Conjoint Analysis Reliability: Empirical Findings

Research Paper No. 795

Stanford University, Graduate School of Business

Prof. David Reibstein**

Prof. John E. G. Bateson*

Prof. William Boulding***

** David J. Reibstein is Associate Professor of Marketing at the Wharton School, University of Pennsylvania, Department of Marketing, Philadelphia, PA 19104 (215) 898-6643.

* John E.G. Bateson is visiting Associate Professor of Marketing, Stanford Business School, Stanford University, Stanford, CA 94305 (415) 497-4260.

*** William Boulding is Assistant Professor of Marketing, Fuqua School, Duke University Durham, NC 27706 (919) 684-3876.

The authors gratefully acknowledge the Support of the Marketing Science Institute and its member companies without whom this research would not have possible

June 1985

The Research paper is not to be copied or referred without the express permission if the authors. Comments welcome.

JL 85P

Conjoint Analysis Reliability: Empirical Findings

ABSTRACT

This paper looks at the comparative reliability of different methodological variants of the conjoint analysis procedure. It differs from previous studies in that it looks at three methods of data collection (Full Profile, Trade-Off Matrices and Paired Comparison), two levels of a key attribute (price) across five different product categories. In addition it tests these manipulations using two different reliability assessment procedures. The results show that all manipulations have a significant effect on the reliability scores and many interaction terms are significant.

Introduction

The term conjoint analysis has come to over a multitute of different methodological procedures. A researcher designing a conjoint analysis study must therefore choose from a large range of alternative procedures. This paper describes a research project which provides information to improve that choice. The research looks at the comparative reliability of the three main methodological variants. That reliability is measured across five products and across a variation in the number of levels of a key attribute. Reliability is measured in two different ways and the research uses a representative sample of consumers.

Since it was first described in the expository paper by Green and Rao (1971), considerable interest has been shown in the application of the additive conjoint analysis model in marketing. The interest in the academic community is evidenced by the stream of articles on the subject. Commercial interest is also high, Cattin and Wittink (1982) estimated that by 1980 some 1000 commercial studies using conjoint analysis had been performed. It is clear that the technique is being used extensively and that major decisions are be being made based on the results.

With the extensive use of conjoint analysis it is clear that the reliability and validity of the procedure should be of major concern. Since reliability is a necessary but not sufficient condition for validity, if the procedure is not reliable then it cannot be valid. From a managerial point of view, however, the issue is often not the absolute reliability. Instead the researchers need to know which of the many variants is more reliable i.e. the 'comparative reliability' of the different procedures.

The name conjoint analysis covers a large range of procedures (Green and Srinivasan 1978). There are many approaches to the decomposition of respondents' part worths that all go under the name "conjoint analysis." The researchers' options when planning such a study are therefore very large.

There are various ways to view the researcher's choice process. There is seldom a strict sequence of choices and the process is often iterative. However there are some basic choices to be made. One of the most basic choices is the method of data collection. Cattin and Wittink (1982) show that the most common procedures are the full profile method and the trade-off matrix. The full profile method was originally proposed by Green and Rao (1971). Respondents are asked to evaluate descriptions or profiles of hypothetical products constructed from factorial designs of the attributes and levels being investigated. In comparison the trade-off matrix approach, originally suggested by Johnston (1974) presents respondents with attributes two at a time. Each pair of attributes is used to create a matrix and respondents are asked to rank the various combinations of levels of the two attributes.

The second basic decision to be made is the nature of the response mode. Cattin and Wittink show that the most common method is the rank ordering of options, this is followed by rating scale and paired comparison. The latter involves presenting the respondents with pairs of full profiles and asking for a simple choice between the two. Thus, a large number of such pairs are presented to respondents and their utilities inferred from the pattern of responses.

Once the data collection method and response mode have been chosen, there are still many choices to be made. The number of attributes to be

used and the number of levels per attribute must be selected. This is integrally linked with the choice of using either a full or a fractional factorial design. The more attributes and levels that are used the more profiles that need to be used in a full factorial design. The rationale for using fractional factorials is generally to reduce respondent fatigue caused by evaluating too many profiles. The fractional factorial designs allow the number of profiles used to be reduced dramatically but require assumptions to be made about the non-significance of many interaction terms (Green 1974).

There are many other decisions which could be mentioned. Assessing all these factors in one study would lead to an impossibly large experimental design. Therefore, earlier work suggesting a framework for future research into the conjoint reliability area was relied upon in selecting dimensions for analysis in this study (see Bateson, Reibstein and Boulding 1985).

The result of this winnowing out of possible experimental manipulations was a smaller, albeit still large, 5X3X2X2 experimental design. The cells contain different products (5), different data collection procedures (full profile, trade-off matrices and paired profile comparisons), different numbers of attribute levels for a key attribute (2 different levels), and a manipulation of the "type" of reliability (2 types)

In the next section of the paper we discuss the nature of conjoint analysis reliability, and what we mean by reliability "types". In particular we suggest that there is no single construct called reliability and instead adopt the generalizability theory framework. We then use this framework to review the limited literature on the comparative reliability of

conjoint analysis. This discussion is used to place in context the large empirical study of comparative reliability which is described and discussed in the rest of the paper.

THE NATURE OF CONJOINT RELIABILITY

In a recent review of the literature, Bateson, Reibstein and Boulding (1985) have identified over thirty reliability studies. They argue, however, that it is impossible to make generalizations from this literature because of the plethora of procedures and approaches used. They suggest that one of the primary sources of this confusion is the lack of clarity over what the construct called "reliability" means in the context of conjoint analysis. To overcome that confusion they advocate the adoption of generalizability Theory.

Generalizability Theory was originally developed in psychology as an extension of the work done on the reliability of multi-item scales. It was developed to explicitly recognize the sources of measurement error under investigation. Up until that point it had been implicitly assumed that the various procedures were measuring a single underlying construct called "reliability". Generalizability Theory, however, recognizes the various sources of unreliability and attempts to measure them separately (Cronbach et al 1963, 1972;1 Gleser et al 1965).

To illustrate this perspective consider a typical reliability check performed on a multi-item scale. Respondents first complete the scale. After some interval they are asked to repeat the exercise. There is much discussion over the appropriate length of the interval. Since it is necessary to avoid memory effects, a long interval seems desirable, but too

long an interval runs the risk that the underlying phenomenon will change. To overcome this, a standard procedure is to use a second set of items form the second administration. This example was suggested by Peter (1979) who argues for the use of Generalizability Theory in marketing. He points out that a traditional reliability approach would involve the computation of a single reliability measure from such a study. Generalizability Theory would argue that the sources of error should be recognized explicitly. The experiment would therefore have to be arranged in such a way that two reliability measures could be computed: reliability over time and reliability over item.

