint data with interactions.

Conjoint tasks utilize the principles from the design of experiments in statistics to ensure the measurability of the partworths. One such design is called the full factorial design. This requires that a respondent rate cards containing every possible combination of levels for all of the attributes. A full factorial design provides enough data to obtain estimates of the effect of all possible interactions among the attributes. A desirable feature of the full factorial design is that the levels of an attribute occur equally frequently with each of the levels of any other attribute. Because of this feature, the attributes are completely unconfounded with one another, thereby allowing independent estimates of all main effects and interactions. Independent estimates allow us to completely separate the unique contribution of one effect from the other.

The properties of an efficient design as well as the construction of such a design are best illustrated by example. Consider a study with two attributes A and B, each having two levels denoted by 1 and 2. The following represents the full factorial design for this simple 2x2 study, which produces four cards:

Card	A	B
1	1	1
2	1	2
3	2	1
4	2	2

The property of independence of estimates is achieved if the levels of an attribute occur equally frequently with each of the levels of any other attribute. Notice that each combination of levels in our example occurs an equal number of times, namely one.

The preceding example involved a full factorial design. A limitation of such designs is that for a moderate number of attributes and levels, an unreasonable number of cards would be required. To illustrate, for a conjoint study with six attributes, two attributes each at two, three and four levels, a total of 576 cards would be necessary. This is an impossible number of ratings for an individual to complete in a conjoint task.

We resort to fractional designs to cope with the problem. These are designs which require only a fraction of the number of cards needed for the full design. We choose designs that permit estimation of all main effects and of some interactions as well. The designs are built to be "orthogonal" (a technical property which ensures independence) and assume that certain interactions are negligible.

The equal frequency condition mentioned earlier is a sufficient condition to produce independent estimates of all main effects and interactions. For main-effects-only designs, however, this condition is not necessary, as shown by Addelman (1962). Addelman establishes a sufficient condition for the estimates of the main effects of any two attributes to be independent: that the number of times a level of one attribute occurs with one of the levels of the other attribute is a proportion of the number of times each level occurs in total. This is more specifically stated using equations.

Let:

n = number of cards
n_i = number of times the *i*th level of attribute A occurs in the design
n_j = number of times the *j*th level of attribute B occurs in the design
n_{ij} = number of times the *i*th level of attribute A occurs with the *j*th level of attribute B.

Addelman's condition for the main effects to be independent (called the proportional frequency condition) is given by:

$$n_{ij} = n_i n_j / n$$

Note in our simple 2x2 design that this condition holds for every card. For example, the number of times A has a level of 1 and B has a level of 2 (1 time) is equal to the number of times level 1 of A appears on any card (2 times) multiplied by the number of times level 2 of B appears (2 times) divided by the number of cards (4): i.e., 1=2(2)/4.

The proportional frequency condition provides us with a simple rule for determining whether a design will provide the independence we desire if we are estimating only the main effects. If we are estimating interactions as well, the equal frequency condition (all combinations of the two attributes involved in an interaction must appear an equal number of times) must be used to evaluate the design for independence.

To illustrate, we compare a main-effects-only design with an interaction-plus-main-effects design (referred to hereafter as an interaction design) for a conjoint study involving four attributes. Three of the attributes occur at two levels (denoted by 1 and 2), and one attribute occurs at three levels (denoted by 1, 2, and 3). The full factorial design would require 24 cards (=2x2x2x3) in this case. However, efficiencies are possible. A main-effects-only design using Addelman's procedure requires 8 cards and is given below.

Card	A	B	C	D
1	1	1	1	1
2	2	2	2	1
3	2	1	1	2
4	1	2	2	2
5	1	2	1	3
6	2	1	2	3
7	2	2	1	2
8	1	1	2	2

Consider level 1 of attribute A and level 3 of attribute D. We find that $n_{13}=1$ since the combination (1,3) of attributes A and D occurs 1 time. We also find that $n_1=4$, $n_3=2$ and the proportional frequency condition is satisfied with

$$1 = 4(2)/8.$$

We can show similarly that the condition is satisfied for every combination of levels of the attributes. This implies that the above design will produce independent estimates of the main effects of any two attributes.

The question arises: Could we use a main effects design to evaluate interactions? Generally the answer is no. The reason is twofold. Typically, in a main-effects-only design, almost all degrees of freedom are accounted for by main effects. We will thus have few, if any, degrees of freedom to accommodate interactions. Second, the interaction effects will not usually be orthogonal to the main effects, as the following example demonstrates.

Consider the interaction between A and D in the previous design. We have the following combinations of levels of attributes A and D and their frequency of occurrence in the main-effects-only design:

AxD	Frequency
1,1	1
2,1	1
2,2	2
1,2	2
1,3	1
2,3	1

We note that the combinations do not occur an equal number of times (that is, the equal frequency condition is not satisfied), indicating that we will be unable to estimate interaction effects independently of the main effects.

Let us now look at the interaction design for this conjoint study involving four attributes. This design was constructed using the method described in Appendix A. The design, which requires 12 cards, is as follows:

Card	A	B	C	D
1	1	1	1	1
2	1	2	2	1
3	2	1	2	1
4	2	2	1	1
5	1	1	1	2
6	1	2	2	2
7	2	1	2	2
8	2	2	1	2
9	1	1	1	3
10	1	2	2	3
11	2	1	2	3
12	2	2	1	3

While this is more cards than Addelman's main-effects-only design, it is still a more efficient design than the full factorial design which requires 24 cards.

Notice that combinations of levels of attributes A and B, A and C, and B and C occur 3 times each while combinations of levels of A, B, and C with attribute D occur 2 times each. In fact, the method used to generate the design ensures that all the levels of an attribute occur equally frequently with each of the levels of the other attributes; the design satisfies the equal frequency condition assuring independent estimates in interaction designs.

The remainder of this section discusses the estimation of partworths and importances once the conjoint data have been collected.

Partworths are calculated for each respondent using ordinary regression with certain constraints imposed to allow us to estimate each partworth uniquely. The regression model predicts the respondent's rating of a card (y_c) from the attribute levels present on the card (x_{ca}). More formally, we say that y_c is the sum of partworths (p_a) associated with each of the x_{ca}'s, plus error in the regression.

$$y_c = \Sigma\ p_a x_{ca} + \text{error}$$

where the sum is taken across all attributes. Note that the sum of partworths in the above equation is called the total utility and is the estimated rating. The weights in this regression (p_a) are in fact the

partworths associated with the attribute levels. To be more general, we must divide the x_{ca}'s and the p_a's into two groups: main effects (x_{cm} and p_m) and interaction effects (x_{ci} and p_i). Then the regression model is:

$$y_c = \Sigma\ p_m x_{cm} + \Sigma\ p_i x_{ci} + \text{error}.$$

Whether we estimate the p_m's only (main-effects-only model) or the p_m's and p_i's (interaction model), the mechanics of carrying out the regression analysis is conceptually the same. The actual regression procedure is outlined in Appendix C.

In the standard main-effects-only conjoint, "attribute importances" are computed based upon differences in total utilities between the "best" and the "worst" possible product configurations. These values reflect the relative size of the difference that an attribute could make in the respondent's total utility for a product. In an interaction study, defining importances in this way may not be practical. First, the concept of attribute importance has no meaning when interactions are present, so we must define main and interaction effect importances instead. To find the "best" and the "worst" possible configurations requires computing total utilities for every one of the possible cards in a full-profile design (for larger designs this can be many thousands of cards) and even then it is questionable as to whether interaction effect importances are being uniquely determined by this approach. As a consequence, we have turned to assessing effect importance using a "percent of variance accounted for" measure, which follows naturally from the definition of conjoint as a regression model. In the present case, the variance being accounted for is the variance in total utilities across every possible card. An effect importance, then, is the relative contribution of the effect to the variance in total utilities, as described in Appendix C.

As an aside, we note that importances, no matter how calculated, can be expected to be much less stable statistically than partworths. The reason is that an importance can be expressed as a sum of random variables. In most cases, such a sum will have much greater sampling error than any one of the random variables (the partworths) which constitute it. Be cautious in your use of importances.

SIMULATED EXAMPLE

To gain a further appreciation of why including interaction effects in individual respondent conjoint models is beneficial, it is instructive to consider some examples. Appendix B shows the formal consequences of ignoring interaction effects in terms of statistical bias and sampling variance. However, statistical theorems may not be sufficient to give you a feel for the magnitude of difference that including interactions can make interpretation-wise. In this and the next section, we present examples comparing main-effects-only models against interaction models.

We begin with an analysis of simulated data in which we have constructed a conjoint problem with a hypothetical respondent population to go with it. Since we constructed this simulated example ourselves, we know what the "right" answers are when comparing the results of the main-effects-only modeling with that including interactions.

To make the example easier to understand, we pretend we are doing a study of automobiles. The attributes for this study are as follows:

Place of Manufacture
1. Foreign
2. Domestic

Number of Doors
1. Two-door
2. Four-door

Performance
1. Low
2. High

Gas Mileage
1. Low
2. High

Price
1. Low
2. Middle
3. High

We constructed a population of partworths to contain both main effects for all five attributes plus interaction effects between Price and each of the first four attributes. Three hundred hypothetical respondents were randomly drawn from a multivariate normal population using the following assumptions:

- Different partworths for each respondent are drawn from the normal population and, once drawn, are regarded as fixed for that individual (the population mean partworths are shown in Table 1; the covariance matrix is diagonal)

- Ratings for the individual are generated from the interaction conjoint model (using the above sampled partworths) with normally distributed errors added in

  - The errors are different for different individuals and are not regarded as fixed, so that the same individual produces different ratings when sampled repeatedly
  - The individual's error has a mean of zero and a variance drawn randomly from a chi-square distribution with degrees of freedom equal to that of the interaction model
  - Partworths and the error term are independent of one another

For our hypothetical study, the standard main-effects-only design requires eight cards. The proper interaction design, constructed following the principles outlined in Appendix A, requires 24 cards.

Using these two design matrices to estimate partworths, we obtain individual partworths for both models. The mean partworths, along with the "true" population mean partworths, are shown in Table 1.

Table 1
Mean Partworths

Effect	Pop.	Interaction	Main
Place of Manufacture			
Foreign	-.08	.03	.52
Domestic	.79	.68	.76
Number of Doors			
Two-door	-.38	-.47	.63
Four-door	1.09	1.18	.64
Performance			
Low	-.50	-.40	.43
High	1.20	1.11	.84
Gas Mileage			
Low	.35	.46	.80
High	.35	.25	.48
Price			
Low	1.85	1.87	.98
Middle	-.05	-.22	1.06
High	-.75	-.58	-.12
Price x Place			
Low,Foreign	.35	.18	—
Middle,Foreign	-.25	-.09	—
High,Foreign	.95	.98	—
Low,Domestic	.35	.53	—
Middle,Domestic	.95	.80	—
High,Domestic	-.25	-.27	—
Price x Doors			
Low,Two-door	.77	.83	—
Middle,Two-door	-.82	-.96	—
High,Two-door	1.10	1.19	—
Low,Four-door	-.07	-.12	—
Middle,Four-door	1.52	1.67	—
High,Four-door	-.40	-.48	—
Price x Performance			
Low,Low	1.35	1.25	—
Middle,Low	-.95	-.80	—
High,Low	.65	.62	—
Low,High	-.65	-.54	—
Middle,High	1.65	1.51	—
High,High	.05	.09	—
Price x Mileage			
Low,Low	-.35	-.43	—
Middle,Low	.35	.49	—
High,Low	1.05	1.01	—
Low,High	1.05	1.14	—
Middle,High	.35	.22	—
High,High	-.35	-.30	—

In Table 1, we see that the interaction model has estimated partworths that, on average, are close to the "true" population values. Of course, we should expect this, given that the population contains interactions by construction. What is interesting is the extent to which the main-effects-only model gives distorted results. In the main effects estimates, three attributes are particularly erroneous: Number of Doors, Gas Mileage, and Price. These main effects are plotted in Figures 5 - 7 for ease of comparison.

