Teichert, Thorsten Andreas

A model of ranked conjoint-data and implications for evaluation

Manuskripte aus den Instituten für Betriebswirtschaftslehre der Universität Kiel, No. 461

Provided in Cooperation with:
Christian-Albrechts-University of Kiel, Institute of Business Administration

Manuskripte
aus den
Instituten für Betriebswirtschaftslehre
der Universität Kiel

Nr. 461

Teichert, Thorsten

A Model of Ranked Conjoint-Data and Implications for Evaluation

Nr. 461

Teichert, Thorsten

A Model of Ranked Conjoint-Data and Implications for Evaluation

Institut für betriebswirtschaftliche Innovationsforschung
Christian-Albrechts-University, Kiel
Olshausenstr. 40, 24 098 Kiel, Germany

November 1997

A Model of Ranked Conjoint-Data and Implications for Evaluation

Abstract

This article examines basic features of ranked Conjoint-data, analyzes the adequacy of evaluation methods and proposes improvements for better utilizing the information provided by ranked data. It is shown that commonly used goodness-of-fit measures provide inadequate proxy measures for assessing rank consistency and internal validity of estimates. In addition, commonly used evaluation methods, such as OLS and LINMAP, are shown to be based on arbitrary propositions which do not fulfill the requisite traits postulated by the model of ranked Conjoint-data. Resulting shortcomings on estimation outcomes are evaluated with means of simulation analyses. New insights into the achievable estimation accuracy are gained and possibilities for improvement are shown.

Table of contents

1 Introduction

Conjoint analysis is a decompositional method of marketing research which is widely used for retrieving consumer's utility functions. It is based on the assumption that there are limits to the respondents' ability to articulate their utility function. In particular, it is assumed that respondents implicitly base their purchasing decisions on a metric, linear-additive utility function, but that they are not able to articulate it. Thus, conjoint analysis asks respondents for a holistic assessment of their preferences with regard to stimuli, which describe products as combination of product characteristics (=variables) and their respective levels. Statistical tools are applied to detect the underlying metric utilities of the single variables (i.e. their part-worth values).

With respect to articulation accuracy, it has traditionally been argued that reliable assessment of preferences can best be obtained in terms of rankings (Green, Rao, 1971, Green, Srinivasan, 1978). It is assumed that higher qualified evaluation tasks, e.g. ratings, overtax respondents' capabilities and thus lead to inconsistent answers. Empirical comparisons have indeed failed to observe improved estimates when refined answering scales were used (e.g. Wittink et al., 1989, Kalish, Nelson, 1991, Steenkamp, Wittink, 1994). Nowadays, rating scales are more frequently employed. However, this seems to be less a consequence of methodological critique but more of practical considerations, such as the ease with which the method is used in telephone-interviews, hybrid models and standardized software packages.

Conjoint analyses which are based on ranked data require a bi-directional transformation of scale: First, respondents implicitly assess the metric utility values of the stimuli presented. This lays the basis for the articulated ranking. In a second step, statistical methods are applied to transform the ranking back into a combination of estimated metric part-worth values. Finally, goodness-of-fit is measured on the individual level as the match between estimated and observed ranking. This analytical framework of ranked Conjoint-Data is visualized in figure 1. The arrows indicate the sequence of the above sketched steps being undertaken in a conjoint experiment. The shadowed areas indicate information which remains hidden for the evaluator.

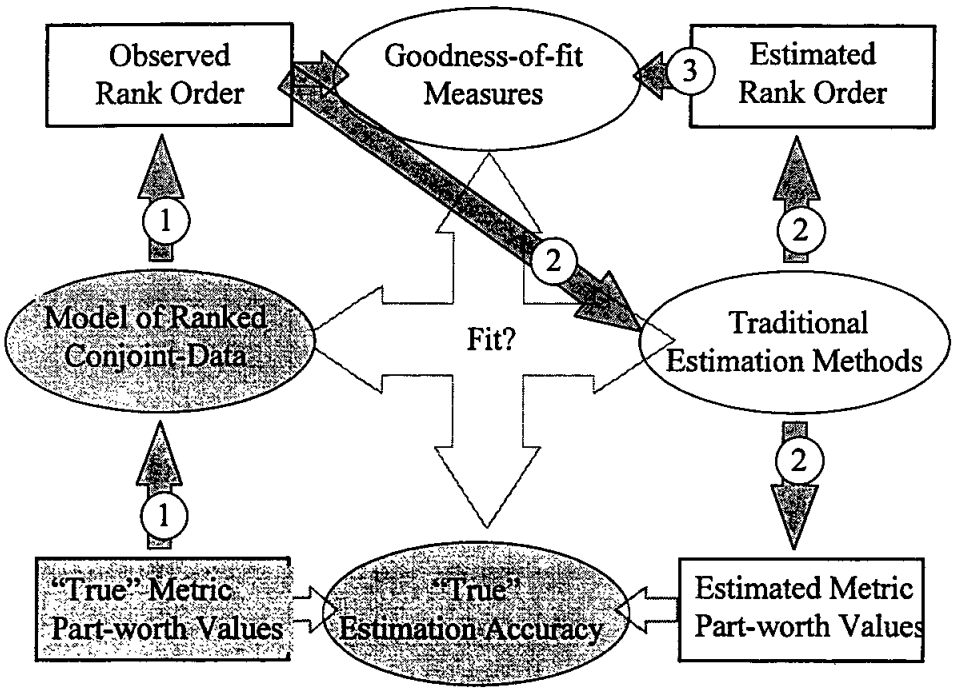

Figure 1: Framework of Ranked Conjoint-Data (Evaluation steps)

From this figure it can be seen that traditional estimation methods and their goodness-of-fit measures focus on the rank order data. This constitutes the basis for potential shortcomings for assessing the internal validity of individual-level estimates: While the evaluator is interested in achieving estimation accuracy between estimated and true metric part-worth values, he is only able to measure the goodness-of-fit between observed and estimated ranking. Latter measure may serve as a substitute of former one only if the scale transformation can be unambiguously reversed and if the estimation methods correctly reshape the transformation process of respondents, i.e. the model of ranked conjoint-data. Thus, the question of fit is of central importance for assessing both the adequacy of estimation methods and the information content of goodness-of-fit measures.

This paper provides a more complete view of the analytical framework and specifies a model of ranked conjoint data. The loss of information resulting from use of rank-order data is quantified. The fit of traditionally used evaluation methods is evaluated both in terms of meeting the model requirements and in its consequences on the estimation accuracy (Figure 1). This enables an in-

depth analysis on the methodological adequacy and on the internal validity of commonly used estimation methods.

There seems to be need for this model-based approach of assessing rank-based conjoint techniques: On the one hand, evaluators can choose between various estimation techniques which base on different propositions, as they range from metric to choice interpretations (Chapman, Staelin, 1982) of rank data. On the other hand, the empirical problems of cross-validating their outcomes are various (Bateson et al., 1987) and yet there is no clear indication on the superiority of one estimation method. Hence, the model-based approach should help to find a reasonable and common basis for comparison. A critical investigation of commonly applied assumptions seems especially relevant, as research work has already quantified the loss of information, which results in rating scales from deviations from equal interval-sizes (Srinivasan, Basu, 1989). In this context, a model of ranked Conjoint data should help to further assess the possibilities and limitations of ranked-based Conjoint-application.

In particular, several propositions are presented which characterize the model of ranked Conjoint-Data. They suggest that the bi-directional scale transformation causes problems in achieving accurate estimates. In addition, traditional estimation methods and their goodness-of-fit measures show to be not in full concurrence with the model of ranked conjoint data. They neglect the underlying pattern of „true" metric utilities and thus valuable information. Implications for evaluation are demonstrated by simulations, which enable a comparison of individual-level estimates with the otherwise unknown „true" metric part-worth values. Systematic and non-systematic shortcomings are distinguished to assess implications on aggregated level evaluations. A new heuristic evaluation procedure is presented which utilizes some of the additional information of the model of ranked Conjoint-Data.

2 The basic features of ranked Conjoint-data

Ranked Conjoint-data do not possess a metric anchorage since they contain only information on relative values. In addition, they are not error-free

translations of the true metric utility values but are characterized by a stochastic error term.

In order to allow for unambiguous calculation of rank inequalities, interval-scale properties are widely assumed. In the following sections, the accuracy of this assumption is tested. The ambiguities in relating preference functions to error-free rankings and those to empirical rankings are subsequently analyzed. Finally, a stochastic model is developed to adequately model the error term. At each stage of analysis, propositions formulate the requisite traits of evaluation procedures. They specify a model of ranked Conjoint-data.

2.1 Rank inequalities

Ranked based conjoint analysis does not ask directly for stimuli's utility, but obtains relative statements on the preferability of any stimulus against all others. The metric stimuli utilities remain ambiguous, since a metric anchorage is missing and rank neighbours provide only lower and upper bounds which by themselves are not fixed but depend on the location of consecutive ranks.