Adopting such a perspective it is clear that a number of different reliability measures have been computed for conjoint analysis:

Reliability over Time asks: "Would the results be the same at a different point of time?"

Reliability of Stimulus Set asks: "Would the results be the same if a different set of stimuli or profiles had been used?"

Reliability over Attribute Set asks: "Would the utilities for a given set of attributes have been the same if these attributes had been included in a study with other attributes?"

Reliability over Data Collection Procedure asks: "Would the results obtained have been the same if a different data collection procedure had been used."

A complete review of the literature using this structure is provided in Bateson, Reibstein and Boulding (1985) but it is perhaps worthwhile to illustrate each of these types of reliability using a published study.

Reliability over Time

As an example of this kind of study, consider the work of McCullough and Best (1979). In their study of preference for apartments they presented their respondents with a card sorting full profile task involving 27 cards. Two days later the same respondents completed the same task. McCullough and Best (1979) assessed the reliability over time of the technique by comparing the results obtained at the two administrations.

Reliability Over Attribute Set

This approach questions whether the part - worths for a given attribute level for an individual depend on the other attributes or levels in the stimuli. Operationally, the tests involve looking at the stability of part worths computed for attributes which are common, when other attributes in the stimuli are varied.

This form of reliability can also be illustrated with the McCullough and Best study of apartments (1979). They surveyed students and asked them to perform two tasks in a single session. The tasks involved the ranking of profiles describing apartments. However, in the second set of profiles one of the four attributes was removed and substituted with another having the same number of levels. McCullough and Best call their approach a test of 'structural reliability', and the second set of stimuli a perturbed form.

Reliability is assessed by comparing the results obtained for the attributes and levels common to both tasks.

Reliability over Stimulus Set

The use of fractional factorial designs has given rise to another major area of conjoint reliability research--reliability over stimuli set. This form of reliability asks whether the answer would have been the same if a different set of stimuli had been used. In its purest form this involves the presentation of different fractional factorial designs to respondents. The part worths estimated in the two halves of the study are then compared. Only three studies have actually used this procedure (Scott and Wright 1976, Parker and Srinivasan 1976, Cattin and Weinberger 1979). Two of these studies (Parker and Srinivasan, 1976 and Cattin and Weinberger, 1979) confounded "reliability over stimuli set" with "reliability over time" by collecting the data for the reliability check at a later time.

A far more common approach to the measurement of reliability over stimuli set is the use of hold out samples. This involves giving the respondent additional stimuli to respond to after the main questionnaire. The number of stimuli is never a full replication and the percentage of the main questionnaire varies dramatically. At the extreme, studies have been done with only one hold-out profile (Tashchian et al 1981).

The hold out studies are almost always called checks of convergent or predictive validity. The distinction between a check of convergent validity and reliability is a fine one. Campbell and Fiske (1959) argue that checks of reliability should involve measurement made with maximally similar measure and checks of convergent validity with maximally different measures.

Since our objective is to compare procedures rather than to assess absolute levels of reliability and validity, the distinction is a semantic one.

Reliability over Data Collection Procedure

Reliability over data collection procedure asks whether the partworths would be different if a different methodological variant had been used. It is with this type of reliability that the distinction between convergent validity and reliability is least clear. Does the comparison of a full profile card sort task conducted by personal interview and a mail trade-off matrix study on the same attributes constitute a comparison of maximally similar or maximally different methods? It is our contention that both involve active evaluation experiments (Scott and Wright 1976) based on full or fractional factorial designs, and therefore, this would be a reliability check.

As an example of such a test consider the study performed by Oppen van Veen and Beazley (1977). They compared the part worths obtained from a full profile card sort with a trade-off matrix (as well as varying the data analysis algorithm used). The comparison of the part worths obtained in these different ways is an assessment of "Reliability Over Data Collection Procedure."

Comparison of Reliability Across Methods

Research on the reliability of conjoint analysis can be broken into two streams. The first, and largest, stream asks whether conjoint analysis is reliable in an absolute sense. The second, and smaller stream, is more

pragmatic and assumes that conjoint studies will be done anyway. This stream therefore asks which is the most reliable technique from amongst all the methodological variants.

Such a comparative emphasis focuses on the needs of the commercial researcher. As Ruch (1977), a senior corporate researcher, points out, it is accepted that all major research techniques have some flaws. He argues, however, that the research will have to be done--the important thing is to understand the nature of the flaws.

Three studies have compared the reliability of the full profile and trade-off matrix approaches. Two of these studies looked at reliability over stimuli set and used a hold-out sample (Jain et al 1979, 1980). In both cases the procedure followed was the same. Respondents performed an initial conjoint task involving either a card sort or completing a trade-off matrix. After some intervening tasks they then evaluated a set of 8 hold out full profiles. The results are slightly confused by this, since a true "reliability over stimulus set" of the trade-off matrix approach would require a complete task replication. The part worth from the two methods are each used to forecasts the ranks of the hold out samples. A comparison of the results indicated that the reliability score was independent of the procedure used.

Segal (1982) made the same comparison but used a measure of reliability over time, collecting two sets of data from the same respondent ten days apart. He compared not only the derived part worth but also the input data and found little difference in reliability between the methods.

A further three studies have looked at the reliability of Hybrid Conjoint Analysis compared with more traditional approaches. In each case the comparison of reliability involved the use of hold out samples. After

computing the partworth for each of the procedures, they were used to forecast the ranking of the hold out sample (Green et al 1982, Akaah and Korgaonker 1983 and Cattin et al 1982). Green (1984) provides a detailed review of these studies.

Two studies have looked at the impact of the number of factors and profiles used (Malhotra 1982, Acito 1979). Both studies simultaneously manipulate the number of factors and the absolute number of profiles. Malhotra measured reliability over stimulus set using a jackknifing approach, and Acito reliability over time, readministering his questionnaire after two and a half weeks. Malhotra used an ANOVA analysis with the standard deviation of the derived parameters as the reliability measure. Acito uses regression on his incomplete factorial design but uses a "distance measure" as the reliability score.

The results of the two studies tend to agree with each other and to be intuitively attractive. Both agree that the number of factors has a statistically significant negative impact on reliability. Similarly, both agree that as the number of profiles presented increases so does the reliability. Malhotra also concludes that the interaction term (factors x profiles) has a statistically significant impact on reliability. Acito's design precludes the estimation of interaction effects.