Figure 5 Main effect Partworths: Place of Manufacture

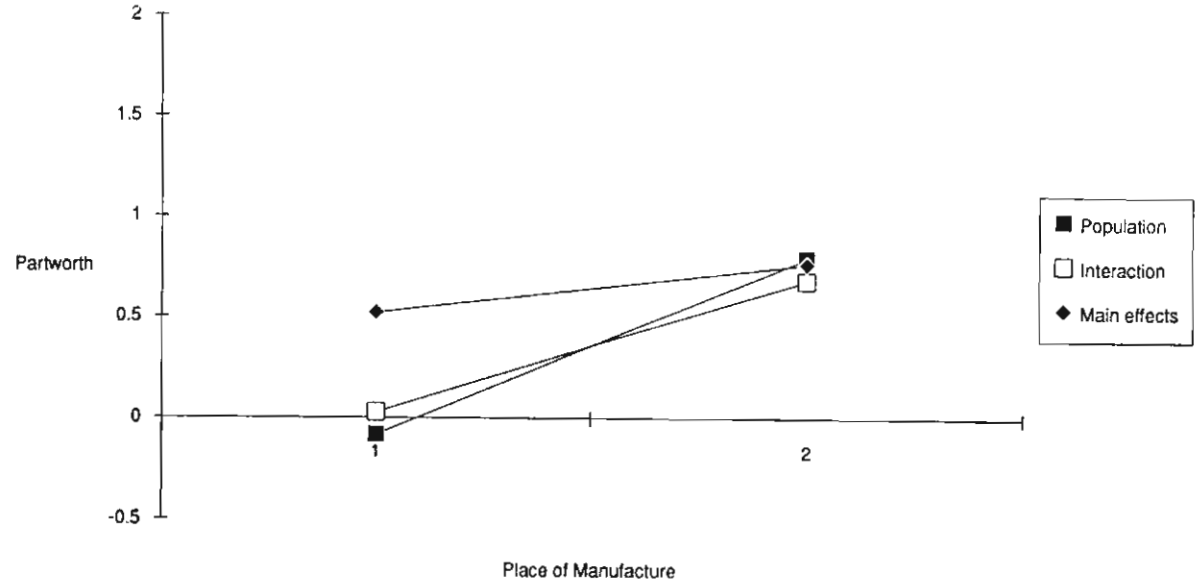


Figure 6 Main effect Partworths: Number of Doors

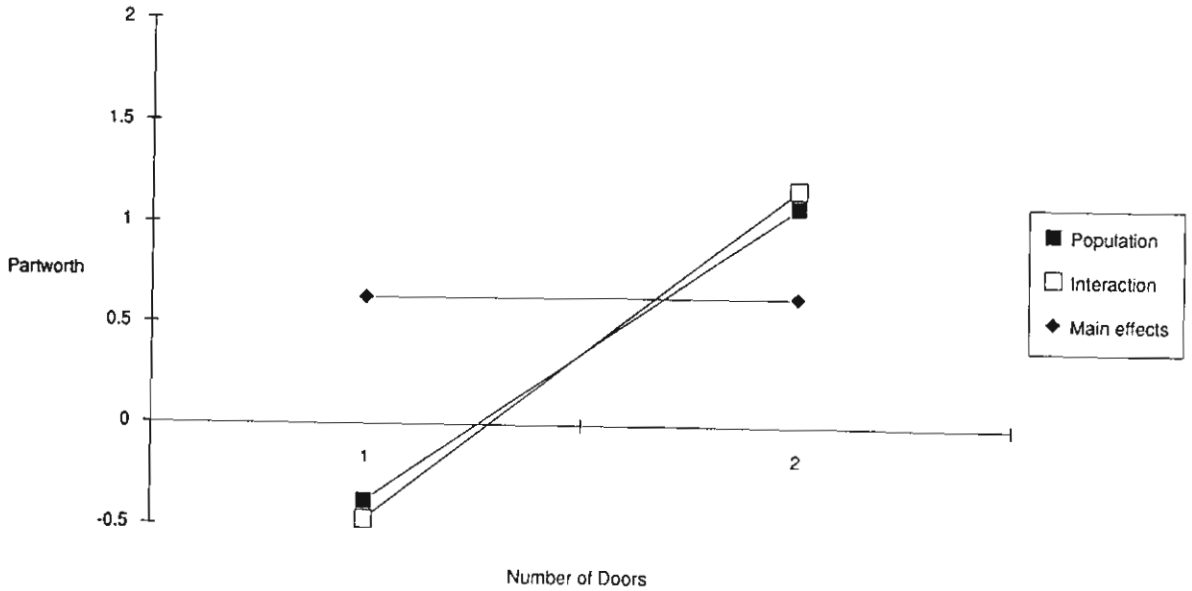
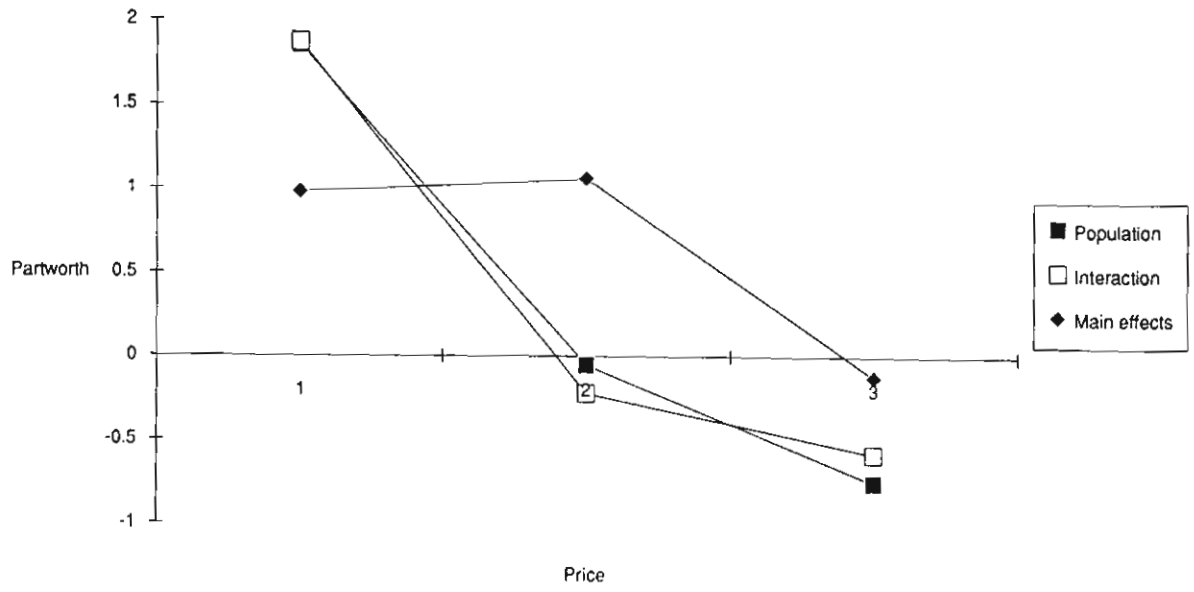


Figure 7 Main effect Partworths: Price



We see that the main-effects-only model would have us believe:

- Respondents aren't sensitive to Number of Doors
- They are sensitive to Gas Mileage
- They actually prefer slightly the middle Price over the low Price

In fact, the population exhibits no such trends in the main effects. However, it must be acknowledged that main effects are not the only effects of an attribute: the presence of interactions in general makes it difficult to specify the total effect of an attribute across both its main effect and any interactions in which it occurs. To meaningfully examine the total effect of an attribute, we must necessarily do so in the context of a specific product. That is, the effect on a product's total utility of varying the levels of an attribute depends upon which specific levels are present in other attributes. There is no single total effect of an attribute: it depends on the presence or absence of other attribute levels.

To compare the abilities of the two designs in predicting shares, we selected two products not contained in either the main-effects-only design or the interaction design. Using the same assumptions as were used in generating the original population, we can derive the exact joint distribution of the two products from which population preference shares can be calculated by applying normal distribution theory. Since we also know the "true" partworths and error variance for each of the 300 respondents in our sample, the "true" shares for the sample can also be determined using normal distribution theory. Note that the "true" sample share and the population share will not be identical because of sampling error, though they should be very close with a sample of size 300.

Finally, for each model, we can compute estimated sample shares by applying normal distribution theory, but using the *estimated* partworths and implied error variances. Results are shown in Table 2 for the least preferred of the two products.

Table 2
Shares for Least Preferred Product

	Share
True Share:	
Population	36
Sample	37
Estimated Sample Share:	
Interaction Model	38
Main-Effects-Only	55

Given that the partworths are biased and more errorful for the main-effects-only model, we expect there to be some deficiency in its ability to predict preference shares. In this particular instance, the main-effects-only model is producing a markedly erroneous estimate that leads to the wrong conclusion: that the product for which the share is being estimated is the preferred one.

REAL DATA EXAMPLE

Although the examination of simulated examples is illuminating, we always like to see how a procedure functions in a real-world application. We present here a full-profile conjoint study of the market opportunities for a new pharmaceutical product. To preserve client confidentiality, we have had to disguise the product category.

A national sample of 319 health care practitioners was administered a conjoint task in which the respondent rated the likelihood of prescribing each card for each of two recent actual patients. The interaction design consisted of 32 cards using the following attributes and levels.

Efficacy
1. Same as "standard" drug
2. More effective

Side-Effect 1
1. Same as "standard" drug
2. None

Side-Effect 2
1. Same as "standard" drug
2. None

Form of Administration
1. IV
2. Oral

Alternative Form Availability
1. Only one form available
2. Both forms available

Dosage Level
1. One
2. Three

Average Cost Per Dose
1. Very Low
2. Moderately Low
3. Moderately High
4. Very High

The 32 card design was selected to allow us to estimate the interactions between Cost and the other six attributes.

The number of cards required in a full factorial design would have been 256, far more than would have been practical. For comparison purposes, 16 cards would have been required for a main-effects-only design.

We did not have the luxury of administering separate studies for main-effects-only and interaction designs. Consequently, we used the ratings of a subset of 16 cards from the interaction design, of which a subset was sufficient for estimating a main-effects-only model.