This has an impact on part-worth estimates which are gained by comparing the preference values at different variable levels. If metric data were existent, estimating contrasts could be evaluated by subtracting average preference values (Box, Hunter, Hunter, 1978). The procedure does not work for ranked data, since these are in mathematical terms only a set of consecutive inequations. Instead, the ranking information can be segregated into all possible combinations of pairwise comparisons. The resulting set of inequalities contains all information which can utilized to determine part-worth estimates. A ranking of n stimuli provides n(n-1)/2 inequalities. Most of them provide consistency checks, few determine the limits of the part-worth estimates.

For example, assume a full-factorial design with three dichotomous variables A,B and C and 2^3=8 stimuli and a fully consistent rank order according to an utility function with part-worth values for each upper level +A=40, +B=35 and +C=25 utility points. This would lead to an error-free rank order as following:

$$R(-A,-B,-C)<R(-A,-B,C)<R(-A,B,-C)< R(A,-B,-C)< R(-A,B,C)<R(A,-B,C)<R(A,B,-C)<R(A,B,C)$$

Out of this rank order, a total of 28 pairwise comparisons can be gained, which contain following non-redundant inequalities (Teichert, 1994):

- The preferred levels of each variable are assessed by comparisons of stimuli pairs which differ only in the level of a single variable, as:

$$R(-A,-B,-C)<R(-A,-B,C) \Rightarrow -C < +C$$

- The importance weights of each variable relative to another one are assessed by comparisons of stimuli pairs which distinguish themselves in the levels of exactly two variables, as: $R(-A,-B,C)<R(-A,B,-C) \Rightarrow +C < +B;$ $R(-A,B,C)<R(A,-B,C) \Rightarrow +B < +A$

- Finally, a variables' importance weight relative to combinations of the other variables are assessed by comparing the closest stimuli which distinguish themselves in all three variables: $R(+A,-B,-C)< R(-A,B,C) \Rightarrow +C > +A-B$

In combination, the ranking provides the following information: (a) $0<+C<+B<+A$ and (b) $A<B+C$. These inequalities determine the part-worth estimates and summarize all information which can be gained from the above rank order. Accordingly, the part-worth values for variables' upper levels, if normalized to 100 utility points, can vary within $A \in \{35;49\}; B \in \{26;48\}; C \in \{3;32\}$. The more variables and stimuli are included, the more complex comparisons are possible. This limits the ambiguity of part-worth estimates, but does not lead to single point estimates.

The inequalities convey an ambiguous relationship between ranked data and the underlying metric utility function. To resolve this ambiguity, it is necessary to recode the inequalities into a system of equations. This requires to formulate assumptions on the metric distribution patterns of observed ranks. Thus, it is commonly assumed that ranked data approximate an interval scale.

2.2 Interval-scale properties

The assumption of interval-scaled properties is widely applied but has never been thoroughly substantiated. The literature relies either on very rough criteria, such as correlation measures between rank ordered and interval-scaled data (e.g. Carmone et al., 1978; Wittink, Cattin, 1981), or refers to an early research

work by Colberg (1977). Colberg's simulation study showed that the „metric determinacy" [1] of ranked data increases with higher number of stimuli. Accordingly, many researchers felt confident in assuming interval-scaled properties when utilizing extended designs with larger number of stimuli (Green, Srinivasan, 1978, 1990).

However, the findings of Colberg have limited generalizibility due to certain limitations of the study, such as the neglect of an error-term, the limited spread of analyzed utility functions[2] and the robustness of the measurement criterion itself[3]. In addition, Colberg observed contradictory findings in a few cases (Colberg, 1977, p. 52). The most severe shortcoming, however, is the fact that metric determinacy constitutes only an indirect proof of interval-scaled property. As will be shown below, there are other effects which can cause a high similarity between estimated and true part-worth values in error-free simulations.

Ranked data approximate interval scale properties if the differences between neighbouring stimuli either become very small or if they become uniform. The more stimuli are included in a design with fixed endpoints, the higher the density of observations and the smaller their average utility difference. However, this is of limited information, since scaling of utility functions is arbitrary. Thus, the design should select equally spaced stimuli in order to indicate interval-scale properties. Orthogonal designs, however, balance the variable levels and not the differences of stimuli utilities. Coincidence of scales is not inherently guaranteed. Based on previous analysis (Teichert, 1994) , we expect systematic differences between three types of preference functions:

- Homogeneous preference functions with similarly important variables.
- Heterogeneous preference functions with variables of different importance weights.
- Dominant preference functions, where preferences are dominated by a single variable.

A simulation is performed to exemplify the interval scale properties. Two design types of differing complexity are compared: an orthogonal 2^{5-2} main-effect design with 8 stimuli and an extended 2^{5-1} design with 16 stimuli. 100 different

utility functions are generated for each of type of preference function[4]. Normalization is applied which forces the sum of absolute part-worth values to 100 utility points. These utility functions also serve as basis for consecutive analyses in the following sections.

Table 1 provides a summary of simulation outcomes. As expected, doubling of the number of stimuli divides the average utility difference between neighbouring stimuli nearly in half. However, the coefficient of variation, which relates the standard deviation to the mean value of utility differences, is differently affected at the types of preference function:

(a) In case of homogeneous preference functions, the coefficient of variation is increased by more than 50%, since the standard deviation of utility differences remains nearly constant (s=17.7 in the extended design versus s=19.6 in the main-effect design). Here, the extended design is likely to include nearly identical stimuli, e.g. if two stimuli are differentiated from one another only in terms of two adversary effects. Thus, homogeneous preference functions lead to deviations from interval scale properties in the extended design.

(b) In contrast hereto, the coefficient of variation remains stable for both the heterogeneous (1.09 versus 0.90) and dominant preference functions (1.21 versus 0.99) This indicates that the extended design provides better interval-scaled estimates. The variation of utility differences is reduced if more stimuli are included, because there is a higher probability that combinations of smaller variables offset the influence of a larger variable (Teichert, 1996). However, if the dominant variable accounts for more than 50% of total utility, then the ranking appears to be inherently shaped by a lexicographic processing. Thus, a larger deviation from interval scale remains.

Table 1: Distribution of metric utility differences between rank-pairs*

Type of preference function	2^{5-2} main-effect design			extended 2^{5-1} design		
	mean difference	coefficient of variation	range of differences	mean difference	coefficient of variation	range of differences
Homogeneous	21.0	0.93	53.1	11.6	1.53	58.3
	(19.6)	*(0.26)*	*(17.6)*	*(17.7)*	*(0.28)*	*(6.9)*
Heterogeneous	22.6	0.90	55.5	12.2	1.09	45.8
	(20.4)	*(0.22)*	*(15.6)*	*(13.1)*	*(0.31)*	*(10.2)*
Dominant	25.6	0.99	70.2	12.9	1.21	56.9
	(25.3)	*(0.28)*	*(21.8)*	*(15.6)*	*(0.37)*	*(22.7)*

**) The data are normalized according to common practice of Conjoint-analysis:* The worst possible stimulus has an utility value of -100; the best possible of +100.

The range of utility differences between rank-neighbouring stimuli can be perceived as a more rigorous test for the interval-scale property. The range amounts up to 25% or more of the total utility interval (ranging from -100 to +100 utility points), which implies that a large proportion of the variance remains hidden within the ranking. Thus, the approximation of ranked data to an interval scale remains questionable. It can not simply be achieved by increasing the number of stimuli.

This leads us to:

Proposition 1: Evaluation procedures should not blindfoldedly follow interval scale assumptions.

2.3 Relationship between metric preference functions and rankings

Since ranked data may fail to approximate the interval scale, the inequalities provided by ranked data can not be unambiguously translated into a set of equations. For sets of inequalities, however, multiple solutions exist. This constitutes a well-known problem in operations research (Stimson, 1969; Brockhoff, 1972) and it was even recognized by early scholars of conjoint analysis (Srinivasan, Shocker, 1973). Thus, different preference functions may well lead to the same ranking.

Since rankings build a set of consecutive inequalities (section 2.1), they provide lower and upper limits for variable estimates. Within these limits all solutions are equally feasible. Therefore, alternative preference functions form estimation intervals. An estimation interval conditional of other variable estimates given can be assessed by modifying the estimate for a single independent variable so long as no rank reversal occurs. In order to obtain overall estimation intervals, one has to replicate this procedure for all possible combinations of the other variables. This, however, requires extensive simultaneous calculations.

An easily applicable, heuristic procedure is proposed to approximate the size of the deterministic estimation intervals. First, the observed rank order is recoded into multiple sets of metric utility values by adding stochastic terms to each rank while keeping the rank order unchanged. Subsequently, standard OLS technique is used for estimating the utility functions. The results are saved if a fit of observed and estimated ranking is achieved and if the estimated utility function deviates by more than 10 utility points from any of the solutions already found.