Leigh et al (1981) provide added support for the proposed effect of number of stimuli. In their study, using small sample sizes, they conclude that less fractionated designs (i.e., more profiles) produce higher reliability. In this study, they measured reliability over time and used a correlation of the B_{ik} as their dependent measure. In a similar 1984 study, using larger sample sizes and a slightly different reliability score, they

were unable to detect any statistically significant impact of number of profiles on reliability.

Both of the Leigh et al (1981, 1984) studies simultaneously manipulated degree of fractionation of the experimental design used and the dependent measure employed. Their 1981 study compares two non-metric procedures (ranking and a paired comparison) with three metric procedures (a graded paired comparison, a dollar metric valuation and a rating scale). The 1984 study drops the dollar metric valuation but uses two forms of graded paired comparison-dollar metric and rating scales. Testing 'reliability over time' they conclude in their 1981 study that comparative judgement procedures outperform absolute judgement procedures. This finding is not supported by their 1984 study which finds no difference between any of the conjoint analysis procedures.

A number of things become clear when reviewing this extremely limited literature on comparative reliability. The first is that compared to the number of methodological variants studied, very little has been done. The studies have hardly begun to address the researchers' problem of which is the most reliable conjoint procedure to use. The Malhotra study is disturbing since it suggests that some of the most common choices do have an impact on reliability. However, the Leigh study (1984) suggests that reliability is less sensitive to these kinds of decisions. The situation is therefore unclear.

The second thing that becomes clear is that the types of reliability studied have also been very limited. Only one study (Malhotra 1982) has compared the "reliability over attribute set" of different procedures and this uses the unusual 'jackknifing' approach. Moreover, each of the studies has only measured one form of reliability. There is no justification for

assuming that different methodological changes will effect the different types of reliability in the same way. If this is not the case then by studying different types of reliability at the same time considerable insights may be gained into the design of new procedures.

The Malhotra (1982) study is also disturbing since it is the only one which has varied more than one methodological variable at a time, and the interaction term was significant. This suggests that these relationships may be more complicated than the simple "main effects" reliability assessment usually made. Finally, it is clear that each of the studies has only limited generalizability. Of the nine studies four use student respondents and the average number of respondents overall is less than 150. Cattin and Wittink (1982) suggest that from their data, most commercial applications have a sample size of between 300 and 500 respondents. There is quite a large spread of products, from H.M.O's to sneakers, but each reliability study uses only one product category.

THE STUDY DESIGN

This research attempts to overcome some of the problems identified in previous studies. It manipulates more than one methodological factor at a time, measures two forms of reliability simultaneously, uses five different product categories and a large representative sample of consumers (not students).

The study looks at the reliability of the three most common data collection procedures: the full profile method; the trade-off matrix method and the full profile using a paired comparison as the measure. It also looks at the impact on reliability of varying the number of levels of one of

the key attributes--price. The design is so constructed that the separate effect of method, number of price levels and their interaction can be measured.

The study measures two forms of reliability, reliability over stimuli set and reliability over attribute set. Although the use of hold out samples to measure reliability over stimulus set is quite common, there are no empirical data available to support the reliability of this procedure. In particular there is no empirical research to guide the selection of an appropriate size for the hold out sample. Only a full replication using a different fractional factorial avoids this problem and allows for the direct comparison of the derived partworths. This approach was therefore adopted.

To measure reliability over attributes we used the perturbation procedure suggested by McCullough and Best (1976). As described earlier, this involves the interchange of one of the attributes in the respondents first task (main task) with a new attribute in the respondents second task (reliability task). This procedure has a number of advantages over the alternative embedding approach suggested by Green and Wind (1973), which can also be thought of as a test of "reliability over attribute set." Embedding analysis often requires a different number of stimuli to cope with the added attributes, and hence confounds changes in number of things at the same time. In addition, the number of parameters estimated in the main and reliability check varies. Perturbation leaves the a number of parameters to be estimated constant, and avoids the problem of confounding.

At this stage it may be worthwhile clarifying those factors that were held constant in the study. All data were collected at the same point of time so that reliability over time was not assessed.

Within product category the number of attributes and the number of levels remained constant across manipulations. This means that the same number of stimuli were used in the main exercise and the reliability checks. Because of this the number of parameters estimated for each product category remained constant, as did the analysis method.

Data Collection

The study is performed across a total of five product categories: telephone service, typewriters, yogurt, retail banking and televisions. These were chosen to represent a broad cross section of product categories purchased by household consumers, and a mixture of goods and services.

The design of such a study involves (1) the identification of determinant attributes (2) the design of the research instrument and (3) the selection of respondents and adminstration.

Identification of Determinant Attributes

The attributes used in the study were identified in consultation with research managers operating in each of the industries. Such attribute batteries constituted a standard instrument in each firm and had been used in many other studies. In each case the senior research manager in the company was asked to identify the six key attributes that previous research had shown to be determinant in consumer choice. This was possible in all cases except banking where only five attributes were identified. In every case price was a key attribute.

The research managers were also asked to identify appropriate levels for each attribute. For the price attribute they were asked to select two cases, the first using three price levels and the second using five price levels. In both cases the same price range was used. Attributes five and six were to be used in the perturbation procedure and so it was necessary for them to have the same number of levels. In this way they could be interchanged within the same fractional factorial design. It is worth noting that these were perceived by the managers as the fifth and sixth "least important" attributes.

The Design of the Research Instruments.

Each respondent was asked to evaluate the main task and reliability check for two product categories. The structure of the questionnaire was identical in every case. Respondents first performed a conjoint task for product A followed immediately by a task for product B. There then followed a battery of demographic and other questions. Finally, respondents completed the reliability check for product A followed by the reliability check for product B.

With such a procedure there was obviously a danger of order effects and respondent fatigue. To overcome order effects, the order of the main manipulations (products and procedures,) were randomized. To minimize respondent fatigue and boredom, it was decided that no respondent should answer questions in part A and part B that dealt with the same product category or used the same conjoint method.

A small computer program was therefore written to design the questionnaires. The first stage was the selection of one of the ten

possible pairs of products. Each product appeared first once and second once in the ten pairs. If the product appeared first, a "reliability over stimulus set" manipulation was employed. If the product appeared second, a "reliability over attribute set" manipulation was used. For each pair one of two levels of the price attribute was selected--three levels of price or five levels of price. For each product/price combination two different conjoint data collection procedures were then selected. This yielded a total of sixty unique questionnaire designs.