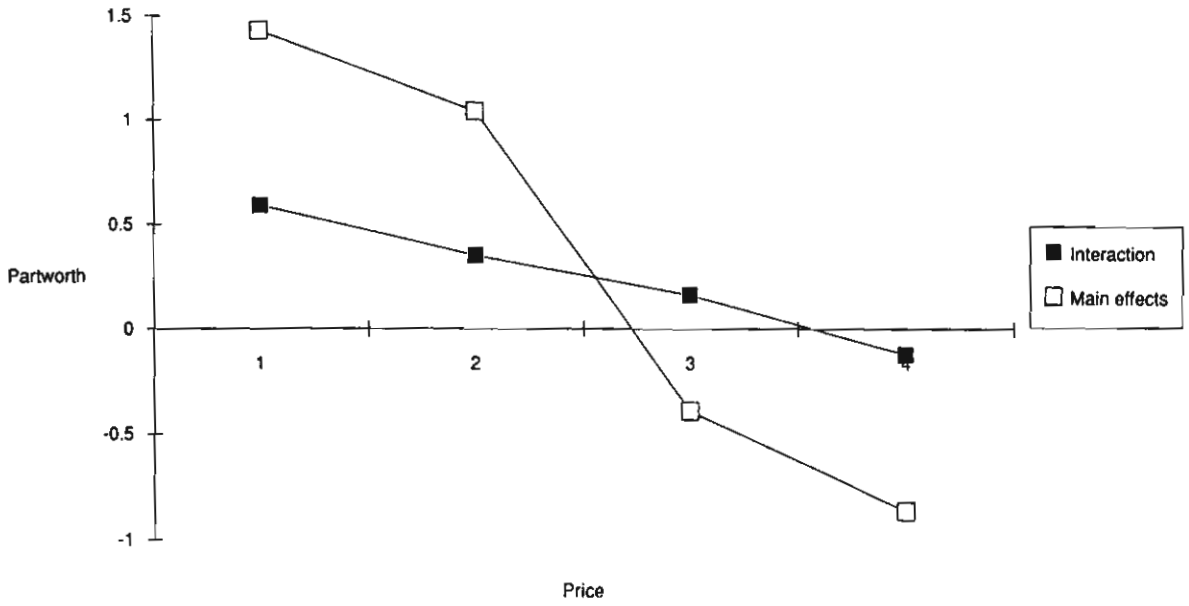
Table 3 shows the mean main effects partworth estimates obtained from both the main-effects-only model and the interaction model.

Table 3
Main Effect Partworths

Effect	Main	Interaction
Efficacy		
1. Same as "standard" drug	.19	.40
2. More effective	1.03	.59
Side-Effect 1		
1. Same as "standard" drug	.51	.43
2. None	.70	.55
Side-Effect 2		
1. Same as "standard" drug	.41	.33
2. None	.80	.66
Form of Administration		
1. IV	.75	.59
2. Oral	.47	.39
Alternative Form Availability		
1. Only one form available	.38	.34
2. Both forms available	.84	.64
Dosage Level		
1. One	.82	.64
2. Three	.40	.34
Average Cost Per Dose		
1. Very Low	1.43	.59
2. Moderately Low	1.04	.35
3. Moderately High	-.39	.16
4. Very High	-.86	-.12

Of particular interest to the client was the main effect of price, shown in graphical form in Figure 8.

Figure 8 Main-effects-only vs. Interaction Partworths: Price

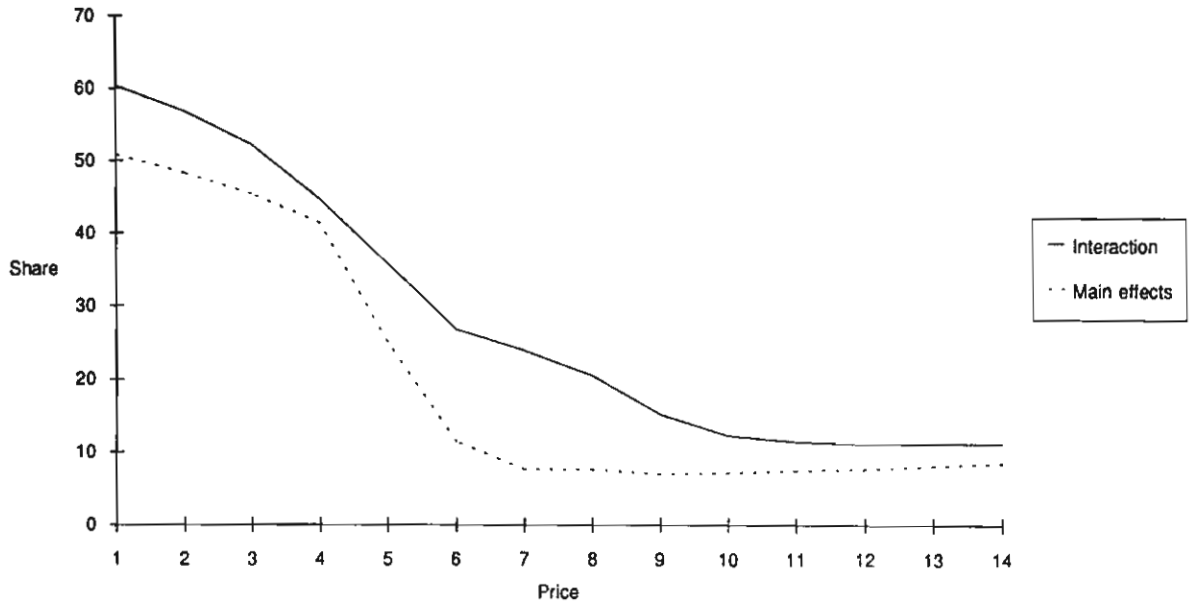


We see a marked difference in main effect price sensitivity, illustrating the difference in parameter estimates between the two models. Once again, we must point out that total price sensitivity depends upon the other attribute levels present in a given product.

To examine total price sensitivity differences, we turn to preference share estimation for the two models. Two competitor products are simulated using the conjoint model, along with a third "new" client product. Preference shares were estimated for all three products using a multinomial probit choice model (Finkbeiner, 1988). The share for the "new" product at each of a series of prices is shown in Figure 9.

The price curve for the interaction model now reflects the total effect of price, albeit in a specific context: all other attributes of the "new" product are held constant, as are all other products. If we changed one of the other attributes of the "new" product or one of the other products, we would expect to see a somewhat different price curve.

Figure 9 Main-effects-only vs. Interaction: Preference shares for "new" product



The two models indicate different price sensitivities for the "new" product, with the main-effect-only model showing a much sharper drop-off. Whether or not the difference would lead to a different pricing strategy depends upon the outcome of profitability analyses not shown here. Suffice it to say that because the preference shares differ so much in the middle price range, we conclude that a much higher price is profitable based on the interaction model than we do based on the main-effects-only model.

CONCLUSION

We have presented the tools needed to design and analyze a full-profile conjoint analysis study, including interactions in an individual-respondent conjoint model. The statistical developments and the simulated and real data examples provide a comparison of results from main-effects-only and interaction models. For some increase in the number of respondent ratings required (or conversely, a decrease in the number of attributes or levels), we have available a tool which, in many cases, will improve the accuracy of our assessments of the dynamics of individual preference judgments. Of course, there are times when a main-effects-only model is sufficient: notably when interactions are largely absent. In addition, there are times when the limitations imposed by the larger interaction designs make them impractical to use. However, interaction designs should always be seriously considered because of the significant accuracy improvements that they make possible.

APPENDIX A: GENERATING INTERACTION DESIGNS

To generate interaction designs, we use the concept of hypercubes of strength "d," [Rao, 1946] also called orthogonal arrays of strength "d." These are combinatorial arrangements which are derived from finite geometrical configurations. The hypercube-based technique described in this section provides fractional designs in which the attributes have equal numbers of levels. Also, the number must be expressed as the power of a prime in order to achieve a fractional design; that is, while all other number of levels between 2 and 9 are permissible, 6 levels will not allow a fractional design. The properties of Abelian groups and hypercubes are combined to give a system of fractional designs for a conjoint study with attributes occurring at different numbers of levels.

A hypercube of strength "d" is defined as follows. Suppose that there are m attributes A_1, A_2, ... , A_m each of which can assume s different values. We may define an ordered set $(i_1,i_2,...,i_m)$ as a combination of m attributes obtained by the selection of i_1-*th*, i_2-*th* ... values of the first, second, ..., attributes respectively. There are s^m such combinations of which a subset of s^t combinations is called an array. An array represented by (m,s,t,d) is said to be of strength d if all combinations of any d of the m attributes occur an equal number of times. When t=2 and d=2 the arrangement (m,s,2,2) corresponds to the existence of (m-2) orthogonal Latin squares of side s. For higher values of t and d, hypercubes can best be considered as ordered combinations to avoid the geometrical representation.

We present a summary of the technique used to derive the designs. Suppose that s, the number of levels, can be expressed as p^n, where p is a prime number. The finite projective k-dimensional geometry $PG(t,p^n)$ consists of the points $(x_0, x_1, ..., x_t)$ where $x_0, x_1, ..., x_t$ are elements of the Galois field $GF(p^n)$, at least one being different from zero.

This finite projective geometry is such that its points are a representation of the possible orthogonal sets of (p^n-1) degrees of freedom in the $(p^n)^{t+1}$ factorial system. Hence $PG(t, p^n)$ is a geometrical representation of (t+1) attributes, each at p^n levels, and their generalized interactions.

To arrange combinations of $(s^t-1)/(s-1)$ attributes, each at $s=p^n$ levels, in blocks of s^t cards, such that no main effects or two-way interactions are confounded with blocks, we establish a one-one correspondence between the attributes and the elements of $PG(t-1,p^n)$. Let A_1, A_2, ..., A_N denote the $N = (s^t-1)/(s-1)$ attributes. With these attributes we may associate a representation of the elements of $PG(t-1,p^n)$.

$$A_1 \; A_2 \; A_3 \; A_4 \; .... \; A_{s+1} \; A_{s+2} \; ... \; A_N$$

$$a_1 \; a_2 \; a_1a_2 \; a_1a_2^2 \; .... \; a_1a_2^{s-1} \; a_3 \; ...(a_1a_2^{s-1},...a_t^{s-1})$$

The exponents 0, 1, 2, ... s-1 are the elements of the Galois field $GF(p^n)$ with addition and multiplication defined within the field. If n=1 then the Galois field is the field of positive integers modulo p.

Let A_i represent any factor and let A_j represent one of the $(s^t-1)/(s-1)$ attributes considered in relation to A_i. Denote by t_{ij}, where t_{ij} is an element of $GF(p^n)$, the sum of the products of the exponents of a_1, a_2, ..., a_t in the expressions of the a representations of A_i and A_j. If the representations of A_i and A_j are

$$a_1^{i1}a_2^{i2}...a_t^{it} \text{ and } a_1^{j1}a_2^{j2}...a_t^{jt}$$

respectively, then $t_{ij} = \Sigma\, i_r j_r$. Choose as the elements corresponding with A_i the following treatment combinations

$(t_{i1}\ t_{i2} \ldots t_{iN})$, $(2t_{i1}\ 2t_{i2} \ldots 2t_{iN})$, ...
$((s-1)t_{i1}\ (s-1)t_{i2} \ldots (s-1)t_{iN})$,

where the representations of the levels of each attribute are elements of $GF(p^n)$.

Designs using arrays of strength d permit estimation of main effects and interactions of order up to k (d>k), assuming that the higher order interactions do not exist. Furthermore, the estimates of main effects and the desired interactions obtained from such designs are mutually independent.