Table 2 summarizes the findings of the simulation example. It shows that the estimation intervals are surprisingly large. The estimation intervals of entire preference functions account on average for more than 90 utility points in the main-effect or 35 in the extended design. Thus, the part-worth values of two alternative preference functions which lead to the same error-free ranking can deviate by up to this amount. This shows that the true preference functions can be quite different. In addition, a more detailed analysis shows that even the ranking of variable importance weights by itself can differ substantially. For example, an estimation interval may include solutions, in which two variables are either the least important or the most important ones (Teichert, 1997). Thus, a high degree of ambiguity in effect estimates has to be diagnosed.

Looking at the individual variables, the largest estimation intervals are to be found for dominant variables. As expected (section 2.2), the information provided by the ranking inequalities is insufficient to closely localize their true value. Finally, the least important variable seems to be especially affected: while the absolute value is lowest, the size of the estimation interval is still above average. With estimation intervals of more than 18 (10) utility points it is

even likely to observe sign reversals for the least important variable which is unlikely to posses an importance weight of more than 10 utility points. Thus it can be concluded that the relative accuracy of variable estimates deviates between variables with different importance weights.

Table 2: Size of deterministic estimation intervals

Type of preference function	2^{5-2} main-effect design			extended 2^{5-1} design		
	Total*	most important variable	least important variable	Total*	most important variable	least important variable
Homogeneous	90.7 *(19.8)*	17.4 *(5.6)*	17.8 *(4.9)*	38.0 *(8.2)*	7.7 *(1.5)*	10.1 *(4.0)*
Heterogeneous	90.0 *(17.7)*	17.9 *(6.1)*	19.2 *(6.3)*	35.7 *(7.4)*	7.3 *(1.7)*	9.1 *(3.5)*
Dominant	123.3 *(19.0)*	33.6 *(5.7)*	23.1 *(6.9)*	57.8 *(8.1)*	17.5 *(3.2)*	10.6 *(2.7)*

*) The total estimation interval is calculated as the sum of the absolute sizes of the estimation intervals of the single variables.

In comparison between main-effect and extended designs, the size of the deterministic estimation interval is greatly reduced by increasing the number of stimuli. There are two reasons for this: First, the average utility difference of two rank neighbouring stimuli becomes smaller and, accordingly, the limits for possible solutions. Second, the more stimuli are used, the more inequalities determine the estimate. In case of error-free data, both effects narrow the size of the estimation interval and center it around the true mean. This explains Colberg's findings as outlined in section 2.2.

In sum, the interval scale properties of ranked data need to be questioned. However, we agree with Colberg's finding that extended designs provide - in case of error-free data - estimates which approximate those gained from interval-scaled data. As shown below, the differentiation between both effects is useful, because it contains valuable information in the case of stochastic data.

Proposition 2: Evaluation procedures should take ambiguity of effect estimates into account and thus should provide estimation intervals.

2.4 Relationship between error-free and empirical rankings

Empirical rank data are not error-free translations of the true metric utility values but are characterized by a stochastic error term. Thus it is unlikely that the observed empirical rankings are identical with the „true", error-free rankings. The observed ranking however is the only information provided by respondents. Both the number and the location of rank reversals are unknown and are implicit within the estimation procedure. This constitutes a severe problem, because an observed ranking may be the result of different errors from different „true" rankings and thus from different underlying utility functions.

For example, an observed faulty ranking of a „true" utility function may be identical with an error-free ranking of another (and thus „false") utility function. This point is elaborated through a simulation of the two most-likely errors, i.e. the reversals between those ranks with most similar metric utility values[5]. Table 3 shows the (non-representative) results of this assessment. It can be seen that the ability to detect even simple rank reversals is by no means guaranteed and that it varies largely. This can be explained by the (non)existence of redundant information: A rank reversal remains hidden, when it is the only inequality which provides the specific information. Reduced designs possess small degrees of freedom and thus next to no redundant information. Accordingly, the ability to detect rank reversals remains low even if more than one rank reversal occurs. In contrast, extended designs provide multiple redundant information. However, the larger their estimation intervals are (Table 2), the less able are they to locate single rank reversals, since single variable estimates can be realigned to a large extent without affecting the entire rank order.

Table 3: Ability to detect a rank reversal for different designs, types of preference function and error-components as measured by the % of observations which lead to a fully consistent ranking

Type of Preference Function:	2^{5-2} main-effect design		extended 2^{5-1} design	
	1 Reversal	2 Reversals	1 Reversal	2 Reversals
Homogeneous	80%	72%	31%	14%
Heterogeneous	88%	77%	20%	8%
Dominant	92%	85%	43%	17%

Empirical data very often could have more than one rank reversal. Thus, inconsistencies are likely to remain unsolved. In fact, a Kendall's tau of 0.8 is often seen as sufficient to classify the observed rankings as internally valid, which corresponds to 3 identified reversals of rank pairs in the case of 8 stimuli and 12 in the case of 16 stimuli. The larger the unresolvable inconsistencies, the less determined the estimate becomes, because the location of reversals remains ambiguous. Identified rank reversals may in fact be correctly ordered rank pairs whereas as correctly identified rank pairs may constitute hidden rank reversals. Such a misleading interpretation of observed rankings is more likely if the number of possible rank reversals increases. Thus, the number of possible „true" rankings and of underlying „true" utility functions are increased with larger number of stimuli.

In sum, it can be concluded that it is unlikely that all rank reversals are accurately identified. This will inevitably lead to defects of both the estimation procedure and of the applied goodness-of-fit measures when they rely on the fit between estimated and observed rankings. Based on the available information an incorrect utility function might be estimated and even a perfect fit might be falsely diagnosed.

Proposition 3: Evaluation procedures should take the inability to detect rank reversals into account. Therefore, they should not utilize the fit between estimated and observed ranking as the only criterion for assessing estimation outcomes.

2.5 Stochastic model

According to the basic idea of conjoint analysis, respondents should rank stimuli using a holistic and comparative assessment of stimuli utilities in accordance with their „true" utility function. Thus, it can be assumed that respondents assess stimuli utilities holistically with an error term centered around their true means and subsequently articulate their ranking in accordance to these judgments. This can be specified by adding a normal-distributed error term to the metric stimuli utility values. Other distributional patterns may be thought of, however, the normal distribution is in line with a comparable model on rating scales (Srinivasan, Basu, 1989) and can base on the arguments elaborated therein.

Figure 2 provides an illustration of this error model, which has been intuitively used in many simulation analyses of rank-order conjoint data (e.g. Carmone, Green, 1981, Wittink, Cattin, 1981, Umesh, Mishra, 1990, Darmon, Rouziès 1991, 1994). It shows that the probability of observing a rank reversal can be assessed by calculating the integral of overlapping normal distributions. Given that the probability distributions of more than two neighbouring stimuli are overlapping, a joint probability integral has to be calculated. This dependence of probabilities is a point of major difference between ranked conjoint data and rating data, in which the assessment of one stimulus would be independent of all others (Srinivasan, Basu, 1989).

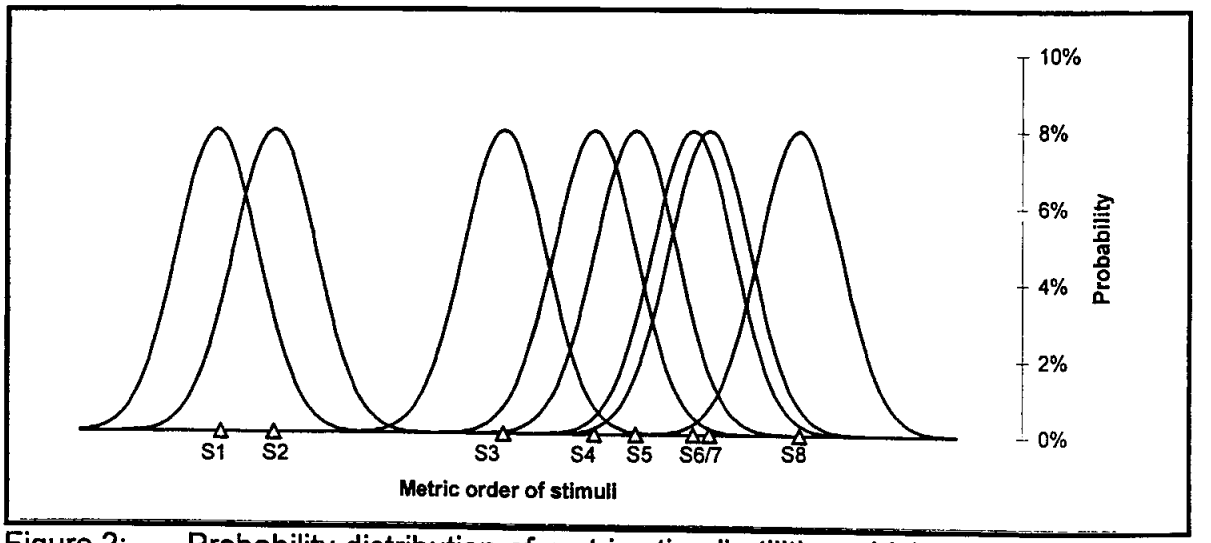

Figure 2: Probability distribution of metric stimuli utilities which constitute an observed ranking

Different assumptions about the decision making process and the error term are to be found in travel behavior research (Hensher, 1994): It is assumed that respondents rank the stimuli from top to bottom by repeatedly choosing the most preferred stimulus. The chosen stimuli is excluded and the choice procedure is repeated to the last. Based on this model, the „Luce and Suppes Ranking Choice Theorem" (Luce, Suppes, 1965) handles a ranking as a sequence of independent choice tasks. This justifies applying the estimation *method of rank explosion* (Chapman, Staelin, 1982), which decomposes ranked data into subsets of independent choices. In contrast, our model would lead to dependent choice tasks[6], which would see the method of *rank-explosion* as falsely diagnosing increased error variance in lower ranks (see analogous empirical findings in: Ben-Akiva et al., 1992; Bradley, Daly, 1994).