The next stage was the selection of a fractional factorial design for each of the unique conjoint tasks. Using the first five attributes only (Four in the case of the banks), a fractional factorial design of twenty five stimuli was identified from standard tables (Adelman 1962). The design was then used to prepare appropriate stimuli for the full profile methods.

The paired comparison requires respondent to choose between pairs of full profiles. The same twenty five profiles were used as in the full profile procedure. To keep the choices to be made by the respondent manageable, a partial block design of 100 pairs was constructed (the minimum size possible far on orthogonal design), using the 25 profiles (Clatworthy 1973).

The trade-off matrices were created using a partially balanced block design (Clatwothy 1973). In this design, each attribute appeared in a table twice, resulting in five trade-off matrices for the respondent task (four in the case of banking)

Once the main questionnaires had been put into place the demographic battery, common to all questionnaires was added. It was then necessary to construct the appropriate reliability checks. The "reliability over stimulus set" test involved the selection of an alternative fractional

factorial design from the tables, in the case of both the full profile and paired-profile comparison. In the case of the trade-off method, the attribute pairings were randomly redrawn, within the constraints of a partially balanced design. In other words, attributes appear with attributes which they had not been paired in the first task. These new stimuli were used in the reliability check.

The "reliability over attribute set" was assessed using the same fractional factorial. Factor five was removed from the questionnaire and replaced with factor six. Thus, for example, the fifth factor for televisions was "remote control yes or no." The sixth attribute was "type of channel selection mechanical/electronic." In the first set of profiles presented to the respondent, the variable "remote control" was used, in the second set it was replaced with the 'channel selection' variable but nothing else was changed.

As indicated earlier, the instrument generation process yielded sixty different questionnaires. Each one was the copied ten times to produce the six hundred questionnaires needed.

Respondent Selection and Administration

The six hundred respondents were selected by intercept in a busy suburban mall in a major U.S. city. They were offered an incentive of ten dollars to complete the questionnaire. If they agreed they were taken to a room in the mall where they were given the questionnaire for self completion. An interviewer was on hand at all times to offer clarification or to answer any questions.

The time taken by respondents to answer the questionnaire varied considerably both with respondent and questionnaire. The mean time taken for the questionnaires containing only the full profile or trade-off tasks was thirty five minutes. Questionnaires containing the paired comparison task look longer and the mean time was approximately fifty minutes.

DATA ANALYSIS

An outline of the data analysis plan is shown in Figure 1. Since each of the sixty questionnaires had a completely different format, data preparation posed an interesting problem. To facilitate key punching a data entry program was produced. Each of the sixty questionnaires carried an identification code which was entered first. The program then generated an exact replica of the questionnaire, page by page on the terminal screen. Into the replica the key punchers were able to copy the respondents answers. The program then unscrambled the randomization to produce the data set.

Insert Figure 1

Since each product used different levels for the five attributes, the data had to be split by product. The different conjoint methods produced different kinds of data so the data set had to be split again. The number of parameters to be estimated varied according to whether three or five price levels were used necessitating another split. Finally the reliability check adopted dictated a different analysis procedure. The final form of

the data was therefore twenty observations in each of sixty cells defined by product (five categories), conjoint method (three categories), number of price levels used (two categories) and reliability method used (two categories.)

The first stage in the data analysis was then to represent each stimuli as a set of dummy variables and submit the data to the ordinary least squares OLS procedure to derive part worths. The whole of the analysis was performed at the individual level. The OLS Algorithm has been shown to yield results close to other, more sophisticated algorithms (Jain et al 1979, Carmone et al 1978, Wittink and Cattin 1981). The partworths were computed for each half of the data separately, (i.e., "main task" versus "reliability task) necessitating a total of 2400 separate regressions. In addition for the reliability over attribute set data a separate 1200 regressions were run using only the attribute common to both halves of the study.

The next stage in the process was to generate a measure of the reliability for each individual. A variety of reliability measures have been suggested in the literature. The measure used in this study was the correlation across attributes and levels of the partworths within individuals. This approach compares the partworths computed for each individuals "main and "reliability" tasks and computes a Pearson Product Moment correlation. This is a standard procedure adopted by many researchers (Etgar and Malhotra 1981, Cattin et al 1982, Weitz and Wright 1979, Jain et al 1980, 1979, Akaah and Korgaonkar 1983, Green et al 1982, Leigh et al 1982, 1984.) In the perturbed form analysis the correlation coefficient was computed on only those attributes and levels common to both halves. The minimum number of observations on which the correlation

coefficient coefficient was computed was seven (banking service in the perturbed form), and the maximum fifteen.

Cattin and Wittink (1982) show in their study that the predominant uses for conjoint analysis are advertising and product policy decisions. In these areas the reliability of the partworths is of the utmost importance. The correlation of the partworths in the two administrations is therefore an appropriate measure.

RESULTS

Each of the two thousand four hundred regressions were used to fit a main effects partworth model. The number of individual partworths fitted varied across product category and price level. For attribute set reliability additional regressions were run. These additional regressions were run using only the attributes common to the main and reliability halves of the questionnaire and less parameters were estimated.

The detailed partworth computed for each attribute and level are not reported here since they were not the focus of this study. The results were shown to the senior research managers in each of the industries. They were asked whether the results were logically consistent. Without exception the results conformed to the managers expectations based on other research.

Table 1 shows the mean and standard deviation of the correlation coefficient for each of the cells in the study. For example the top left cell in the table shows the mean and standard deviation of the correlation coefficients of the twenty individuals who performed a conjoint task for televisions using the full profile procedure, three price levels and a test

of stimuli set reliability. Across those twenty individuals the mean product moment correlation of the main and reliability partworths was 0.79 and the standard deviation 0.19.

INSERT TABLE 1

To aid interpretation a weighted average was computed for each product category across and within reliability method. Consideration of these marginals provides considerable insight into the effects of the manipulations. Across all product categories it appears that "stimuli set" reliability is considerably higher than attribute set reliability. Since the "stimuli set" reliabilities are so high the results suggest that the partworths obtained may be independent of the fractional factorial. The relatively low levels of attribute set reliability can be interpreted in a number of ways. They could be due to the fact that the partworths obtained for any given attribute are not independent of the other attributes used, and a linear additive model is not appropriate. Alternatively they could be due to the use of a main effects only model in a situation were at least some of the interaction terms are significant.

The variations across product category are less clear and seem to depend upon the type of reliability being measured. Using reliability over stimulus set there seems to be little difference in the scores across products. However if we look at reliability over attribute there is a considerable variation in the mean reliability score across products. Telephone and Yogurt in particular produce low scores. Later we will show other differences between products and discuss possible reasons of this.