APPENDIX B: THEOREMS ON BIAS AND INCREASED VARIANCE

Theorem 1. Given an RxK contrast-coded design matrix, X, Rx1 dependent variable vector, y_i (a respondent's rating of R cards), a Kx1 parameter vector, β_i (partworths), and an Rx1 vector of random errors, ε_i, such that:

(i) $y_i = X \beta_i + \varepsilon_i$,

(ii) β_i is independently distributed with mean β_o and positive definite covariance $\Sigma_{\beta\beta}$,

(iii) ε_i is independently distributed with mean 0 and variance $\sigma_i^2 I$,

(iv) $\text{cov}(\beta_i, \sigma_i) = 0$ for any i.

Let X_M be a version of X (with L rows and K_M columns) for a main-effects-only design with the minimum number of cards (L) needed to estimate only the main effects. Let

$$X_M^* = [\, X_M, X_{M(I)} \,],$$

i.e., the X_M matrix expanded to include the implied interaction terms. Partition the Kx1 vector β_o into a K_Mx1 main-effects vector β_M and a K_Ix1 interaction effects vector β_I:

$$\beta_o' = [\, \beta_M', \beta_I' \,],$$

$K_M+K_I = K$. Further, define the least squares main effect estimator for individual i as:

$$\hat{\beta}_{i(M)} = (X_M'X_M)^{-1}X_M'y_i$$

and define the estimate of the mean main effects across a sample of N individuals as:

$$\hat{\beta}_M = (1/N) \; \Sigma \, \beta_{i(M)}.$$

Then, the **bias in using $\hat{\beta}_M$ as an estimator of β_M** is given by

$$D\beta_I$$

where $D = (X_M'X_M)^{-1}X_M'X_{M(I)}$.

Proof. Since we have assumed a two-stage process, we find $E(\hat{\beta}_M)$ in two steps. First,

$$E(\hat{\beta}_M|\beta_i) = (1/N)\,\Sigma\, E(\hat{\beta}_{i(M)}|\beta_i, i=1, \ldots, N)$$

$$= (1/N)\,\Sigma\, E[(X_M'X_M)^{-1}X_M'(X_M^*\beta_i + \varepsilon_i)\,|\beta_i]$$

$$= (X_M'X_M)^{-1}X_M'X_M^*\,\Sigma\,\beta_i/N.$$

Then,

$$E(\hat{\beta}_M) = E\,(X_M'X_M)^{-1}X_M'X_M^*\,\Sigma\,\beta_i/N$$

$$= (X_M'X_M)^{-1}X_M'X_M^*\,\beta_o.$$

Since $X_M'X_M^* = [\,X_M'X_M,\ X_M'X_{M(I)}\,]$ and $X_M'X_{M(I)}$ is not generally equal to 0, we have

$$(X_M'X_M)^{-1}X_M'X_M^* = [\,I,\ D\,].$$

Therefore,

$$E(\hat{\beta}_M) = [I, D]\beta_o = \beta_M + D\beta_I$$

implying that the bias is

$$D\beta_I.$$

Theorem 2. Assume conditions (i)-(iv) of Theorem 1. Suppose further that we can partition $\Sigma_{\beta\beta}$ into main effect(M) and interaction effect(I) partitions:

$$\begin{bmatrix} \Sigma_{\beta\beta\,(MM)} & \Sigma_{\beta\beta\,(MI)} \\ \Sigma_{\beta\beta\,(IM)} & \Sigma_{\beta\beta\,(II)} \end{bmatrix}$$

Consider X_M and X_M^* defined as in Theorem 1. Let X_I be the JxK contrast-coded matrix for an interaction design (where J is the number of cards needed to estimate the interaction model; J>L generally) and partition it into K_M columns for main effects ($X_{I(M)}$) and K_I columns for interactions effects ($X_{I(I)}$) as was done for X_M^*:

$$X_I = [\,X_{I(M)},\ X_{I(I)}\,].$$

Define $\hat{\beta}_{i(M)}$ and $\hat{\beta}_{(M)}$ as in Theorem 1 so that:

$$\hat{\beta}_M = (1/N) \Sigma \hat{\beta}_{i(M)}.$$

In an analogous fashion for the interaction model, let the main effects estimates derived from the interaction model be:

$$\hat{\beta}_{il(M)} = (X_{l(M)}'X_{l(M)})^{-1} X_{l(M)}'y_i$$

and then define the mean across N individuals as:

$$\hat{\beta}_{l(M)} = (1/N) \Sigma \hat{\beta}_{il(M)}.$$

Then **the sampling variance of $\hat{\beta}_M$ as an estimator of β_M is greater than the sampling variance of $\hat{\beta}_{l(M)}$**, i.e. :

$$V(\hat{\beta}_M) > V(\hat{\beta}_{l(M)}).$$

Proof. We obtain $V(\hat{\beta})$ using the rule

$$V(\hat{\beta}) = V[E(\hat{\beta}|\beta_i, i=1,\ldots,N)] + E[V(\hat{\beta}|\beta_i, i=1,\ldots,N)]$$

from which we find

$$V(\hat{\beta}) = (1/N)[(X'X)^{-1}X'X^* \Sigma_{\beta\beta} X^{*'}X(X'X)^{-1} + \sigma^2(X'X)^{-1}]$$

where $\sigma^2 = (1/N) \Sigma\sigma_i^2$. As before, X is the contrast-coded main effects component and X^* is the matrix expanded to include interaction. From this, we obtain

$$V(\hat{\beta}_M) = (1/N)[(X_M'X_M)^{-1}X_M'X_M^* \Sigma_{\beta\beta} X_M^{*'}X_M(X_M'X_M)^{-1} + \sigma^2(X_M'X_M)^{-1}].$$

Upon letting $D = (X_M'X_M)^{-1}X_M'X_{M(I)}$, we obtain

$$(X_M'X_M)^{-1}X_M'X_M^{*}\,\Sigma_{\beta\beta}\,X_M^{*\prime}X_M(X_M'X_M)^{-1} = [\,I,\,D\,]\,\Sigma_{\beta\beta}\,[\,I,\,D\,]'.$$

Thus, with $D_M = X_M'X_M$,

$$\hat{V}(\beta_M) = (1/N)[\,\Sigma_{\beta\beta\,(MM)} + D\Sigma_{\beta\beta\,(IM)} + \Sigma_{\beta\beta\,(MI)}D' + D\Sigma_{\beta\beta\,(II)}D' + \sigma^2 D_M^{-1}].$$

Similarly, letting $D_{I(M)} = X_{I(M)}'X_{I(M)}$, we find

$$\hat{V}(\beta_{I(M)}) = (1/N)[(X_{I(M)}'X_{I(M)})^{-1}X_{I(M)}'X_I\Sigma_{\beta\beta}X_I'X_{I(M)}(X_{I(M)}'X_{I(M)})^{-1} + \sigma^2(X_{I(M)}'X_{I(M)})^{-1}].$$

$$= (1/N)[\Sigma_{\beta\beta(MM)} + \sigma^2 D_{I(M)}^{-1}]$$

since $X_{I(M)}'X_{I(I)} = 0$. Thus,

$$\hat{V}(\beta_M) - \hat{V}(\beta_{I(M)}) = (1/N)[D\Sigma_{\beta\beta\,(IM)} + \Sigma_{\beta\beta\,(MI)}D' + D\Sigma_{\beta\beta\,(II)}D' + \sigma^2(D_M^{-1} - D_{I(M)}^{-1})].$$

The sum of the first three terms inside the brackets is positive definite using standard results on the positive definiteness of the sum of a matrix plus its transpose and a quadratic form. $(D_M^{-1} - D_{I(M)}^{-1})$ is also positive definite since X_M has fewer rows than $X_{I(M)}$.

Note that in the proof of the preceding theorems we have not used a normality assumption. Given the stated assumptions, these results apply no matter how the partworths or errors are distributed.

APPENDIX C: PARTWORTHS AND IMPORTANCES FOR AN INTERACTION STUDY

Partworths are calculated for each respondent using ordinary least squares regression. A dummy-coded design matrix, D, with J rows (one for each card) and K columns (one for each level of each effect) is constructed with elements d_{jk}, where $d_{jk}=1$ if effect level k is present on card j and $d_{jk}=0$ otherwise. The dummy-coded design matrix D, which has rank L with L<K is decomposed into its first L principal components. An ordinary least squares regression of the ratings (y) on these principal components is performed. Partworths (p) are obtained by back transformation of the principal components regression coefficients. The total utility, z, is then estimated from z=Dp.

Importance indices for each respondent are computed in the following manner. Let Δ be an Nxr design matrix in orthogonal contrast form [Anderson and Houseman, 1942] where N is the number of possible profiles in the full factorial design and r is the number of unique effects parameters to be estimated and must be equal to L. Since Δ is orthogonal and its columns represent contrasts, we note that

$$(1/N)1'\Delta = 0.$$

Furthermore, $C=(1/N)\Delta'\Delta$ is a diagonal matrix with the r diagonal elements given by the mean of the squares of the contrasts.

For the cards in the conjoint study, we can construct an orthogonal contrast-coded design matrix X (where X contains a subset of the rows of the full factorial Δ). We regress the ratings, y, on X to obtain coefficients, u, from the regression model y=Xu+e. It can be shown that there is an exact transformation between X and D, so that Xu=Dp(=z, the total utility). By extension, we can use u with Δ to estimate total utilities for every possible card represented in Δ: $w=\Delta u$. In orthogonal contrast coding, Δ can be partitioned into sections corresponding to each effect k:

$$\Delta = [\Delta_1, \Delta_2, \ldots],$$

with $\Delta_k = [\Delta_{jkl}]$ and Δ_{jkl} being the contrast code for the l*th* parameter of the k*th* effect for card j. Correspondingly u can be similarly partitioned:

$$u = [u_k],$$

with $u_k = [u_{kl}]$, u_{kl} being the estimate of the l*th* parameter of attribute/interaction effect k for the respondent.

Therefore,

$$w = \Delta u = \Sigma \Delta_k u_k = \Sigma w_k,$$

where $w_k = \Delta_k u_k$. This implies that the variance of w is just the sum of variances of w_k (since Δ is orthogonal). The variance of w_k is

$$s_k^2 = \Sigma \; c_{(k)ll} u^{*2}_{kl},$$

where $c_{(k)ll}$ is the l*th* element of the diagonal of $(1/N)\Delta_k'\Delta_k$, which is the k*th* partition of C. Then the variance of w is $s^2 = \Sigma s_k^2$. Defining the respondent's importance for effect k (I_k) as the percent of the total variance in w accounted for by that effect, we have:

$$I_k = 100x(s^2_k/s^2).$$

REFERENCES

Addelman, S. (1962). "Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments," *Technometrics*, 4, 21.

Anderson, R. L. and E. E. Houseman (1942). "Tables of orthogonal polynomial values extended to N=104," Agricultural Experiment Station Research Bulletin 297. Iowa State College, Ames, Iowa.

Finkbeiner, C. T. (1988). Comparison of Conjoint Choice Simulators. *Proceedings of the Sawtooth Software Conference* . Sun Valley, Idaho.

Louviere, J. J. and G. G. Woodworth (1983). "Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data," *Journal of Marketing Research*, 20, 350.