Comparing both models, the Luce and Suppes approach seems to be less adequate for describing respondents´ behavior. It would require respondents to reassess subsets of the n stimuli over and over again, leading to ½*n(n+1)-1 comparisons, while simultaneously forgetting about possible misperceptions in the former stages and neglecting already ranked stimuli. This sees respondents as very persistent and extremely forgetful at the same time. In addition, respondents would neither be allowed to pre-sort the stimuli nor to re-evaluate their ranking, despite the fact that both aspects are often encouraged in interviewers´ guidelines in order to enhance the reliability of answers (e.g. Schrader, 1990, Teichert, 1993).

The normal distributed model better reflects the hypothesized process of respondents´ assessments. In fact, Figure 2 can be viewed as an analogy to a desirable process of respondents´ assessment. The location of stimuli´s probability distributions may be seen as the outcome of presorting the stimuli on a continuous line. The respondent now refines his judgment by comparing the utilities of neighbouring stimuli, considering the better and the worse alternatives. Each stimulus is given exactly one perceived utility value and stimuli are ranked according to their resulting order on the x-axis.

Overall, the model of normal distributed stimuli evaluations is in accordance to the basic idea of Conjoint-analysis. Thus, the further analysis will be based on this error model. This allows us to calculate probabilities to observe any possible ranking under the premise of a given metric preference function and a distributional assumption. Since an exact calculation would require extensive enumeration of overlapping integrals from normal distributions, we recommend and apply an approximation shown in the appendix.

Given a normal distributed error model, rank reversals are to be expected, when neighbouring ranks possess similar metric utilities, and they are unlikely, when neighbouring ranks are highly different. Thus the probability of observing rank reversals depends not only on the error term but also on the distribution of metric stimuli utilities, which deviate between different experimental settings (section 2.2). The true ranking has the highest probability of occurrence, when the stimuli are distributed equally, i.e. interval-scaled. This joint probability is cut in half, when one rank pair is nearly identical, and divided by four, when two rank pairs are similar. Thus the probability of observing the true ranking is highly sensitive to the number of minor stimuli differences. To provide a sketch of this effect, probability values are calculated for the utility functions of the simulation example. Table 4 shows that there are high variations in the probability of observing the true ranking, which can be traced back to different causes:

- The size of the error-term has a non-linear negative effect on the probability of observing the „true" ranking. The metric differences between neighbouring ranks build threshold-values: the ranking remains unchanged as long as the error term does not exceed the smallest utility difference of a rank pair.
- The probability of observing the „true" ranking diminishes with increased number of stimuli. Since a larger number of stimuli leads both to more rank pairs and to smaller utility differences between rank neighbors, it is more likely to observe at least one rank reversal.
- Lower probabilities are observed for the homogeneous type of preference function. Here, the probability of observing a rank reversal is higher despite

similar interval-scale properties, because this type possesses the highest range of differences (Table 1) and thus is likely to contain similar rank pairs.

- The high standard deviation within the single cells shows that the casewise distribution of metric stimuli utilities highly influences the probability of obtaining the „true" ranking. Thus individually different rank recoveries are to be expected even in identical experimental settings.

Table 4: Probability of observing an error-free ranking (standard deviation)

Type of preference function	2^{5-2} main-effect design size of ND(0,σ) error term σ = 5	σ = 10	σ = 20	extended 2^{5-1} design size of ND(0,σ) error term σ = 5	σ = 10	σ = 20
Homogeneous	43.2% *(20.7%)*	18.2% *(10.1%)*	3.5% *(1.9%)*	0.4% *(0.6%)*	<0.1% *(<0.1%)*	<0.1% *(<0.1%)*
Heterogeneous	49.3% *(19.4%)*	22.0% *(9.6%)*	4.7% *(1.9%)*	3.8% *(4.4%)*	0.1% *(0.2%)*	<0.1% *(<0.1%)*
Dominant	47.8% *(22.7%)*	23.7% *(12.3%)*	6.3% *(2.8%)*	4.7% *(5.9%)*	0.2% *(0.3%)*	<0.1% *(<0.1%)*
TOTAL	46.7% *(21.0%)*	21.3% *(11.0%)*	4.8% *(2.5%)*	2.9% *(4.6%)*	0.1% *(0.2%)*	<0.1% *(<0.1%)*

In sum, both the probability of observing the true ranking and the number of rank reversals do not only depend on the size of the error term but also on the metric differences between neighbouring stimuli. Thus, the ranked data do not contain all information which could be utilized by the evaluator. The stochastic model and the pattern of possibly underlying metric stimuli utilities provide further information.

Proposition 4: Evaluation procedures should relate the ranking information to the probable pattern of underlying metric utilities.

Based on this proposition, evaluation methods should not only aim at replicating the observed ranking, i.e. minimize the number of rank reversals. Since the observed ranking should be looked at as a sequence of equally normal-distributed comparisons of rank pairs, consideration should be given to the joint probability of observing both the estimated correct rank pairs and the rank reversals.

Alternative solutions should be valued accordingly. An estimate, which diagnoses rank reversals of similar stimuli and correctly classifies highly different rank pairs, should be preferred to an estimate, which diagnoses a cluster of similar stimuli without a rank reversal. The latter estimate is questionable, since similar stimuli are likely to result in some rank reversals, while there is a high probability of obtaining the combination of rank reversals and correctly classified rank pairs from the first solution.

In sum, evaluation procedures should apply a two-sided optimization procedure which maximizes the joint probability of observing the ranking: (a) observing the correct rankings and (b) observing the rank reversals.

Proposition 5: Evaluation procedures should be based on the total probability of observing the actual ranking, that is the probability of observing the specific combination of rank reversals and correctly ordered rank pairs.

3 Implications for evaluation

The observed ranking constitutes the only information provided by respondents. Accordingly, estimation procedures and their goodness-of-fit measures are directed towards replicating the observed ranking through the estimated one. In the following section, it will be analyzed whether or not the methods hereby adequately reflect the propositions of the model of ranked Conjoint data.

Since the „Luce and Suppes Ranking Choice Theorem" already showed to be not in concurrence with the stochastic model applied here (section 2.5), the choice-based evaluation techniques are excluded. Instead, the analyses focus on the metric ordinary least squares regression analysis (OLS), the most often used evaluation method (Wittink et al., 1994), which is based on a well-known robustness (e.g. Carmone et al., 1978, Cattin, Bliemel, 1978), and on the statistically more adequate, non-metric Linear Programming Approach LINMAP (Shocker, Srinivasan, 1977). These two methods represent the metric and the non-metric evaluation alternatives. The findings can be easily applied to variations of these estimation approaches.

Simulation analysis is applied to assess implications of revealed shortcomings on the validity of estimates. This allows us to benchmark estimation outcomes against otherwise unknown true metric utility functions. To be specific, (total) estimation accuracy is measured as the (sum of) deviation(s) of estimated variable part-worth values from their true metric part-worth values. Findings are compared against commonly used validity proxy measures of goodness-of-fit. Subsequently, the influence of stochastic pattern and of the methods on estimation performance is analyzed.

3.1 Adequacy of goodness-of-fit measures

Commonly used goodness-of-fit measures, as Kendalls τ (Kendall, 1962) or Spearman's ρ, do not fulfill the propositions of the model of ranked Conjoint-Data: They strictly rely on the ranking information (which is against proposition 4) and disregard differences in the quality of observed rank reversals (which is against proposition 1). These procedures would be adequate if two rankings were compared in isolation. However, since the estimated ranking is based on estimated metric utility values, the rank comparison implicitly applies a less sophisticated error model for the estimated metric utility values which is equivalent to an interval-scale assumption[7]. This causes goodness-of-fit measures to neglect valuable information. These measures do not differentiate between solutions which lead to identical rank recoveries (which is against proposition 3) but which possess different patterns of underlying metric utilities (which is against proposition 5). Thus goodness-of-fit does not constitute an exact measure of ranking consistency, but a robust approximation of it.