To test the impact of the various manipulations, an ANOVA was

performed. Table 2 summaries the results for the main effects and interaction terms including product as a variable. The vast majority of the terms are highly significant. Among the main effects the reliability method used has the largest impact. This is in line with the results in Table 1. However, all of the main effects are significant including the manipulation of price level.

The complexity of the interaction terms makes direct interpretation of the results difficult. The inclusion of 'product category' as a variable also pre-supposes as an "a priori" model of its impact on reliability. This was not the case and the products were chosen to represent a cross-section of applications areas. To overcome this and to aid interpretation the ANOVA was re-run for each product category separately. The results are shown in Table 3. To improve readability only the 'Sum of Squares,' the 'F Statistic' and the significance level of the F statistic have been included.

INSERT TABLE 3

Table 3 shows some of the reasons for the complexity of the original ANOVA There are different patterns of significant main effects and interaction terms across products. When the products are included in the ANOVA the model attempts to represent this pattern though the interaction terms.

Clearly with the size of the sample it is important to separate statistical and practical significance. However a number of clear findings do emerge from the tables. The data collection procedure effect is highly significant in all but the typewriter product category. The price level effect reaches significance in only two product categories, television and

telegraph. Across all product categories the reliability method effect is highly significant.

Of the two may interactions very few are significant although one--data collection procedure with reliability method--is highly significant, in the yogurt product category. Similarly the three way interaction for the yogurt product category reaches significance.

To provide further insights into the directional relationships a dummy variable regression was performed, representing each of the manipulations as a dummy variable. In line with the full factorial nature of the study all interaction terms were included. Table 4 shows the standardized regression weights and R-squared for each of the product categories separately.

INSERT TABLE 4

Once again there are some common patterns across product categories but many differences. There is a clear pattern as regards the main effects of the different data collection procedures. In every case the trade-off matrix is significantly less reliable then the paired comparison procedure (the base case). In three out of five products there is no significant difference between the performance of the full profile procedure and the paired comparison and both procedures outperform the trade-off matrices.

For telephone and yogurt however the pattern is somewhat different. In these cases the full profile procedure performs significantly worse than the paired comparison approach and in the telephone product category worse than the trade-off matrix. These result contradict previous findings. Earlier studies (Jain et al 1979, 1980 and Segal 1982) found no differences

in the reliability scores between the full profile and trade-off matrix approaches. Leigh et al (1984), in addition, found no differences between a number of dependent measure types including paired comparisons.

As shown clearly in all of the ANOVA results the reliability method chosen has a major impact on the score obtained. The results in Table 4 show that in all but the yogurt category the attribute set reliability, the base case, produces significantly lower reliability scores. The yogurt results are complicated by a number of significant two way interaction terms. These tend to support the findings in the other product categories since they produce highly significant and positive beta weights for those interaction terms involving stimulus set reliability.

The minor variation in the number of price levels used produces no significant main effects although in a number instances the interaction terms containing this effect are significant. Given that this was a variation in the number of levels of only one out of five attributes it is perhaps not surprising that no significant effects were found.

A total of six significant interaction terms confirm the findings of Malhotra (1982) that the impact of the data collection procedures may be more complex than a single main effects model of reliability can capture. Malhotra showed that in looking at the impact of different manipulations on reliability it was necessary to look at interaction terms.

Perhaps the most disturbing result is the different pattern of significant effects found for the different product categories. The variation in the significance of the full profile effect and the reliability method manipulation stand out. The yogurt product category in particular has unusual results when compared to the others. One explanation for this

may be the nature of the attributes chosen. Table 5 shows the R-Square results obtained from the perturbed form for the different product categories:

INSERT TABLE 5

The top two figures show the R-Squared obtained when the regression were run with the full number of attributes. There are two different results since one attribute is different in the second case. Below the two figures is the R-Squared run only with the attributes common to both halves i.e., with one less attribute. Clear patterns emerges from this table. With the exception of Yogurt and to a lesser extent banks there is little difference between the three numbers. This indicates that the fifth attribute did not make a significant contribution to explain the subjects responses to the stimuli. Moreover the two variants of the last attributes were equally poor at explaining the respondents choices.

In the case of yogurts and to a lesser extent banks there is a large reduction in the R-Squared when using only the common attributes. The last attribute clearly contributed significantly to the explanatory power of the model. Such a variation in the importance of the last attribute could be hypothesized to have a major impact particularly in the perturbed form type of procedures.

DISCUSSION

In the review of the literature of the comparative reliability of conjoint analysis a number of shortcomings were identified. The first

shortcoming was the relative scarcity of such studies. This study has added to the limited knowledge base and has done so in a number of unique ways.

This is the first study that has tested the same manipulations across a number of products. The significant differences shown in all of the analyses show that this was an appropriate decision. Our product categories were not chosen with any a priori theory in mind. It does, however, appear that the results obtained from such studies may not be independent of the product used.

We have suggested one explanation for the variation shown:that the differing levels of importance of the attribute exchanged in the perturbed form analysis may have a major impact. Such an explanation would not stop generalization from one product category to another. However is it also clear that the nature of the yogurt attributes may be different since they attempted to describe taste and textures using words. Other product categories had much 'harder' attributes. Future studies must clearly address the impact of the product category on the results obtained.

The second major feature that distinguished this study was the incorporation of two different reliability procedures. This was based on the generalizability approach which argues that there is no such thing as a single construct called 'reliability'. The significantly different levels of reliability measured with the two procedures attests to the validity of the generazability approach. This argues strongly for the adoption of this perspective in future studies.

Unfortunately, this study has also served to contradict a number of existing findings. This study clearly shows that the type of data collection procedure does have a significant impact on the reliability score, independent of the type of reliability tested. A number of earlier

studies had shown that this was not the case. As mentioned in the literature review, the Jain et al studies (1979, 1980) use a cross-substitution strange method to compute the reliability index and it may be this that caused the discrepancy. The Leigh et al (1984) study tests reliability over time and this may be the cause of the difference. All of the earlier studies use student respondents. Since there is some indication that reliability may increase with the education level of the respondent (Taschian et al 1981), the use of students may have biased the results.

Unlike most other studies this one was constructed to investigate the interaction terms amongst the various manipulations. The high number of them that were significant suggests that future studies should adopt this procedure. The effects illustrated in this study show clearly that single manipulation models for comparative reliability studies are no longer appropriate.