Rao, C. R. (1946). " On hypercubes of strength d and a system of confounding in factorial experiments," *Bull. Cal. Math. Soc.*, 38, 67.

Comment on Finkbeiner and Lim

Joel C. Huber
Duke University

In their paper, Carl Finkbeiner and Pilar Lim provide a fine tutorial demonstrating what interactions can mean to conjoint and how a proper experimental design can resolve this problem. Particularly valuable is their demonstration of the distortion in partworth values and market share if one assumes that there are no interactions when they actually exist. On reading the paper, however, the conjoint user is put in a bind. While the cost of ignoring interactions may be high, so is the cost in design time and respondent effort of a study accounting for these interactions. Thus it is important to delineate those contexts in which one can expect interactions to be important and to specify a number of ways to deal with interactions, given they are believed to be a problem.

WHEN ARE INTERACTIONS A PROBLEM?

In a word, rarely. Interactions are a serious problem in conjoint where stimuli are holistic and thus inherently interactive, or where the respondents are familiar with the task and motivated to process the profiles in a configural way. Consider the case given in the Finkbeiner and Lim paper in which price sensitivity is higher for low-priced cars than high-priced ones. This interaction makes sense; a $2,000 difference should have greater impact on a $10,000 Fiesta than a $15,000 Taurus. However, registering pleasure at Taurus over Fiesta and displeasure over a higher price over a lower one is hard enough when one is evaluating, say, 32 profiles, each defined on 7 attributes. Responding interactively by making a stronger response to a price difference for the low- over the high-end car may require greater adaptability than respondents are willing or able to give. In my own work, I have included cross-terms to uncover interactions in a number of conjoint studies. I have been consistently disappointed by their insignificant magnitude even when there were good reasons to expect them to occur.

There are, however, two contexts in which interactions can be expected to be a problem. The first case is where the stimuli are presented as a whole or in pictorial terms. Interactions predominate for such stimuli and generally frustrate any attempt to decompose judgments into additive main effects. By contrast, main effects are quite adequate to characterize most verbally described stimuli. Verbal descriptors tend to lead respondents to respond in an additive fashion simply because it gets them through the task.

Even verbal descriptors can elicit interactions if respondents have sufficient familiarity with the product class so that interactions are easy to process. In the automobile example such configurality would occur if respondents knew automotive prices so well that they could immediately recognize a $2,000 difference as large for a Fiesta, but small for a Taurus. The interactions found in the drug study reported by Finkbeiner and Lim are consistent with this requirement, since it is likely that the "health professionals" were well acquainted with the profiles being tested. Note, however, that familiarity with the general class of products is not sufficient — respondents must be able to quickly respond to the actual profiles presented. Suppose, in the automobile example, that prices were twice their normal level, and that a $14,000 price attached to a Taurus has immediate meaning, while a $28,000 price does not. Faced with unfamiliar attribute levels, even an expert is likely to have difficulty responding configurally, thereby suppressing interactions.

In addition to needing expertise in the domain of the attribute levels tested, respondents need to be motivated to respond configurally. Thus, respondents who are tired or insufficiently compensated for the conjoint task may be less likely to produce interactions. Motivational considerations also suggest that making the test too difficult, either with too many profiles or too many attributes within each, may lead the respondent to revert to non-interactive responses as a way to get through the task.

In summary, interactions are problematic less because they bias the main effects, but because they are so difficult to tease out of respondents. To the extent that we believe that interactions are important in the market choice process simulated with conjoint, we need to develop ways to estimate interactions without taxing respondents too greatly.

HOW TO DEAL WITH INTERACTIONS

Given one desires to account for them, interactions can be estimated either within or across respondents. Both strategies have their advantages. For interactions to be uncovered across respondents, they must exist homogeneously within a defined segment or group. While it may be true that the only managerial useful interactions are those that dependably occur within a segment, having to estimate these interactions at that level limits the ability to simulate choice at an individual level.

There are a number of strategies for cross-respondent estimation of interactions. One method gives each respondent a different main effect design so that interactions are estimable within relatively small groups. For example if 9 profiles are used to estimate the main effects of 4 attributes each at three levels, then these fractional factorials can be arrayed in 9 blocks so that all possible interactions can be estimated at the group level. A second process that is less elegant but somewhat more flexible is to randomize the assignments of levels for each respondent. That is, for any fractional factorial the assignment attribute levels to the particular design levels is arbitrary. For example, the main effects design is just as efficient regardless of whether a level 1 is given to 20 MPG and level 2 to 30 miles per gallon, or the reverse. However, the confounding of higher-order interactions with

main effects and with each other depends on the assignment used. Thus, by randomizing the assignment of levels to attributes for each respondent, these biasing effects tend to cancel across respondents, making it possible to estimate all possible interactions for any group of sufficient size (say 20-30 respondents). This process requires coding to keep track of the particular assignment for each respondent, something that would have been infeasible 15 years ago but is relatively easy with computers today.

Analysis of such pooled data is tedious but straightforward. Think of a large regression which combines data for all respondents from a segment into one large set. Each respondent is allowed to have individually-determined main effects while sharing the interactions terms with the rest of the segment. One is thus able to make different predictions at the individual level. Note that the critical assumption of this aggregate interactive model is that the interactions are consistent within the segment analyzed.

The other way to deal with interactions is to estimate them at the individual level. This strategy is preferable, where possible, as it preserves the integrity of the individual preference model that is so central to conjoint simulators. If only a few factors are expected to interact then sometimes it is possible to estimate these directly by forming a super-variable. For example, the interaction of three price levels and two quality levels can be estimated by including a 6-level price x quality attribute. That strategy is often used in programs such as ACA, but may be less useful for full-profile designs where a 6-level attribute can drastically complicate finding a main-effects design with a reasonable number of profiles. Further, the solution requires the unlikely assumption that only the two crossed attributes interact, and that the other attribute pairs do not interact at all.

What is particularly appealing about the design that Finkbeiner and Lim present in their paper is that it can estimate the price interaction with each of the other factors. However, this solution is not for everyone. To my knowledge there do not exist catalogues of such designs, and Finkbeiner and Lim had to generate theirs through an arduous process involving Galois fields. Further, like their main-effects brethren, these designs only exist for limited combinations of levels. It is no accident that the 5 factors crossed with cost are defined at only two levels. Increasing just one to three levels would have required more than 32 the profiles used. Thus, even using great cleverness, there are limits to the degree to which one can account for interactions on a within-respondent basis.

Perhaps the best compromise is to design studies so that certain interactions can be estimated within respondents, but to vary across respondents those attribute pairs whose interactions are estimable. This strategy permits one to test whether interactions matter at both a group and an individual level. The finding of strong but heterogeneous interactions signals that both the simple main effects models and the aggregate interactive model are seriously flawed. This may not be good news to hear, but failing to design studies which can detect interaction is the survey research equivalent to shooting the messenger before he gets a chance to speak.

RECOMMENDATIONS

This comment began with the statement that interactions occur rarely in conjoint exercises. While that statement still holds, it is important to echo Finkbeiner and Lim's warning of how dangerous it is to use main effects models in the face of interactions. Not only will be the results be biased, but the analyst will never know the bias. Main effects models are particular risky when the same experimental design is used across respondents. In that case, the bias takes the same pattern across respondents: a pattern that appears to reveal a consistent relationship when in fact there is only a consistent bias.

Thus, it is important that the conjoint user be constantly vigilant for possible interactions. These interactions occur when profiles are evaluated as a whole, either because they are presented pictorially or because the respondent has the experience and motivation to process them configurally. However, since a main effects design cannot be used to evaluate its own critical assumption, the wise conjoint user constantly tests against interactions. The easiest of these tests occurs across subjects, where respondents receive different fractional factorials which, when aggregated, can reveal interactions. A second group of tests are within respondents. These tests should take the form of regularly checking a couple of interactions whenever a conjoint is run. They also involve more complex designs such as those presented by Finkbeiner and Lim.

Most users will probably find that the interactions are inconsequential. When that happens, breathe a sign of relief. However, it is important to remain watchful, because where interactions are concerned, the real risk is assuming them away and never knowing that your results are chronically biased. Where interactions are concerned, it's what you don't know that can hurt you, rather than the reverse.

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

Since the introduction of Sawtooth Software's Adaptive Conjoint Analysis (ACA) in the mid-eighties, its reception by both academics (Carmone 1987; Huber and Hansen 1986) and industry practitioners (Cerro 1988; Stahl 1988; Zimmermann 1988) has been very favorable. ACA is a computer-based system (Johnson 1987) for collecting conjoint data that uses a type of hybrid model (Green 1984) for updating self-explicated partworths with information obtained from graded paired comparisons of partial profiles. The profiles consist of two to five attributes each. The principal novelty of ACA is that the program customizes the paired comparisons that a respondent sees, based on partworths estimated from preceding stimulus evaluations.

Several validation studies of ACA have been implemented (e.g., Finkbeiner and Platz 1986; Agarwal 1987; Agarwal 1988). These early studies showed roughly equivalent validation (over holdout profiles) for ACA versus traditional, full-profile conjoint.

Recently, however, researchers have raised questions about some of ACA's features. Agarwal (1989), for example, has reported a validation study in which the addition of the paired comparisons step resulted in predictive validities that were not significantly higher than those found from the self-explicated data alone. In a separate study, Agarwal and Green (1989) reported that the paired comparisons data provided some signal beyond the self-explicated data. Still, the traditional, full-profile conjoint model cross-validated better than ACA, as did a simple self-explicated model that used finer scale gradations than are possible in the current version of ACA.

In a more methodologically-oriented paper, Green, Krieger, and Agarwal (1991), hereafter referred to as GKA, have discussed some concerns about ACA and have offered suggestions for improving its versatility and validity. In particular, GKA question the *ad hoc* way in which ACA links the respondent's self-explicated responses with his or her paired comparison responses. GKA ran additional analyses of the Agarwal and Green data to demonstrate various methodological points in their discussion. However, see Johnson (1991) for a rejoinder to conclusions rendered in both Agarwal and Green and GKA.

Like many of the ACA cross-validated experiments, the GKA analysis was based on business student subjects. The stimuli consisted of student apartment descriptions that varied on six attributes (each at three levels): public transport time to campus; noise level; safety; apartment condition; size of living/dining area; and monthly rent.

Other aspects of the GKA study, relevant to the present paper, should be noted:

1. Given the monotonic nature of the apartment attributes (including lower rent preferred to higher rent, greater apartment safety preferred to less safety, and so on), the GKA task was probably less difficult than would be the case if non-monotonic attributes were considered.
2. The student subjects received fifteen graded paired comparisons, each comprised of only two attributes per partial concept.
3. The sets of holdout profiles showed relatively wide separation in utilities, leading to fairly easy profiles to evaluate.

All in all, the question arises as to whether some of the directional findings of GKA's study could be replicated in a new validation study that involved different types of subjects and more difficult evaluation tasks.

THE CURRENT STUDY

The current study entailed a sample of 96 industry practitioners of conjoint analysis. Respondents were drawn from marketing research firms and industrial firms. All respondents had some knowledge of conjoint analysis (although not necessarily of ACA).