A more severe shortcoming stems from the basis of comparison: Goodness-of-fit measures relate the observed to the estimated ranking and not the true ranking. Since rank reversals may remain hidden (proposition 2) and estimation procedures tend to nearly maximize the fit between observed and estimated rankings (see section 3.4), we expect goodness-of-fit measures to systematically overestimate the true consistency. Simulations are performed in order to assess this shortcoming. The findings (Table 5) show that the more rank reversals occur, the more the estimated fit[8] overestimates the consistency

of the true ranking. Similar findings were found in a study of Stallmeier (1993) on graded paired comparisons and on rating data: Commonly used goodness-of-fit measures consistently overestimate the true fit and sometimes led to measures twice as high. The fit between observed and estimated ranking can at best serve as an upper limit of the consistency of answers, but should not be regarded as an unambiguous indicator. Since all of these shortcomings are inherent in commonly used indices of validity and reliability, it is questionable whether they reflect the true state of nature correctly[9].

Corollary 1: Goodness-of-fit measures provide ambiguous clues on the consistency of observed rankings.

Table 5: „True" versus estimated goodness-of-fit using OLS regression: Mean Kendalls τ (standard deviation)

	Intervall classes for true goodness-of-fit (Kendalls τ)			
	τ < .70 (n=66)	.70 < τ <= .80 (n=174)	.80 < τ <= .90 (n=329)	.90 < τ <= 1.0 (n=331)
True goodness-of-fit	0.63 *(0.06)*	0.75 *(0.03)*	0.85 *(0.03)*	0.94 *(0.03)*
Estimated goodness-of-fit using OLS	0.76 *(0.09)*	0.84 *(0.08)*	0.92 *(0.06)*	0.95 *(0.04)*
Mean difference b/w estimated and true fit	0.13	0.09	0.07	0.01
Minimum difference	-0.04	-0.03	-0.13	-0.18
Maximum difference	0.50*	0.26	0.19	0.10

*) In the extreme case, the true Kendalls τ equals 0.317, while the Kendalls τ value based on the estimate equals 0.817 and thus could even be considered as fulfilling minimum requirements in regard to consistency.

3.2 Accuracy of estimated internal validity

The goodness-of-fit is usually interpreted not only as a measure of respondents' consistency but also as an indicator for the internal validity of the derived estimate (Müller-Hagedorn et al., 1993). In this sense, goodness-of-fit is a two-fold proxy measure, since the objective of the estimation procedure is not to reshape the observed, stochastic ranking but to assess the metric utility

function, which underlies the unknown error-free ranking. However, as Table 5 shows, a reasonable fit between observed and estimated ranking may coincide with a bad fit between observed and error-free, true ranking. Thus, recovery of the observed ranking and recovery of the underlying utility function does not necessarily coincide. Accordingly, we expect goodness-of-fit measures to provide a distorted picture of the internal validity.

The model of ranked Conjoint-data could cover more shortcomings of the goodness-of-fit as an indicator of internal validity. This stems from the dual nature of ranked data, in which rank reversals may contain valuable information for estimating the underlying metric utility function by revealing those rank pairs which are of similar utility (see section 2.5). Rank recovery can be achieved also by applying an estimation procedure which is not based on the model of ranked data. Recovery can be gained even from random data (e.g. Mullet, Karson, 1986; Weisenfeld, 1989). Finally, there exists a well known non-comparability of goodness-of-fit measures across different experimental settings (Green, Wind, 1973; Umesh, Mishra, 1990).

Our simulation findings (Table 6) are in line with this. They show that there is no clear relationship between metric estimation accuracy and achieved rank recovery, measured by Kendalls τ. A high number of rank inconsistencies decreases estimation accuracy under low-error conditions but enhances it under high-error conditions. This is because few reversals of highly different rank pairs are likely to cause a higher distortion of estimation outcomes than many of them, as stochastic misrankings are likely to offset each others influence on estimation outcomes.

Even if goodness-of-fit measures are used only to identify outliers they deliver information of questionable quality. For example, a cut-off-level of Kendalls τ = 0.8 divides the simulated ND(0,20) extended design subsample into 112 rejected observations with an average deviation of 22.1 utility points and 188 accepted observations with average deviation of 25.8 utility points between estimated and „true" utility function.

In sum, the concept of measuring internal validity by the proxy goodness-of-fit has to be questioned. It can neither validate the applied utility model nor can it

provide a safe-guard against unreliable answers. Furthermore, methodological studies which base their evaluations on improved goodness-of-fit measures may provide misleading conclusions, since maximizing goodness-of-fit does not necessarily concur with achieving superior estimation accuracy.

Corollary 2: Ranking consistency is an inadequate proxy of internal validity.

Table 6: Comparison of Goodness-of-fit measures under different error terms

Mean values (n=300) *(standard deviation)*	2^{5-2} main-effect design size of ND(0,σ) error term				extended 2^{5-1} design size of ND(0,σ) error term			
Preference function	σ = 0	σ = 5	σ = 10	σ = 20	σ = 0	σ = 5	σ = 10	σ = 20
Kendalls τ								
True	1.00 *(0.00)*	0.93 *(0.07)*	0.90 *(0.08)*	0.80 *(0.13)*	1.00 *(0.00)*	0.91 *(0.06)*	0.86 *(0.07)*	0.76 *(0.09)*
OLS	0.99 *(0.02)*	0.98 *(0.03)*	0.97 *(0.05)*	0.94 *(0.07)*	0.99 *(0.01)*	0.92 *(0.05)*	0.89 *(0.06)*	0.81 *(0.07)*
LINMAP	1.00 *(0.00)*	0.98 *(0.03)*	0.97 *(0.05)*	0.94 *(0.07)*	0.99 *(0.01)*	0.94 *(0.04)*	0.90 *(0.05)*	0.84 *(0.06)*
Total deviation DEV *								
OLS	23.2 *(9.6)*	23.9 *(9.8)*	24.9 *(10.2)*	31.9 *(13.3)*	25.5 *(9.4)*	22.9 *(8.8)*	22.1 *(8.4)*	24.4 *(9.0)*
LINMAP	26.8 *(15.0)*	27.5 *(13.4)*	29.6 *(13.6)*	36.2 *(15.4)*	22.0 *(9.8)*	19.0 *(10.1)*	17.6 *(9.1)*	21.6 *(9.8)*
Korrelation between Kendalls τ and DEV								
OLS	n.a.	0.03	-0.01	-0.05	n.a.	-0.21	-0.17	0.17
LINMAP	n.a.	0.12	0.00	-0.03	n.a.	-0.20	-0.18	0.16

*) The total deviation measures metric estimation accuracy and is calculated as the sum of the absolute deviations of the metric variables' part-worth estimates from their true values.

3.3 Influence of response quality

The simulation results listed in Table 6 show that the three error-levels ND(0,5), ND(0,10) and ND(0,20) possess realistic dimensions: Comparing the resulting

estimated goodness-of-fit measures with cut-off levels applied in real-life conjoint applications (e.g. Mullet, Karson, 1986; Sattler, 1994), the upper error level seems to simulate a sample with bad consistency, the medium error level a reasonable one and the lower error level one with high consistency. However, even a medium error level of ND(0,10) leads to confidence intervals of metric stimuli utilities which can highly exceed the differences between rank pairs[10]. This indicates a high noise in empirical Conjoint-data and suggests a high influence of the error term on estimation accuracy.

In contrast it can be seen from Table 6 that the accuracy of preference function estimates remains nearly stable across different error levels. An error term of up to 10 points exerts only minor influence on the quality of estimation outcomes[11]. The higher number of rank reversals, documented by lower Kendalls τ values, do not affect average estimation accuracy. More interestingly, even the variations in estimation accuracy stays nearly stable. This shows that not only aggregated but also individual estimates possess similar quality at different error-levels. The evaluator should thus not be overly concerned about the consistency of the observed ranking.

The estimation accuracy of the extended design even improves slightly when adding an error term (Table 6). This is because minor rank reversals provide information by revealing those rank pairs which are of similar utility. This, however, requires rank reversals to be detected and interpreted as such. Accordingly, the information of the error component can be used better in extended designs, because rank reversals are more likely to remain hidden within the main-effect design (see section 2.4). Similarly, LINMAP is better able to utilize the stochastic information, since it does not force observed rankings into interval-scaled differences (see section 3.4).

Corollary 3: Estimation accuracy is insensitive to the size of the ND error term and thus to variations in the quality of responses.