Table 1: Mean (and Standard Deviation) of Product Moment Correlations of Part Worths Analysed by Conjoint Method, Number of price Levels, Type of Reliability and Product Category

	FULL PROFILE		TRADE - OFF MATRIX		PAIRED COMPARISON		
	3 Levels of Price	5 levels of price	3 Price Levels	5 Price Level	3 Price Levels	5 price Levels	WT AVG
Television							
Reliability Over							
Stimuli Set	.79(.19)	.72(.25)	.60(.18)	.54(.18)	.77(.23)	.76(.17)	.70
Attributes Set	.42(.21)	.54(.18)	.30(.23)	.39(.30)	.45(18)	.45(.24)	.43
WT AVG	.61	.63	.45	.47	.61	.61	
Banks							
Reliability over:							
Stimuli Set	.66(27)	.69(.30)	.65(23)	.74(.12)	.57(.30)	.53.(.34)	.64
Attribute Set	.17(.57)	.30(.40)	.04(.26)	.21(.23)	.64(.27)	.66(.17)	.27
WT AVG	.42	.50	.35	.27	.61	.60	
Typewriters							
Reliability over:							
Stimulus Set	.57(.37)	.72(.27)*	.57(.21)	.64(.12)*	.81(.17)	.81(.13)	.69
Attribute Set	.15(.26)	.45(.22)	.26(.34)	.25(.19)	.42(.23)	.51(.19)	.34
WT AVG	.36	.59	.40	.45	.62	.66	
Telephone							
Reliability over:							
Stimulus Set	.72(.24)	.78(.18)*	.61(.27)	.68(.16)	.74(.46)	.78(.19)	.72
Attribute Set	.07(.37)	.16(.28)*	-.26(.18)	.01(.23)*	.16(.25)	.12(.23)*	.04
WT AVG	.40	.47	.18	.35	.45	.45	
Yogurt							
Reliability over:							
Stimulus Set	.77(.17)	.72(.24)	.63(.21)	.74(.12)	.71(.21)	.72(.20)	.72
Attribute Set	-.01(.43)*	-.02(.31)	-.08(.18)	.17(.24)	.06(.30)	.04(.33)	.03
WT AVG	.38	.35	.28	.46	.39	.38	

Table 2 Analysis of Variance of Reliability Score

	Sum of Squares	df	Mean Square	F	Signif of F
Main effects	78.8	8	9.9	154.6	0.000
DATCOL	5.9	2	3.0	46.8	0.000
PDTCAT	6.5	4	1.6	25.6	0.000
PRILVL	0.6	1	0.7	10.3	0.001
RBLMTD	65.7	1	65.8	1032.2	0.000
2-way interactions	14.3	21	0.7	10.7	0.000
DATCOL PDTCAT	2.4	8	0.3	4.7	0.000
DATCOL PRILVL	0.3	2	0.1	2.4	0.088
DATCOL RBLMTD	1.7	2	0.9	13.6	0.000
PDTCAT PRILVL	0.4	4	0.1	1.9	0.109
PDTCAT RBLMTD	9.2	4	2.3	36.3	0.000
PRILVL RBLMTD	0.1	1	0.1	2.0	0.158
3-way interactions	8.3	22	0.4	6.0	0.000
DATCOL PDTCAT PRILVL	1.3	3	0.2	2.6	0.009
DATCOL PDTCAT RBLMTD	6.6	8	0.8	13.1	0.000
DATCOL PRILVL RBLMTD	0.1	2	0.1	1.1	0.333
PDTCAT PRILVL RBLMTD	0.2	4	0.1	0.9	0.439
4-way interactions	0.9	8	0.1	1.9	0.060
DATCOL PDTCAT PRILVL RBLMTD	0.9	8	0.1	1.9	0.060
Explained	102.50	59	1.7	27.3	0.000
Residual	72.1	1132	0.1		
Total	174.6	1191	0.1		

Where DATCOL: Data Collection Precedure
PDTCAL: Product Category
PRILVL: Number of Price Levels
RBLMTD: Reliability Method

TABLE 3 ANALYSIS OF VARIANCE OF RELIABILITY SCORE RUN SEPARATELY FOR EACH PRODUCT CATEGORY

	TELEVISION			BANKS			TYPEWRITERS			TELEPHONES			YOGURT		
	SUM OF SQUARES	F		SUM OF SQUARES	F		SUM OF SQUARES	F		SUM OF SQUARES	F		SUM OF SQUARES	F	
MAIN EFFECTS	28.93	130.25	0.000	5.53	30.21	0.000	28.55	107.80	0.000	9.59	42.99	0.000	11.87	30.47	0.000
DATCOL	1.76	15.87	0.000	1.34	14.68	0.000	0.02	0.15	0.863	1.87	16.79	0.000	3.32	17.47	0.000
PRILVL	0.37	6.67	0.010	0.00	0.04	0.845	0.12	1.87	0.173	0.62	11.10	0.001	0.00	0.00	1.000
RLBMTD	26.79	482.49	0.000	4.17	91.15	0.000	28.42	429.30	0.000	7.08	126.91	0.000	8.30	87.49	0.000
2 WAY INTER ACTIONS	0.62	2.22	0.053	0.26	1.15	0.34	0.76	2.29	0.047	0.54	1.93	0.090	8.09	17.05	0.000
DATCOL PRILVL	0.31	2.75	0.066	0.02	0.22	0.80	0.55	4.18	0.017	0.49	4.42	0.013	0.25	1.33	0.266
DATCOL RLBMTD	0.27	2.39	0.094	0.07	0.76	0.47	0.17	1.26	0.285	0.00	0.00	1.997	7.81	41.44	0.000
PRILVL ALBMTD	0.04	0.67	0.414	0.17	3.78	0.05	0.04	0.54	0.465	0.05	0.80	0.371	0.05	0.54	0.462
3 WAY INTERACTION															
DATCOL PRILVL RLBTD	0.173	1.55	0.214	0.13	1.39	0.25	0.07	0.53	0.590	0.14	1.25	0.287	0.61	3.21	0.04
EXPLAINED	29.72	48.66	0.000	5.92	11.76	0.000	29.37	40.34	0.000	10.27	16.74	0.000	20.27	19.41	0.00
RESIDUAL	12.38			10.32			15.03			12.60			21.34		
TOTAL	42.10			16.31			44.40			22.87			41.81		

Where DATCOL: Cata Collection Procedure
PRILVL: Number of Pricle Levels
RLBMTD: Reliability Method

TABLE 4 STANDARDIZED DUMMY VARIABLE REGRESSION OF RELIABILITY SCORE AGAINST THE MANIPULATIONS ANALYZED BY PRODUCT CATEGORY