The stimulus set consisted of eight attributes (at four levels each) of automobile profiles. Table 1 shows the attribute and level descriptions. As noted from the table, with the exception of attributes C (fuel economy) and D (base price), the attributes are non-monotonic. That is, we might well expect individual differences in the preference orderings of body style, country of manufacture, and exterior color. Also, with the implied price tradeoffs, the extra-cost options of engine type, radio, and warranty could easily lead to individual variation in preference orderings.

Evaluation Tasks

Each subject was given an ACA computerized questionnaire, which consisted of four phases:

1. In Phase I each subject was asked to order the levels of each attribute, in turn, in terms of preference, other things equal.
2. In Phase II each subject was asked to rate each of the attributes on a four-point importance scale, where "4" denotes highest importance.
3. In Phase III each subject was presented with twenty ACA-generated paired (partial) concepts. In the first ten of the twenty pairs, each concept in the pair was described on two attributes. In the second ten pairs, each concept was described on three attributes. For each pair the subject could select "left concept strongly preferred to right" (coded 1) up to "right concept strongly preferred to left" (coded 9). Indifference between the two concepts received a middle code value of 5; in short, a 9-point equal interval scale was used.

4. In Phase IV, the calibration stage, each subject received a very poor full profile followed by a very good profile, each based on the subject's already-computed partworths. Subjects rated each full profile on a 0-100 likelihood-of-purchase scale. (Since the calibration profile responses are used only to rescale the original partworths, they are not analyzed further.)

Holdout Profiles

Following the ACA exercise, each subject received a set of twelve randomized full profile cards. This design was based on a standard Plackett-Burman orthogonal plan in which each attribute was set at two levels. Table 1 shows the levels that were picked for the holdout sample design. Table 2 shows the specific orthogonal design levels. As can be noted from the starred levels of Table 1, we tried to choose levels that would be presumed to be relatively close in utility; this made the overall evaluation task reasonably challenging to the subjects. For each holdout profile, the subject was asked to rate it on a 0-100 likelihood-of-purchase scale, if he or she were in the market for a new car.

The present study thus differs from the GKA study in the following respects:

1. Eight (rather than six) attributes at four (rather than three) levels each.
2. Six out of eight attributes being non-monotonic versus none for the GKA study.
3. Twenty paired comparisons (half with two attributes and half with three attributes per concept) versus only fifteen paired comparisons (all with two attributes) for the GKA study.
4. Marketing research suppliers and industry researchers versus business student subjects.

It should be mentioned, however, that the scope of the present validation study is limited. Primarily, we wished to see if the paired comparisons section added any appreciable signal to the self-explicated exercise and, if so, the comparative extent of its contribution, relative to that found in the GKA study.

In the GKA paper, the authors considered comparisons of not only the full (updated) ACA model with the self-explicated model, but also comparisons involving only the first 8 paired comparisons. In a similar spirit, in the current study, we later present comparative results for 0 pairs (self-explicated only), 7 pairs, 14 pairs, and all 20 pairs.

STUDY RESULTS

In the GKA study, the authors found that subjects tended to expand the Phase III paired comparisons response scale (so as to use its full 1-9 point range), relative to scale values that would be predicted by the ACA model.

Table 3 shows comparative results for the two studies. As GKA observed, there is a fairly high incidence of incorrect predictions, involving even preference direction; 27 percent of the time, subjects' partworths failed to predict the direction of response (that is, left preferred to right or vice versa). In the current study, we find that the incidence is higher, at 34 percent.

More importantly, given that a subject's partworths predicted the correct direction of concept preference, GKA found that almost three quarters of the time the actual response was more extreme than that predicted. Subjects tended to use the whole range of the 1-9 point scale. In so doing, their responses were biased. The observed departure from the null case of 50 percent was highly significant ($p < 0.001$). The same is true in the current study. In fact, the effect is even more extreme: approximately 86% of the time subjects report more extreme scale values than would be predicted by the ACA model.

So far, then, the current study results support those of the GKA study. If anything, the extent of the unreliability/bias is even higher in the current study.

How Well Do the ACA Models Predict Each of the Paired Comparisons? We next considered the question of actual prediction error (as opposed to directional bias). Table 4 shows these results. As noted from the table, GKA found that the mean absolute errors of prediction were approximately the same when the self-explicated model was used to predict all 15 paired comparisons versus the case when each paired comparison was predicted by all the information available up to that point (including the subject's responses to preceding pairs). The current study supports the finding that mean absolute errors are not significantly affected by the cumulative information, entailing all 20 paired comparisons.

In contrast, if one is allowed to find a best-fitting linear transformation of the paired comparison responses (as implied by a correlation model), the correlations differ significantly across models. In the GKA study the correlation, involving the fully updated model, was significantly higher than predictions obtained from the self-explicated data alone ($p < 0.05$); see Table 4.

Similar findings are noted in the current study. In general, as the amount of cumulative information increases, the predictions get better (as measured by correlations). This result is also significant ($P < 0.05$). Hence, we have further evidence in support of GKA's claim that ACA does not contain a satisfactory procedure for making the Phase III paired comparisons raw response scale commensurate with the Phases I/II self-explicated partworths. Apparently, an additional transformation (for example, linear) is needed to bring these two data sources into satisfactory agreement.

How Well Do the ACA Models Predict the Holdout Profile Evaluations? While Table 4 suggests that the paired comparison responses of both studies contain some signal, the main point of the exercise is to see how well each model cross-validates with the holdout profile responses. Results are shown in Table 5 for three commonly-used validation measures: correlations, first-choice hits, and rank position hits.

The GKA study indicated that the cross-validated correlations generally increased with the additional information provided by the paired comparisons. However, the correlations obtained with all 15 pairs were not significantly higher than those obtained from the first 8 pairs. In the current study the gains are also quite modest (about three correlation points), although their sample trend is in the anticipated direction.

Insofar as first-choice hits are concerned, the effects were also modest (less than four percentage points) for 15 versus 0 pairs in the GKA study. The current study shows somewhat better performance for 20 versus 0 pairs and 14 versus 0 pairs, but there is essentially no difference between the 14 and 20 pair cases. In all cases the cross-validations are lower here than those found by GKA.

Rank position hits in the GKA study showed very little improvement with additional pairs — about one percentage point in the most extreme case. In the current study the most extreme case (20 versus 0 pairs) shows only a 1.6 percentage point gain over the self-explicated data alone.

Hence, while the paired comparisons add something over the self-explicated model, the gains are modest, particularly in the case of rank position hits. As can be noted, all predictions in the current study are substantially lower than their GKA counterparts; however, directional correspondence is similar between the two studies.

CONCLUSIONS

Considering the two studies together, we can make the following (tentative) conclusions:

1. There is a consistent tendency for subjects to "stretch" their paired comparisons judgments to "fill up" the whole 1-9 point scale. In both studies the tendency was highly significant. In the more difficult tasks associated with the current study, this biasing tendency reached an incidence of 86 percent. As GKA (1991) have pointed out, paired comparisons with three or more attributes tend to result in closer utilities between concepts, as well as making the evaluation task more difficult.
2. As Table 4 shows, the paired comparisons data provide some signal beyond the self-explicated data, assuming that correlation (rather than mean absolute error) is used to measure agreement between actual and predicted responses.
3. In terms of cross-validation (Table 5), the pairs provide modest improvements over the self-explicated model. In terms of correlation, the current study shows a maximal gain (in the 20 pairs case) of three correlation points. First-choice hits show a maximal improvement of about seven percentage points. For rank position hits the improvement is very small — less than two percentage points.

In conclusion, both studies show that gains from the paired comparisons section (the mainstay of the ACA methodology) are less than impressive. These findings, coupled with Agarwal's empirical findings (1989) of no improvement at all, suggest that applied researchers might want to consider more carefully the gains versus costs of adding the paired comparisons section in applied ACA studies. (We now understand that the paired comparisons section is optional in version 3.0 of ACA; this new version was introduced in late 1989. (Version 2.0 was used in the GKA and current studies.)) In particular, users of ACA might wish to consider conducting pretests of the survey

(with a holdout set of profiles) to get feedback on the amount of additional information provided by the paired comparisons section. Pretests could also provide guidance on the number of paired comparisons to include.

The authors would like to acknowledge the support of the Citibank Fellowship (received from the Sol C. Snider Entrepreneurial Center) and the SEI Center for Advanced Studies in Management, at the Wharton School.

REFERENCES

Agarwal, Manoj K. (1989), "How Many Pairs Should We Use in Adaptive Conjoint Analysis?," *1989 Winter Educators' Conference Proceedings.* American Marketing Association, 7-11.

_____ (1988), "Comparison of Conjoint Methods," *Sawtooth Software Conference Proceedings.* Ketchum, ID: Sawtooth Software, 51-57.

_____ (1987), "An Empirical Comparison of Traditional Conjoint and Adaptive Conjoint Analysis," Working Paper, State University of New York at Binghamton.

_____ and Paul E. Green (1989), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results," *International Journal of Research in Marketing* (1991), 8, in press.

Carmone, Frank J. (1987), "Review: Adaptive Conjoint Analysis," *Journal of Marketing Research*, 24 (August), 325-327.

Cerro, Dan (1988), "Conjoint Analysis by Mail," *Sawtooth Software Conference Proceedings.* Ketchum, ID: Sawtooth Software, 139-144.

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

Green, Paul E. (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, 21 May, 155-159.

_____, Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28 (May), forthcoming.

Huber, Joel and David Hansen (1986), "Testing the Impact of Dimensional Complexity and Affective Differences of Paired Concepts in Adaptive Conjoint Analysis," in *Advances in Consumer Research*, eds. M. Wallendorf and P. Anderson, Provo, UT: Association for Consumer Research, 14, 159-163.

Johnson, Richard M. (1987), "Adaptive Conjoint Analysis," Sawtooth *Software Conference Proceedings*. Ketchum, ID: Sawtooth Software, 253-265.

_____, (1991), "Rejoinder to Green, Krieger and Agarwal," *Journal of Marketing Research*, 20 (May), forthcoming.

Stahl, Brent (1988), "Conjoint Analysis by Telephone," *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software, 131-138.

Zimmermann, Robert (1988), "Complex Computer Interviews," *Sawtooth Software Conference Proceedings*. Ketchum, ID: Sawtooth Software, 145-150.

TABLE 1
Attributes and Levels Used in ACA Study

A. Body Style 1. Sedan* 2. Station wagon* 3. Van 4. Convertible	E. Engine Type 1. 4-Cylinder 2. 4-Cylinder Turbo + $500* 3. 6-Cylinder + $600* 4. 6-Cylinder + $1000
B. Country of Manufacture 1. U.S.-made* 2. German-made 3. Japanese-made* 4. Swedish-made	F. Exterior Color 1. Maroon* 2. Silver 3. White 4. Light blue*
C. Fuel Economy 1. 20 MPG average* 2. 25 MPG average 3. 30 MPG average* 4. 35 MPG average	G. Radio 1. AM only 2. AM/FM + $100* 3. AM/FM stereo + $250 4. AM/FM stereo, tape deck + $350*
D. Base Price 1. $10,000 2. $13,000* 3. $16,000* 4. $19,000	H. Warranty 1. 1-year* 2. 2-year + $150 3. 4-year + $350* 4. 6-year + $600

* Levels picked for the holdout sample design.