Even in case of error-free conditions, the estimated preference functions deviate visibly from the true ones, with the sum of part-worth estimates lying off from their true values by an average above 20 points. These limitations in

estimation accuracy can be attributed to inherent shortcomings of ranked data and of the applied estimation procedures.

Furthermore, the estimation accuracy gained from the extended design does not deviate largely from the one gained from the main-effect design. This indicates that the estimation procedures do not utilize the information provided by the ranking in an efficient manner. Otherwise, a significantly increased estimation accuracy would be expected in case of applying the extended design, since this design delivers twice as much information and accordingly cuts the size of the estimation interval nearly in half (Table 2). The controversial findings point at shortcomings of the estimation procedures.

3.4 Adequacy of estimation methods

Traditional estimation procedures deliver only point estimates. They do not contain information on the size of the estimation interval (which is against proposition 2). The estimation procedures which are based on different propositions about the basic principles of ranked data use different search algorithms and, thus, lead to divergent estimates. Thereby, neither estimation techniques fulfills all of the required traits postulated by the model of conjoint data.

The metric estimation technique OLS assumes use of interval-scaled ranked data (which is against proposition 1). The observed ranks are simply interpreted as metric numbers. An unique solution is achieved by minimizing the squared sum of deviations between estimated and observed metric values. Thus, each rank reversal is valued as equally important (which is against proposition 4), and a rank reversal encountering a skipped rank is valued as more important than the sum of two reversals of neighboured ranks. In minimizing the sum of squared errors, information from both rank reversals as well as from correctly ordered rank pairs determines estimation outcomes (proposition 5). In this regard, the OLS algorithm departs from a mere replication of the observed ranking (proposition 3) but may falsely diagnose rank reversals even in error-free conditions (see Table 6).

LINMAP does not state assumptions on the scale properties of the ranked data (proposition 1). It uses the method of linear optimization to minimize the sum of metric corrections which have to be added to force the estimated error-free stimuli values into the observed rank order. LINMAP will provide an estimate with no rank reversals, as long as there is such a solution, since in this case the necessary metric corrections are zero and, thus, minimized. Accordingly, it is likely that rank reversals are falsely resolved (which is against proposition 3). However, even in error-free conditions a perfect fit is not necessarily achieved, since ties are neglected (Albers, 1984). Under error-free conditions, LINMAP leads (in 192 out of 300 cases) to an estimate which falsely identifies ties in the extended design (Table 6).

In minimizing the metric distance of any rank reversal, LINMAP utilizes the information content from different distances between rank pairs (proposition 4). However, the distance between two correctly classified stimuli is not analyzed (which is against proposition 5). Thus the algorithm ceases as soon as it finds one of the possible solutions. The location of this solution within the estimation interval is arbitrary (which is against proposition 2) and depends on the initial values given by the algorithm[12].

The comparison between OLS and LINMAP reveals strength and weaknesses of both methods, which requires a differentiated assessment of their relative performance: OLS wrongly bases on the interval-scale proposition, but correctly bases its evaluation on the entire ranking, i.e. it considers the patterns of metric differences between all rank pairs simultaneously. From this basic differences between both estimation procedures we can derive suppositions on their relative performance.

The interval-scale assumption causes OLS to overvalue the information provided from closely neighbouring ranks. Furthermore, the squared term causes skipping of ranks to exert a high impact on estimation outcomes. Highly distorted estimates are thus to be expected, if closely neighbored rank pairs cause skipped ranks. Accordingly, the propositions of OLS seem to be best

satisfied in case of small designs and small error terms, because a skipping of closely neighbored ranks is less likely to occur under these conditions.

LINMAP, on contrary, can be expected to underperform under the above mentioned conditions. In cases of relatively small numbers of rank reversals, LINMAP provides less determined solutions because it does not fully utilize the information of correctly ordered rank pairs. This leads to an arbitrary choice of any solution within the larger estimation intervals of smaller designs (section 2.3). Furthermore, LINMAP is likely to resemble suboptimal solutions which avoid rank reversals but cluster similar stimuli against more realistic solutions, which encounter rank reversals of similar stimuli and correctly classified stimuli of larger metric differences.

In order to assess the implications of these shortcomings on estimation performance, our simulation findings are split by type of preference function. This allows us to localize the origin of the (non-representative) performance differences of the estimation methods found in the aggregated analysis (Table 6). Since similar patterns could be found for all error levels, Table 7 summarizes only the findings of the deterministic case for ease of reference (see section 2.3). These results and further detailed analyses show that the relative estimation accuracies of OLS and LINMAP behave as expected:

a) The accuracy of the OLS-estimate turns out to depend highly on the adequacy of the interval-scale assumption. A correlation analysis between estimation accuracy and coefficient of variation of rank pairs' utility differences reveals a correlation coefficient of r=0.49 for OLS across all 600 deterministic observations. There is no such correlation if LINMAP is used (r=0.04), because LINMAP does not rely on the interval-scale assumption. Correspondingly, Table 7 shows that OLS is less able to utilize the information provided by additional observations, if interval-scale properties are not improved. Accordingly, the accuracy of the OLS-estimate on average even deteriorates in extended designs, if the observations stem from homogenous preference functions, because interval-scale properties are not achieved (section 2.2).

b) The accuracy of the LINMAP-estimate is strongly affected by the size of the estimation interval (see section 4 for description of the calculus) with a correlation coefficient of r=0.55 across all deterministic observations. In contrast, the OLS-algorithm tends to center its estimate within the estimation interval, which leads to less dependency of estimation accuracy on the size of the estimation interval (r=0.33). As observed in other simulation studies (Darmon, Rouziès, 1991, 1994), the LINMAP estimate underperforms in case of main-effect designs which possess especially large estimation intervals. It improves in extended designs the more the size of the estimation interval is reduced, and thus most in case of dominant preference functions (Table 2).

Table 7: Estimation accuracy of estimation methods in the deterministic model: Mean deviation from true part-worth values (standard deviation)

	2^{5-2} main-effect design			extended 2^{5-1} design		
Type of preference function	Total*	most imp. var.	Least imp. var.	Total	most imp. var.	least imp. var.
OLS						
Homogeneous	21.3 *(8.0)*	-0.1 *(4.4)*	-1.2 *(5.1)*	29.8 *(6.4)*	10.2 *(2.5)*	-10.5 *(2.4)*
Heterogeneous	19.0 *(7.1)*	0.9 *(5.0)*	-0.4 *(5.0)*	18.6 *(8.0)*	6.0 *(3.6)*	-6.0 *(3.3)*
Dominant	29.2 *(10.0)*	-8.4 *(6.3)*	0.0 *(3.4)*	27.0 *(9.6)*	-10.3 *(7.9)*	-2.9 *(2.9)*
LINMAP						
Homogeneous	22.3 *(10.0)*	-1.9 *(4.9)*	1.1 *(5.3)*	23.5 *(6.3)*	7.7 *(2.6)*	-8.2 *(2.4)*
Heterogeneous	19.2 *(9.0)*	-1.9 *(5.4)*	0.9 *(4.4)*	14.3 *(6.7)*	3.5 *(3.4)*	-4.0 *(3.3)*
Dominant	38.9 *(15.8)*	-15.5 *(8.4)*	3.3 *(5.4)*	28.2 *(10.1)*	-9.5 *(8.3)*	-3.5 *(2.8)*

*) The total deviation measures metric estimation accuracy and is calculated as the sum of the absolute deviations of the metric variables' part-worth estimates from their true values.

In sum, the revealed shortcomings of traditional estimation methods showed to exert a higher influence on the accuracy of estimated preference functions than

either the design applied or the quality of answers being obtained. The relative estimation performance can thus be related to failures in fulfilling the requisite traits of the model of ranked Conjoint-data. This hints towards the existence of systematic shortcomings of traditional estimation procedures.

Corollary 4: Estimation accuracy can be strongly influenced by shortcomings of estimation procedures.

3.5 Accuracy of estimated part-worth values

Given that estimated part-worth values deviate from the true ones as outlined above, it is of interest whether there is some systematic bias in the estimates. Hence, Table 7 contains information on the least and the most important variables. Individual-level estimates are on average unbiased given that the mean estimate does not deviates from the mean true value. Shortcomings may then distort individual-level analyses, but do not affect aggregated outcomes.

This, however, does not hold true for all variables in the experimental conditions applied. First of all, the dominant variable is strongly underestimated regardless of design applied or estimation method chosen. This is not surprising, as the experimental design assumes a compensatory utility function and as it is not designed to provide information on the extent of dominance. The robustness of conjoint-experiments against violations of the linear-additive, compensatory preference function thus has its limitations. It is unlikely that the influence of a dominant variable is accurately estimated.