	TV	BANKS	TYPE-WRITER	TELEPHONE	YOGURT
MAIN EFFECTS					
DATCOL					
FULL PROFILE	-0.10(1.39)	-0.09(.55)	-0.08(0.82)	-0.42(13.50)**	-0.53(23.53)**
TRADE-OFF MATRIX	-0.47(31.59)**	-0.31(6.19)*	-0.16(3.20)*	-0.25(4.73)*	-0.67(37.37)**
PRILVL					
5 PRICE LEVELS	-0.05(0.28)	-0.06(0.19)	-0.03(0.11)	0.14(1.34)	0.03(0.05)
RLBMTD					
SIMULUS SET	0.69(60.55)**	0.56(18.47)**	0.76(64.1)**	0.62(26.6)**	-0.08(0.44)
TWO WAY INTERRACTIONS					
DATCOL/PRILVL					
FULL PROFILE/5 PRICE LEVEL	0.11(1.36)	0.22(2.55)*	0.01(0.02)	0.26(4.41)*	0.10(0.61)
TRADE-OFF MATRIX/ 5 PRICE LEVEL	0.27(8.32)*	0.16(1.37)	0.24(5.79)*	-0.11(0.76)	-0.24(3.91)*
DATCOL/RBLMTD					
FULL PROFILE /STIMULUS SET	0.05(0.31)	0.10(0.52)	0.11(1.28)	0.04(0.12)	0.50(16.50)**
TRADE-OFF MATRIX/ STIMULUS SET	0.25(6.83)*	0.00(0.00)	0.05(0.25)	-0.09(0.49)	0.60(23.79)**
PRILVL/RBLMTD					
STIMULUS SET/ SERVICE LEVEL	0.07(0.43)	0.03(0.05)	0.03(0.06)	-0.12(0.64)	-0.07(0.25)
THREE WAY INTERACTIONS					
TRADE-OFF MATRIX/ 5 PRICE LEVELS/ STIMULUS SET RELIABILITY	-0.17(3.04)*	-0.18(1.49)	-0.11(1.04)	0.14(1.15)	0.27(4.50)*
FULL PROFILE/5 PRICE LEVEL/ STIMULUS SET RELIABILITY	-0.06(0.42)	-0.23(2.56)	-0.04(0.17)	-0.06(0.23)	-0.02(0.02)

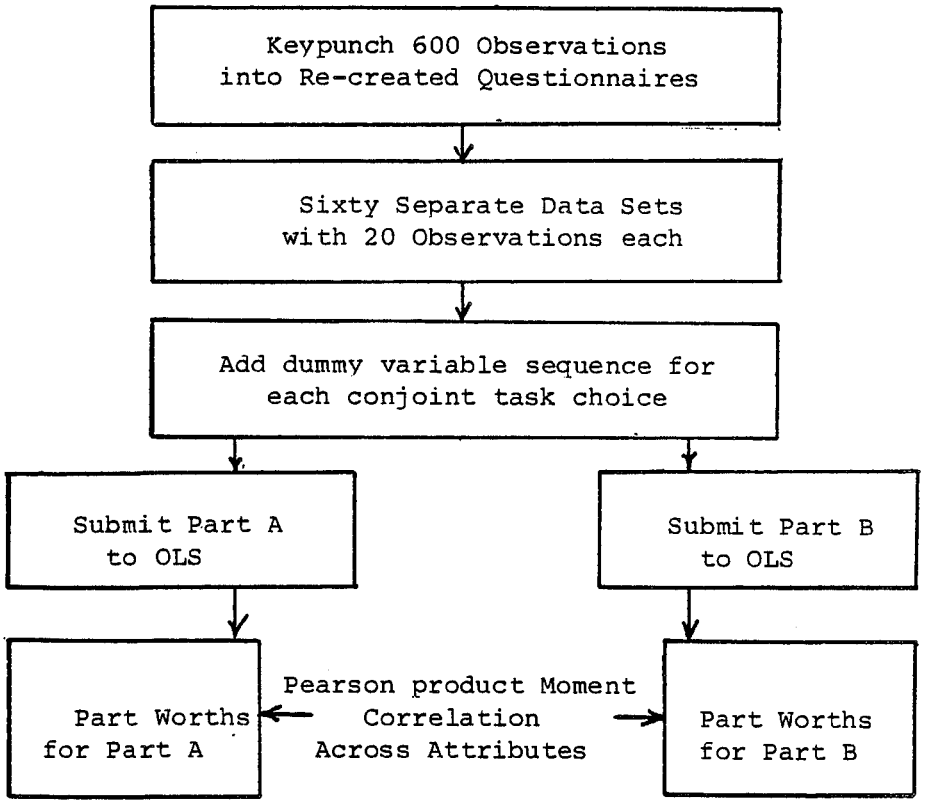
* Significance at the 0.05 level
** Significance at the 0.01 level
Results show standardized regression coefficient and F Statistic

Table 5 = Average R^2 of the Various Perturbed Form Models Analyzed by Product Category, Conjoint Methods, and Number

	FULL PROFILE		TRADE-OFF MATRIX		PAIRED COMPARISON	
	3 Levels of Price	5 Levels of price	3 Levels of Price	5 Levels of Price	3 Levels of Price	5 Levels of price
Television						
5 Attributes	.87/.85	.91/.91	.63/.71	.60/.75	.26/.28	.25/.27
4 Attributes	.87	.87	.56	.56	.27	.25
Banks						
5 Attributes	.69/.77	.76/.74	.63/.59	.64/.66	.22/.26	.23/.26
4 Attributes	.40	.47	.33	.41	.06	.08
Typewriters						
5 Attributes	.78/.79	.85/.87	.61/.63	.64/.63	.24/.24	.24/.25
4 Attributes	.70	.80	.50	.51	.22	.22
Telephone						
5 Attributes	.78/.78	.84/.83	.65/.64	.60/.56	.21/.26	.19/.22
4 Attributes	.73.	.79	.59	.08	.08	.09
Yogurt						
5 Attributes	.81/.79	.89/.85	.73/.67	.69/.64	.33/.34	.35/.35
4 Attributes	.46	.46	.49	.51	.19	.18

There were 2 different sets of 5 Attributes since one attribute was switched

Figure 1 Data Analysis Procedure



REFERENCES

Acito, Franklin (1979), "An Investigation of Reliability of Conjoint Measurement for Various Orthogonal Designs," in Preceedings Southern Marketing Association 1979 Conference eds R.S. Franz, R.M. Hopkins, A. Toma, University of Sothewestern Louisiana 175-78

-------, Arun Jain (1980), "Evaluation of Conjoint Analysis Results: A Comparison of Methods, Journal of Marketing Research, (Feb.), 106-112.