TABLE 2
Orthogonal Design Matrix Used to Construct Holdout Profiles

Card #	A	B	C	D	E	F	G	H
1	1	1	1	2	2	1	2	1
2	2	3	1	3	3	4	2	1
3	1	3	3	2	3	4	4	1
4	2	1	3	3	2	4	4	3
5	1	3	1	3	3	1	4	3
6	1	1	3	2	3	4	2	3
7	1	1	1	3	2	4	4	1
8	2	1	1	2	3	1	4	3
9	2	3	1	2	2	4	2	3
10	2	3	3	2	2	1	4	1
11	1	3	3	3	2	1	2	3
12	2	1	3	3	3	1	2	1

TABLE 3
Summary of Subjects' Evaluations of the ACA Pairs

	GKA Study (N=170)	Current Study (N=96)		
		7 Pairs	14 Pairs	20 Pairs
Fraction with Wrong Direction*	.268	.342	.345	.350
Standard Error	.009	.008	.010	.009
Fraction In Which Actual Response Is More Extreme Than Predicted	.744	.861	.873	.856
Standard Error	.012	.011	.018	.016

TABLE 4
Performance of Models in Predicting Phase III Responses; Correlations and Mean Absolute Errors

	GKA Study (N=170)		Current Study (N=96)	
	Self-Explicated Alone	All 15 Pairs (Updated)	Self-Explicated Alone	All 20 Pairs (Updated)
Mean Correlations	.335	.524	.200	.398
Standard Error	.012	.015	.013	.020
Mean Absolute Errors	1.808	1.865	1.930	1.894
Standard Error	.037	.041	.047	.051

TABLE 5
Cross-Validation Performances of Various Prediction Models on Holdout Profile Sets

	GKA Study* (N=170)			Current Study** (N=96)			
	0 Pairs	8 Pairs	15 Pairs	0 Pairs	7 Pairs	14 Pairs	20 Pairs
Correlations							
Mean	.543	.600	.618	.229	.236	.245	.255
Standard Error	.029	.019	.014	.019	.018	.018	.017
First Choice Hits (%)							
Mean	58.4	60.4	62.0	25.5	27.4	32.4	32.3
Standard Error	3.780	3.751	3.723	4.448	4.552	4.777	4.769
Rank Position Hits (%)							
Mean	16.0	16.7	17.1	14.3	15.1	15.5	15.9
Standard Error	2.812	2.861	2.905	3.573	3.654	3.694	3.732

* Based on 16 holdout profiles

** Based on 12 holdout profiles

Comment on Green, Schaffer, and Patterson

Dick R. Wittink
Johnson Graduate School of Management
Cornell University

INTRODUCTION

When conjoint measurement was introduced into the marketing literature, the technique was proposed as a numerical method for decomposing overall preferences for objects into part utilities for the objects' characteristics (Green and Rao 1971). In its original form, objects could be defined on two or more attributes, and a respondent's preference rank order for the objects was the basis for obtaining numerical values. In marketing, researchers' attention quickly focused on efficient methods (designs) for the construction of the objects to obtain a maximum amount of information for the quantification of an assumed preference structure for respondents.

Johnson (1974) introduced the tradeoff matrix approach. In this method respondents were asked to provide a preference rank order of all objects defined on two attributes at a time. Subsequently, Green and Srinivasan (1978) in a review of the collection of decompositional preference measurement approaches introduced the term conjoint analysis. This was done to emphasize the fact that in marketing and elsewhere applications involved numerical estimation of parameters for a prespecified preference model. Conjoint measurement, on the other hand, concerned itself more with the existence of measurement scales and the extent to which preference judgments satisfy the conditions required for alternative composition rules.

Because alternative methods became available for data collection and analysis, simulation studies and empirical comparisons were completed. Green and Srinivasan (1978) provide a review of some studies and discuss advantages and disadvantages of full profiles relative to tradeoff matrices. Cattin and Wittink (1982) provide information on the commercial use of alternative data collection methods. In an update on the commercial applications, Wittink and Cattin (1989) suggest that the tradeoff matrix approach is used less frequently. This decrease may be due to the finding that respondents sometimes provide patterned responses but also to the fact that ACA (ACA System for Adaptive Conjoint Analysis) is a replacement of the tradeoff matrix approach (MacBride and Johnson 1979).

The connection between ACA and tradeoff matrices can be seen by considering the fact that a preference ordering of the six cells in a (3 x 2) matrix implies fifteen paired-comparison preference judgments. And paired comparison judgments can be used to infer a preference order for the cells in a tradeoff matrix. Interestingly, however, as shown in Currim *et al.* (1981), the entire rank order for the six cells can be inferred from the preference judgments for three specific paired comparisons, if the preference orders of the attributes' levels are known. And if the process were computer interactive one paired comparison might be sufficient!

Another potential difficulty with the tradeoff matrix approach is that in a study with many attributes, respondents have a number of matrices to consider that are defined on two unimportant or almost irrelevant attributes. If one knows in advance which attributes are unimportant to all respondents, one could eliminate those attributes from the study. Chances are, however, that respondents disagree substantially about the relative importance of individual attributes. Therefore, by using a computer-interactive approach one can elicit information on the relative importance of all attributes from each respondent, and concentrate on those in subsequent data collection that are especially relevant to the respondent in question. Thus, a computer-interactive approach allows the data collection task to be individual-specific, which increases efficiency and should enhance respondent interest in the task. Naturally, the advantages of such a computer-interactive approach increase as the number of attributes included in a study goes up and respondent heterogeneity in the role of the attributes becomes greater.

ISSUES UNIQUE TO ACA

In ACA self-explicated data are used to obtain an initial set of partworths. These initial partworths are defined such that the difference between the values for the best and worst levels of an attribute corresponds to the stated importance of the attribute. Intermediate attribute levels obtain values that correspond to the preference order for the levels. For example, if the attribute warranty has three levels and its importance equals 2 on a 4-point scale, the initial partworths are +1, 0 and -1, respectively for the best, intermediate and worst levels. Johnson (1987) provides a detailed description of ACA.

One may argue that the paired comparison intensity judgments should be used 1) to differentially stretch the attributes' importances, and 2) to modify the assumption of equal utility distances between successive attribute levels. However, there is nothing in the estimation procedure that prevents the updated partworths from violating the assumed or specified preference order for the levels. Especially for attributes with known preference orders for the levels, it is very likely that imposing order constraints on the final partworths (Srinivasan *et al.* 1983) will improve the results. Green *et al.* (1991) in their discussion of ACA focus primarily on the extent to which the paired comparison intensity values are commensurate with the initial (or previously updated) partworths. They suggest that ACA results can be improved by differentially weighting the self-explicated data and the paired comparison preferences. Johnson (1991) notes that the nature of differential weights for optimal improvement in predictive validity varies between applications. Green *et al.* also mention that more precise measurement scales for the self-explicated data may further enhance the performance of ACA.

EMPIRICAL STUDIES

MacBride and Johnson (1979) compared a computer-interactive approach to tradeoff matrices. The computer-interactive approach used is a forerunner of ACA. The study, involving 38 attributes, was completed by the John Morton Company. Data were collected in stages, such that for all respondents the conjoint task was preceded by an importance ranking of the attributes. Half the respondents (n_1 = 39) were then given a respondent-specific set of matrices involving 10 attributes to be completed by a paper-and-pencil task. The other half (n_2 = 39) completed a computer-interactive interview involving paired comparison preference intensity judgments. As in ACA the pairs of objects selected were based on previous judgments provided by the same respondent. The holdout (validation) task consisted of a ranking of five product concepts. The results showed a 61% first choice hit rate for the electronic versus a 51% hit rate for the paper-and-pencil approach. The difference was not statistically significant. Other measures suggested a higher perceived ease of the task, higher degree of interest in the task, lower amount of time used and higher level of task involvement for the electronic approach. Thus, the results indicate considerable potential for the computer-interactive approach. Of course, the difference in performance between the electronic and paper-and-pencil approaches might be greater if the tradeoff matrices had been defined on the same (sub)set of attributes for all respondents, as is traditionally done.

Finkbeiner and Platz (1986) compared ACA with the full profile approach. They used checking accounts with six attributes. All attributes were monotone; thus, the opportunity for respondent heterogeneity was severely limited in this study. About half the subjects in a convenience sample completed a paper-and-pencil full profile sorting of sixteen cards. The other half performed the ACA task. The results indicate that ACA took longer ($p < .05$) but was more interesting ($p < .05$), on average. ACA was also rated as somewhat more difficult, but this evaluation represented a combination of the difficulty of the task and the difficulty of the questions. Another problem in the study is that two attributes had to be combined in the ACA task. This difference in the two data collection methods may have contributed substantially to differences in average partworths. The average attribute importances suggest that the full profile method results are much further from a uniform distribution (for example, the average deviation from uniformly distributed importances is twice as high for full profile as for ACA).

To assess predictive validities, Finkbeiner and Platz obtained desirability ratings for four profiles as well as choices from a set of four and from a set of six objects. Importantly, in the validation task all objects were presented in terms of paragraph descriptions. The use of paragraph descriptions makes it less likely that the validation task favors one method over another. The desirability ratings were better predicted, using (squared) correlations, in the full profile than in the ACA task. For the choice data, the first-choice shares (aggregate predictions) showed comparable predictive results for the two methods. However, ACA with 8 pairs appeared to perform substantially below ACA with 16 pairs. Similarly, individual predictions showed comparable results: full profile was better than ACA in the four-item and ACA was better than full profile in the six-item choice set. Neither of these differences was statistically significant.

Agarwal and Green (1989) compared ACA with a self-explicated approach, using holdout profiles. Undergraduate business students provided data on apartments based on six monotone attributes. However, unlike Finkbeiner and Platz, Agarwal and Green used a within-subject design. One advantage of within-subject comparisons is that the power of statistical tests is enhanced. On the other hand, there may be systematic order effects due to the sequence of the tasks. Agarwal and Green collected ACA data, followed by self-explicated data and two sets of full-profile evaluations. Differences in performance between alternative methods on holdout tasks may, therefore, be due to differences in the methods themselves and to the order in which data for the methods were collected. Johnson (1991) expresses concern about the inability to separate these effects in Agarwal and Green.

The results of the study are that the weighted self-explicated model outperforms ACA using product-moment correlations as well as percent hits. However, the authors do not provide statistical significance for the observed differences. This, combined with the possible task order effect in the design of the empirical study, makes it impossible to provide conclusions about the relative performance of ACA versus self-explicated data, even if one only considers an application context with six monotone attributes.

Huber *et al.* (1991) compared the predictive validity of ACA and full-profile data on holdout choices. Respondents were asked to complete both full profile evaluations and ACA responses. However, to avoid systematic effects due to the order of administering data collection tasks, half the respondents did full profiles before ACA and the other half did full profiles after ACA. Also, the number of attributes used was five (four monotone) for half the respondents and nine (seven monotone) for the remaining respondents.

They found higher overall predictive validity for ACA than for full profile. This difference occurred for the five- as well as the nine-attribute tasks. The order of administering the tasks made no difference for ACA. However, for full profiles the predictive validity was much lower when it was the first task (relative to it being the second task and relative to ACA being the first task).