The least important variable is underestimated, if the extended design is applied. Since there are analogous findings to this in other studies, it seems that the less important variables are systematically underestimated (Darmon, Rouziès, 1991, 1994; Teichert, 1996). This should cause us to reconsider the origin of estimation differences to self-explicated data, which typically put more emphasis on less important variables (e.g.: Schoemaker, Waid, 1982). It may well be the case, that the empirically observed differences do not only stem from psychological effects of the evaluation task, but that they are also caused by inherent methodological properties of decompositional ranked data

Simultaneously, the most important variables tend to be overestimated. In combination with the underestimation of less important variables, conjoint estimates tend to augment the differences in variable importance weights. This effect is most strongly observed in case of homogeneous preference functions. Here, the range of OLS-estimated variable importance weights exceeds those of the simulated „true" preference functions by 152% (LINMAP: 116%) in case of the extended design. While the extent of this bias is based on the chosen experimental conditions[13], similar findings are documented for narrow preference function ranges in a simulation study of Darmon, Rouziès (1989)[14]. Thus, respondents which put equal weight on each variable can hardly be revealed. Conjoint-estimates are likely to identify distinct preference functions even if respondents are indifferent between variables.

Finally, differences between individual preference functions show to be overestimated as well. A comparison of the differences of individual utility functions provides evidence for this systematic bias: In case of the extended design, the differences of the simulated „true" homogenous preference functions are only half as high as their estimated values[15]. Such patterns may cause distortion of aggregated analyses, e.g. clustering techniques may identify market segments which do not exist.

The revealed distortions relate to the experimental conditions applied and do not necessarily constitute generalizable patterns. Different experimental conditions may lead to different distortions. However, for the conditions chosen it could be shown that the distortions of individual estimates do not level out on average but lead to systematic differences. Thus, the findings lead us to:

Corollary 5: Shortcomings of estimation procedures can cause systematic differences between true and estimated utility function.

3.6 Summary of findings

Table 8 provides a summary of the findings on the adequacy of traditional evaluation methods with the model of ranked Conjoint-data. From this it can be seen that LINMAP and OLS possess in many aspects complementary strengths

and weaknesses. Accordingly, the analyses revealed a changing superiority: OLS showed to perform better in case of smaller designs and smaller error term; while LINMAP outperformed OLS in case of larger designs and larger error terms.

By and large, these traditional evaluation methods fail to meet most of the requisite traits postulated before. Thus the theoretical basis of the evaluation methods possesses major shortcomings, if the proposed model of ranked Conjoint-data holds true.

Table 8: Adequacy of traditional evaluation methods with ranking model

Propositions of ranking model for evaluation procedure:	OLS-technique	LINMAP	goodness-of-fit measures
(1) Do not rely on interval scale	NO	YES	NO
(2) Consider ambiguity of estimates	NO	NO	NO
(3) Consider missing ability to detect rank reversals	(not explicitly)	NO	(NO)
(4) Utilize information of underlying metric pattern	(not in full concurrence to error model)		NO
(5) Base evaluation on entire ranking	(YES, but based on interval scale)	NO	NO

Simulation analyses showed two major areas of resulting shortcomings which in combination question the adequacy of traditional evaluation outcomes: First, goodness-of fit measures are likely to fail in serving as a quality measure, since they neither constitute an adequate proxy of internal validity (corollary 2) nor do they even provide unambiguous clues on the consistency of observed rankings (corollary 1). Second, the estimation accuracy of traditional estimation methods is highly questionable. It does not depend as much on the size of the error term (corollary 3) but results from shortcomings of the estimation procedures (corollary 4), which even may cause systematic biases (corollary 5). In combination, the evaluator is confronted with a high degree of inherent

uncertainty about the determinacy of estimates gained from ranked-based Conjoint analysis.

4 Possibilities for improvement

As a main shortcoming, traditional estimation procedures do not adequately reflect ambiguity of effect estimates. They deliver only point estimates, which are not strictly derived from the model of ranked Conjoint-data but based on arbitrary assumptions. However, in order to justify choosing any of the possible solutions, it is necessary that this one possesses superior qualities over the others. Hence, in the following section it will be analyzed whether or not a superior estimate can be gained based on the information available from the ranking.

In order to extract the entire information available from ranked Conjoint-data, one has to reconsider its very basic nature: One knows neither the true utility function nor the true, error-free ranking, but rather observes a stochastic ranking. The best-fitting point estimate is thus the one which leads to the observed ranking with the highest probability under a given error term. A prerequisite for gaining a best-fitting point estimate is that the estimated utility functions lead to the observed ranking with significantly different probabilities and, vice versa: It is necessary that an estimated preference function leads to a particular ranking with a higher probability than to any other ranking.

As shown before (section 2.3), the ranked data provide estimation intervals. In the absense of an error term, the alternative solutions are all equally possible, as long as they meet the requirements of the ranking. Thus it is impossible to derive a best-fitting point estimate for error-free data. The same holds true for stochastic data, if the error model is unknown.

Under a given error model and size of the error term it becomes possible to calculate the probability that an estimate leads to the observed ranking. Hence a two-stage procedure can be applied: First, a stochastic estimation interval can be evaluated. Subsequently the probability of obtaining the observed ranking can be calculated for each possible solution using the approximation method

described in the appendix. However, a sketch on methodological properties of stochastic rankings questions the utility of this approach:

a) It is only possible to derive a best-fitting point estimate once the size of the error term is known. It is unlikely that one and the same solution fits best under various sizes of the error term. This is due to the non-linear influence of the error-term. Only if one of the solutions leads to perfect interval-scaled differences between rank pairs, then this one may prove to be the overall best solution.

b) Preference functions may lead to a variety of possibly observed rankings with nearly equally high probability: According to the error model, out of all possibly observed rankings the deterministic estimate always has the highest probability of being observed[16]. Thus, its probability can serve as an upper limit to other probabilities and its inverse serves as an estimate of the minimum number of possibly observed rankings. Going back to the example of Table 4, we expect to observe, on average, more than 200 different rankings with less than 0.1% probability each in the extended design, given an medium-level error term. Thus, the probability of observing one specific rank order approaches zero both with increasing number of stimuli and with increasing size of the error term. In this case, the probability distributions can not provide clues on the superiority of alternate solutions. Point estimates may remain arbitrary.

Since it is unlikely to identify a best-fitting estimate, the center of gravity of each variables' estimation interval can be considered as an alternative, robust point estimate. In this case, the middle of variables' estimation intervals serves as point estimates for the specific variable. This conservative approach is common practice in operations research (Davidson, Suppes, 1957). It limits the consequences of potentially wrong estimates, because the maximum possible estimation error is minimized.

In order to assess the benefits of this procedure, stochastic estimation intervals are evaluated analogous to the proposed heuristic method (see section 2.3)[17]. Table 9 lists the estimation intervals found under the various error levels. From this it can be seen that the size of the estimation intervals is influenced only

slightly by the quality of answers obtained. While the random effect narrows the estimation interval of main-effect designs slightly, it widens the estimation interval in extended designs. Overall, the findings show (section 3.3) that the stochastic pattern does not highly influence the achievable estimation accuracy. This was expected. Accordingly, the consequences of estimation ambiguity on point estimates (section 3.6) hold true for Conjoint-analyses, independent of the quality of answers obtained.

Table 9: Limits of stochastic estimation intervals

Mean total estimation interval* *(standard deviation)*	2^{5-2} main-effect design size of ND(0,σ) error term				extended 2^{5-1} design size of ND(0,σ) error term			
Preference function	σ = 0	σ = 5	σ = 10	σ = 20	σ = 0	σ = 5	σ = 10	σ = 20
OLS								
Homogeneous	90.7 *(19.8)*	87.4 *(29.1)*	78.1 *(27.3)*	100.2 *(31.3)*	38.0 *(8.2)*	47.7 *(11.4)*	46.2 *(15.8)*	51.4 *(16.1)*
Heterogeneous	90.0 *(17.7)*	93.7 *(21.1)*	101.8 *(21.4)*	104.1 *(33.0)*	35.8 *(7.4)*	47.4 *(11.5)*	46.7 *(13.9)*	53.0 *(16.0)*
Dominant	123.3 *(19.0)*	115.8 *(27.0)*	88.7 *(19.0)*	120.0 *(27.4)*	57.8 *(8.1)*	53.9 *(18.2)*	61.3 *(19.6)*	69.1 *(22.2)*

*) *The total estimation interval* is calculated as the sum of the sizes of the estimation intervals of the single variables.

Further results show that the estimates in the center of the estimation intervals provide only limited improvements (Table 10). They deliver superior results in the main-effect design but fail to outperform LINMAP in the extended design. Furthermore, they do not lead to smaller variations in the accuracy of individual estimates. This failure can partly be explained by the simplified procedure applied for determining the estimation interval[18]. However, it also indicates that any point estimate within the estimation interval remains arbitrarily.