Adelman, S. (1962), "Orthogonal Main-Effect Plans for Asymetrical Factorial Experiments," Technometrics, 4, 21-49.

Akaah, Ishmael P., Praddep K. Korgaonkar (1983), "An Empirical Comparison of the Predictive Validity of Self-Explicated, Huber-Hybrid, Traditional Conjoint, and Hybrid Conjoint Models," Journal of Marketing Research, (May) 187-97.

Bateson John, David Reibstein, William Boulding (1985) "Conjoint Analysis Reliability and Validity: A Framework for Future Research," Working Paper No 792, Graduate School of Business, Stanford University.

Campbell, Donald R and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait--Multimethod Matrix," Psychological Bulletin, 5b, 81-105.

Cattin Phillipe, Marc Weinberger (1979) "Some Validity and Reliability Issues in the Measurement of Attributes Utilities," in Advances in Consumer Research Vol. VII, J. C. Olsen Ed, Ann Arbor: Association for Consumer Research, 780-3.

---------- Garard Hermet, and Alain Pioche (1982). "Alternative Hybrid Models for Conjoint Analysis: Some Empirical Results," in Analytic

Approaches to Product and Marketing Planning: The Second Conference, R. K. Srinivasan and A. D. Schocker Eds, Cambridge, MA: Marketing Science Institute 142-52.

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

Carmone, Frank J, Paul E. Green and Arun K. Jain (1978) "The Robustness of Conjoint Analysis: Some Monte Carlo Results" Journal of Marketing Research, 15 (May) 300-3.

Clatworthy, Willard H. (1973), Tables of Two-Associate-Class Partially Balanced Designs, Washingtn DC: U.S. Department of Commerce.

Cronbach J. J., N Rajaratnam and G. C. Gleser (1963), "Theory of Generalizibility: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, 16 (November), 137-63

----------, G. C. Gleser, H. Nanda and N. Rajaratnam (1972) The Dependability of Behavioral Measurement: Theory of Generalizability for Scores of Profiles New York: John Wiley and Sons, Inc.

Etgar, Michael, and Naresh K. Malhotra (1981), "Determinants of Price Dependency: Personal and Perceptual Factors," Journal of Consumer Research, 8 (September), 217-22.

Gleser, G. C., L. J. Cronbach and N. Rajaratnam (1965), "Generalizability of Scores Influenced by Multiple Sources of Variance." Psychometrika, 30, (December) 395-418.

Green, Paul E, and Vithala R. Rao (1971), "Conjoint Measurement for Quantifying Judgemental Data," Journal of Marketing Research, 8, 355-63.

----------, Yoram Wind (1973), Multi-Attribute Decisions in Marketing: A Measurement Approach Hinsdale, Ill.: The Dryden Press.

---------- (1974), "On the Design of Choice Experiments Involving Multfactor Alternatives," Journal of Consumer Research, 1 (September). 61-68.

----------, and V. Srinivasan (1978). "Conjoint Analysis in Consumer Research: Issues and Outlook," Journal of Consumer Research. 5 (September), 103-23.

----------, Stephen M. Goldberg, and James B. Wiley (1982), "A Cross Validation Test of Hybrid Conjoint Models," Advances in Consumer Research, Vol. X, Ann Arbor: Association fo Consumer Research, 147-50

---------- (1984), "Hybrid Models of Conjoint Analysis: an Expository Review," Journal of Marketing Research, 21 (May) 155-69.

Jain, A. K., Franklin Acito, Naresh K. Malhotra, and Vijay Mahajan (1979), "A Comparison of Internal Validity of Alternative Parameter Estimation Methods in Decompositional Multi-attribute Preference Models," Journal of Marketing Research. 16 (August). 313-22.

----------, Naresh K. Malhotra, Christian Pinson (1980), "Stability and Reliability of Part-Worth Utility in Conjoint analysis: A Longitudinal

Investigation," European Institute of Business Administration Working Paper 80/05.

Johnson, Richard M. (1974), "Trade-Off Analysis of Consumer Values," Journal of Marketing Research, (May) 121-27.

Leigh, Thomas W., David B. MacKay, and John O Summers (1981), "On Alternative Experimental Methods for Conjoint Analysis," Advances in Consumer Research, Vol VIII, Ann Arbor; Association for Consumer Research, 317-22.

---------- (1984), "Reliability and Validity of Conjoint Analysis and Self Explicated Weights: A Comparison," Journal of Marketing Research, 21 (November) 456-62.

Malhotra, Naresh K. (1982), "Structural Reliability and Stability of Nonmetric Conjoint Analysis," Journal of Marketing Research, (May), 199-207.

McCullough, James L. and Roger Best (1979), "Conjoint Measurement: Temporal Stability and Structural Reliability," Journal of Marketing Research. (February), 26-32.

Oppedijk van Veen, W.M., David Beazley (1977), "An Investigation of Alternative Methods of Applying the Trade-Off Model," Journal of Marketing Research Society, 19 (January), No 1, 2-11.

Parker, Barnett, R. and V. Srinivasan (1976), "A Consumer Preference Approach to the Planning of Rural Primary Health-Care Facilities," Operations Research. 25 (Sept.-Oct.), 991-1025.

Peter, J Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16 (February), 6-17.

Ruch D (1977), "Managerial Implications of Analytic Approach to Market Planning" in Analytic Approaches to Product and Market Planning, A D. Shocker, ed., Cambridge, MA: The Marketing Science Institute, 481-4.

Scott, Jerome E. and Peter Wright (1976), "Modelling an Organizational Buyer's Product Evaluation Strategy: Validity and Procedural Considerations." Journal of Marketing Research. 13 (August), 211-24.

Segal, Madhav N. (1982), "Reliability of Conjoint Analysis: Contrasting Data Collection Procedures," Journal of Marketing Research, (Feb.), 139-143.

Taschian, Armen, Roobina O. Taschian, Mark E. Slama (1981), "The Impact of Individual Differences on the Validity of Conjoint analysis" Advances in Consumer Research, Vol. IXA, A. Mitchell, ed. Ann Arbor: Association for Consumer Research, 363-6.

Weitz Barton and Peter Wright (1979), "Retrospective Self-Insight on Factors Considered in Product Evaluation," Journal of Consumer Research, (Dec), Vol 6, No. 3, 280-94.

Wittink, Dick R. and Phillipe Cattin in (1981), "Alternative Estimation Methods for Conjoint Analysis: A Monte Carlo Study," Journal of Marketing Research. (February), 101-6..