The Green, Schaffer, and Patterson Study

Finally, the paper by Green, Schaffer and Patterson (1990) is a further study of the commensurability of the self-explicated data and paired comparison preference judgments in ACA. Green, Schaffer and Patterson (GSP) mention that Agarwal and Green (1989) show better cross-validation for full profiles than for ACA. However, this conclusion is unwarranted given 1) the possible systematic effect on holdout task results due to the order of the task, and 2) the use of a (similar) full profile task on which ACA and full profile results are compared.

GSP note that the use of monotone attributes is a restrictive aspect of the Agarwal and Green study. They use a sample of industry practitioners of conjoint analysis to obtain data on automobiles with eight attributes, each defined on four levels. All but two of the attributes are non-monotone. Thus, there is greater opportunity for respondent heterogeneity in the quantification of the assumed

preference structures. The ACA task was followed for each subject by a purchase likelihood evaluation of twelve full profiles. The profiles were chosen to be "relatively close" in overall utility. The primary objective of GSP is to see how the paired comparison data add to the quality of the holdout task predictions. GSP use twenty paired comparisons, half defined on two, the remainder on three attributes. Holdout predictions are determined separately for the use of zero, seven, fourteen, and all twenty pairs in the computation of ACA partworths.

GSP observe that the direction of paired comparison preferences is not predicted by the partworths 34 percent of the time. One interpretation of this is that there is an inconsistency between the self-explicated and the paired-comparison data. Another interpretation is that perfect consistency would imply that there is little incremental value in the paired comparison judgments.

To understand the occurrence of inconsistencies between initial (updated) partworths and paired comparison judgments, I use an example with two attributes, A and B, each having three levels. Suppose that $u(A_1) \geq u(A_2) \geq u(A_3)$ and $u(B_1) \geq u(B_2) \geq u(B_3)$, where u stands for utility and $\geq$ means at least equal to. If the self-explicated importances of A and B are 1 and 4, the inference is that $[u(A_1) - u(A_3)] = 1$ and $[u(B_1) - u(B_3)] = 4$. In ACA, this results in the following initial partworths:

Attribute Level	Initial Partworth
A_1	0.5
A_2	0
A_3	-0.5
B_1	2.0
B_2	0
B_3	-2.0

If these partworths correctly reflect the respondent's preferences, then any subsequent paired comparison judgments are redundant. This should occur under the following conditions: 1) the true difference in utility between B_1 and B_3 is exactly four times as important as the difference between A_1 and A_3, and 2) the true differences in utility between successive levels of each attribute are exactly equal. The paired comparison judgments should then be consistent with the preferences predicted from the (initial) partworths. For example, the prediction is that $u(A_3 B_1) > u(A_1 B_2)$, where $A_i B_j$ is an object with levels i for A and j for B. In general, consistency between the actual direction of paired comparison preferences and the predicted direction should occur if the attributes have true pairwise importance ratio's belonging to the set [1.0, 1.3, 1.5, 2.0, 3.0, 4.0] and have equal utility differences between all successive levels. In addition, both the self-explicated and the paired comparison judgments have to be perfectly valid. In practice, any of these conditions may be violated.

To see how "inconsistencies" can occur, suppose that the same respondent's true partworths are the following:

Attribute Level	True Partworth	Initial Partworth
A_1	0.8	0.5
A_2	-0.3	0
A_3	-0.5	-0.5
B_1	1.8	2.0
B_2	1.3	0
B_3	-3.1	-2.0

Based on the true partworths shown, it is reasonable to assume that the respondent provides self-explicated importances of 1 for A and 4 for B. Thus, the initial partworths are the same as before. If ACA now asks the respondent to evaluate profile $A_3\ B_1$ versus profile $A_1\ B_2$, the predicted preference order is $u\ (A_3\ B_1) > u\ (A_1\ B_2)$, since $1.5 > 0.5$, but the true preference order is $u\ (A_3\ B_1) < u\ (A_1\ B_2)$, since $1.3 < 2.1$. This example shows that a respondent who expresses a preference that directionally differs from the predicted preference may have a partworth structure that could not be fully captured by the self-explicated data.

To be more precise, consider all possible pairs that require a respondent to trade attributes off against each other. The following nine satisfy this requirement. I also show the true and predicted difference in utilities for these nine pairs:

Pair	True Difference	Initial Predicted Difference
$u\ (A_1\ B_2)$ - $u\ (A_2\ B_1)$	0.6	-1.5
$u\ (A_1\ B_2)$ - $u\ (A_3\ B_1)$	0.8	-1.0
$u\ (A_1\ B_3)$ - $u\ (A_2\ B_1)$	-3.8	-3.5
$u\ (A_1\ B_3)$ - $u\ (A_2\ B_2)$	-3.3	-1.5
$u\ (A_1\ B_3)$ - $u\ (A_3\ B_1)$	-3.6	-3.0
$u\ (A_1\ B_3)$ - $u\ (A_3\ B_2)$	-3.1	-1.0
$u\ (A_2\ B_2)$ - $u\ (A_3\ B_1)$	-0.3	-1.5
$u\ (A_2\ B_3)$ - $u\ (A_3\ B_1)$	-4.7	-3.5
$u\ (A_2\ B_3)$ - $u\ (A_3\ B_2)$	-4.2	-1.5

Of these nine pairs, two show a positive true but a negative predicted difference. That is, two out of nine or 22 percent are "inconsistent." This is close to the 27 percent observed by GKA (Green, Krieger and Agarwal) and 34 percent observed by GSP. If one allows for systematic and random noise in the self-explicated and/or the paired comparison preferences, the incidence of opposite signs for true and predicted differences will increase. Naturally, as the partworths get updated in ACA, the incidence of opposite signs should decrease as the interview progresses. However, by pairing profiles with approximately equal predicted utilities, the chance of opposite signs occurring increases (relative to random pairings).

GSP also observe that in case of consistent signs between true and predicted differences, 86 percent of the time the preference intensity scale values are more extreme than predicted. Similarly, GKA found the judgments to be more extreme almost 75 percent of the time. Johnson (1991) points out that regression analysis provides fitted (predicted) values that have less variation than the actual values for the criterion variable, unless $R^2 = 1$. In my example, the absolute value of the true difference is greater than the absolute value of the initial difference in six out of seven cases (for pairs with the same signs). However, this is due to the assumed structure of the true partworths. So what explains GSP's finding that the paired comparison judgments are "too extreme"?

The notion that the judgments are too extreme is based on the idea that more extreme values are expected to occur half the time. Indeed, if the predictions from the initial (or updated) partworths are unbiased, and the preference intensity judgments are commensurate with the previous data, then randomly selected pairs should not have a more extreme preference intensity scale value more than half the time. However, the pairs are not selected randomly. In ACA, pairs are selected based on closeness in overall predicted utility. And for pairs with small predicted differences in overall utility, the incidence of extreme intensity scale values is expected to be greater than 50 percent (due to noise). Similarly, I would expect that if one selects pairs with the highest predicted differences in overall utility, the incidence of extreme intensity scale values will be less than 50 percent. This indicates that to test for lack of commensurability, an adjustment must be made for the systematic selection of pairs in ACA. Nevertheless, even though 50 percent is not the correct value to use for the null hypothesis, it is still possible that differential weights for the self-explicated and the paired comparison data may improve ACA results.

Perhaps the most interesting result is contained in Table 4 of GSP. In this table GSP show that the mean absolute error in predicting paired comparison judgments is not significantly reduced when the updated partworths are used instead of the initial partworths (based on self-explicated data only). Yet when a systematic difference in scale values between the self-explicated and paired comparison data is allowed, closer agreement between the predicted and actual values occurs. Thus, it appears that ACA results can be improved if these two sets of data can be weighted in a manner that optimizes predictive validity (the weights being different between applications, and perhaps unique to each respondent).

In terms of first choice hits, the addition of the paired comparison data is shown to improve performance. For example, GSP find a 25.5 percent hit rate when the partworths are based only on self-explicated data and a 32.3 percent hit rate for the final partworths (for a 27 percent improvement). Although these hit rates appear to be very low, the difficulty of the task (twelve profiles being designed to be very close in overall utility) is primarily responsible for this. Also, nothing is known about the reliability of the holdout data. Imperfect reliability places an upper bound on the hit rate that is less than 100 percent (see Huber *et al.* 1991). Finally, improvement in the hit rates is only possible for respondents whose partworths could not be captured by the self-explicated data alone.

CONCLUSION

I disagree with GSP that the 34 percent of the time the direction of observed and predicted preference judgments is inconsistent indicates unreliability. Such inconsistencies can also occur when the "true" partworths differ from the initial estimates. I also disagree with GSP that the 86 percent of the time the preference scale values are more extreme than the (same sign) predicted values indicates bias. Extreme scale values occur also because the pairs are not randomly chosen. My perspective on the role of the paired comparisons is that the preference intensity judgments allow for 1) nonlinear effects for continuous attributes and unequal utility differences between successive levels for other attributes, and 2) differential stretching of the implied attribute importances.

I do agree with GSP that it may be worthwhile to have differential weights for the self-explicated and paired comparisons data. One way to accomplish this is for ACA to have a calibration task that is used not only to modify the partworth scales so that various choice rules can be applied in market simulations, but also to estimate optimal differential weights, for each respondent separately.

REFERENCES

Agarwal, Manoj K. and Paul E. Green (1989), "Adaptive Conjoint Analysis Versus Self-Explicated Models: Some Empirical Results," Working Paper, State University of New York at Binghamton, December.

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

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.

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, October.

Green, Paul E. and Vithala R. Rao (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8 (August), 355-63.

Green, Paul E. and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (September), 103-23.

Green, Paul E., Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research* (in press).

Green Paul E., Catherine M. Schaffer, and Karen M. Patterson (1990), "A Validation Study of Sawtooth Software's Adaptive Conjoint Analysis," Working Paper, August.

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*, Ketchum, ID: Sawtooth Software, January.

Johnson, Richard M. (1974), "Tradeoff Analysis of Consumer Values," *Journal of Marketing Research*, 11 (May) 121-7.

Johnson, Richard, M. (1987), "Adaptive Conjoint Analysis," *Sawtooth Software Conference Proceedings*, Ketchum, ID: Sawtooth Software, June, 117-130.

Johnson, Richard M. (1991), "Comment on Green *et al.*," *Journal of Marketing Research* (in press).

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

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.

Wittink, Dick R. and Philippe Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing,* 53 (July), 91-6.

PROCEEDINGS VOLUMES FROM PREVIOUS CONFERENCES

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 Computer 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

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 of 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

1989

VOLUME I

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

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.

VOLUME II

STARTING A PC-BASED CATI (COMPUTER-AIDED TELEPHONE INTERVIEWING) FACILITY

Selecting the Hardware

Part 1: Networks
Del Talley, The Wirthlin Group

Part 2: File Servers, Workstations, and Requests for Bids
Charles Bolton, Consultant to PG&E and Freeman, Sullivan & Company

Selecting CATI Software
Andrew Norton, ARC Professional Services Group

Designing the Facility
Richard Miller, Consumer Pulse, Inc.

Changes in Organizational Management

Part 1: Staffing
Karla Haugan, Lawrence & Schiller

Part 2: Impact on Management and the Bottom Line
Jane Thompson, The Research Spectrum

Since 1989, conferences are held every 18 months; there are no 1990 Proceedings.
To order any Proceedings volume, please contact our Ketchum, Idaho office:
208/726-7772 (fax: 208/726-5156).