Table 10: Performance of center of gravity

Mean deviation from true part-worth values *(standard deviation)* Preference function	2^{5-2} main-effect design size of ND(0,σ) error term				extended 2^{5-1} design size of ND(0,σ) error term			
	σ = 0	σ = 5	σ = 10	σ = 20	σ = 0	σ = 5	σ = 10	σ = 20
Total deviation* DEV	19.8 *(9.4)*	21.6 *(10.1)*	23.3 *(9.8)*	31.5 *(12.6)*	18.2 *(9.0)*	19.6 *(8.7)*	20.0 *(8.4)*	24.5 *(9.1)*
Relative performance**								
.compared with OLS	+15%	+10%	+6%	+1%	+29%	+14%	+10%	+0%
.compared with LINMAP	+26%	+21%	+21%	+13%	+17%	-3%	-14%	-13%

*) The total deviation measures metric estimation accuracy and is calculated as the sum of the absolute deviations of the metric variables' part-worth estimates from their true values.

**) Relative performance is measured as percentage of decrease of mean deviation from true importance weights. Negative numbers indicate larger deviations of estimates than those gained from the traditional estimation methods.

In sum, the above outlined analyses started with the ambitious objective of finding a solution which leads with highest probability to the observed ranking. Following this, the possibility of evaluating a conservative estimate which minimizes the risk of misled judgment were analyzed. Even this approach, however, provided only a limited information gain compared over traditional point estimates. Thus there remains a high degree of ambiguity in point estimates gained from ranked Conjoint-data.

5 Conclusions

The model of ranked Conjoint-data and simulation findings indicate limitations of ranked Conjoint-data for gaining accurate utility-function estimates. As a very simple, but often overlooked rule one should recall that information which is not provided in the ranking can not be regained by sophisticated statistical procedures.

While above analyses focused on individual-level estimates, it is unlikely that improvements of estimation accuracy can be gained by applying existing aggregation methods. This is because the superiority of individual-level estimates has repeatedly been documented in empirical conjoint applications (e.g.: Moore, 1980, Green, Srinivasan, 1990, Krieger et al., 1996). However,

there are some other possibilities for improving common practice which attract further attention.

Goodness-of-fit measures as Kendalls τ or Spearman's ρ are inadequate proxies for assessing estimation accuracy. Improved measures should reflect the dual nature of Conjoint-data. They should not only rely on the ranking itself, but should encompass the underlying metric stimuli utilities as well. Validity measures based on the normal distributed error model should constitute a more rigid test of model validity than commonly used measures. The applicability of such an approach has still to be proven, since rankings are sequences of overlapping normal distributions and thus lead to infinitely small combined probabilities. However, looking at single probabilities would discard valuable information on joint probability. Thus, it has to be evaluated whether meaningful cut-off-levels can be found.

Traditional estimation techniques as OLS or LINMAP are neither superior as they do not fulfill the requisite traits of the model of ranked Conjoint-data. A simple heuristic method of deriving point estimates on average outperforms the traditional estimation methods. This clearly indicates possibilities for improving estimate accuracy, if an estimation method could be developed which better fulfill the proposition of the model of ranked-Conjoint data. As an alternative, it is worthwhile to refine the heuristic procedure of calculating the center of estimation interval.

In addition, it seems feasible to use information above and beyond simple point-estimates. However, probabilistic information are likely to overtax the information processing capabilities of practitioners. They offset the advantages of the Conjoint method against self-explicated models, which lie especially in its striking appeal of easiness and understandability. Thus, at least a lower and an upper estimation bound should be documented in individual-level analyses.

The stochastic information provided by the estimation intervals should be of larger use in aggregated analyses than for the individual analyses studied here. For example, a probabilistic clustering could be applied to gain improved estimates for consumer segments. Furthermore, aggregated estimates, which regard the individual observations as replications, could utilize distribution

patterns of observed rankings as valuable information. They should simultaneously assess a shared utility function and an overall level of the error-term. A best estimate could then be defined as the solution, which leads to a distribution of possible rankings which fits best with the distribution of observed rankings.

Footnotes

---

[1] Colberg (1977) defines metric determinacy as the coefficient of determination (i.e. the correlation coefficient) between „true", i.e. given, and estimated part-worth values of a utility function.

[2] Utility function are generated using an averaging mechanism of random part-worth values. Since each design is tested with only 40 different utility functions, there is only a slight chance that heterogeneous or dominant utility functions are enclosed in the simulations.

[3] The coefficient of determination indicates a good fit between estimated and true utility values even in those case where the estimated values possess little explanatory power: Assumed a preference function contains three dichotomous variables and consists of following true variable weights: 40%, 33%, 27%. Then an estimated preference function with equal variable weights (33,3%) delivers a coefficient of determination equals 0.975. Colberg observes for this type of preference function slightly lower coefficients of determination (0.9715 for 4 stimuli and of 0.9664 for 8 stimuli) in his simulations. This implies that the estimates derived from Conjoint-simulation contained less accurate information than those build on the simple assumption of equal importance.

[4] Preference functions were generated as following: Three basic pattern were defined with the following importance weights of the five variables: Dominant preference function: 60%, 10%, 10%, 10%, 10%; homogeneous preference function: all variables equal 20%; heterogeneous preference function: 30%, 25%, 20%, 15%, 10%. From these basic pattern 3*100 different utility functions were generated by modifying each variable by two random terms as follows: (a) addition of an incremental term of +/- 10% and (b) random change of sign of the corrected importance weights to specify the direction of part-worth values.

[5] The 3*100 error-free rank orders of the simulation example in section 2.1 serve as basis. They are corrected by the two most-likely errors, i.e. the reversals between those rank pairs with most similar metric utility values. Then a heuristic assessment of whether the error is

visible or whether the faulty observations constitute a *different, fully consistent ranking* is performed: Analogeous to section 2.2, the observed rank order is recoded into multiple sets of metric utility values and evaluated by standard OLS technique. If at least one of 500 simulations provides an estimate with perfect fit, the observation is marked as one which constitutes a fully consistent ranking.

[6] Given that we observe in the example described in Figure 2 a reversal of ranks 5 and 6, then stimulus 5 is likely to be highly overvalued. This would lead to a significant increase of the probability of subsequently observing a reversal between stimuli 5 and 7.

[7] *Kendall's Tau* is an ordinal measure which counts the number of rank reversals and sets them in relation to the total number of rank pairs. This implicitly assumes that each rank reversal is equally severe. In this regard, there is congruence with the metric measure *Spearman's Rho*, which calculates the metric correlation between observed and estimated ranking. Both procedures disregard differences in the quality of rank reversals, which is equivalent with an implicit interval-scale assumption of ranks.

[8] The estimated fit is calculated based on individual OLS-estimates. Since OLS does not maximize the fit of observed and estimated ranking, this comparison provides a lower bound for the estimated fit. Higher values are likely to be obtained by LINMAP. See section 3.4 for detailed explanations and Table 6 as example.

[9] Shortcomings of the goodness-of-fit measures are especially expected in calculating the predictive validity of small hold-out samples, which possess few ranking information.

[10] A ND(0,10) error term implies that stimuli are evaluated with 95%-confidence intervals of around +/- 20 utility points around their true means. An interval of size 40 utility points is equivalent to 20% of the entire normalized utility interval (ranging from -100 to +100 utility points). In contrast, a rank pair accounts on average for 1/15 of the entire utility interval, thus rank neighbored stimuli deviate on average by only 13,3 utility points.

[11] The analyses indicate that the error term of 20 points can be regarded as a cut-off level. Such an error is likely to cause not only minor rank reversals but also skipped ranks. This leads to a high ambiguity of estimates which rely on few rank pairs, i.e. main-effect design. The extended design is less affected, because different rank reversals are more likely to offset each others influence on estimation outcomes.

[12] Srinivasan and Shocker (1973) recognize that alternate optima may exist for their linear programming technique LINMAP and that they could be enumerated. However, there is no evidence for any application of such a procedure.

[13] The range of the simulated „true" preference functions (see footnote 4) amounts on average 13.7 utility points in case of homogeneous preference functions (23.3 for the heterogeneous and 58.8 for the dominant preference functions). The average bias of the estimated

preference function range can be recalculated from Table 7 by subtracting the mean deviation (of estimated from true part-worth values) of the least important variable from the mean deviation of the most important variable. In case of the homogeneous preference functions, the OLS-estimate overestimates the variable range by on average 20.7 utility points (= -(-10.5)+10.2), which comes to 152% of the average preference function range.

[14] Adverse effects are shown in a second study (Darmon, Rouziès, 1994). Unfortunately, the authors do not comment on the revealed differences. Furthermore, their variable definitions are not consistent to the presented results (the range recovery index contains the factor 100 in the denominator, whereas the numbers indicate that it is used in the nominator), thus it is unfortunately impossible to reconstruct their findings.

[15] The differences of individual preference functions are measured by the average standard deviation of importance weights across all variables. In case of homogeneous preference functions and extended design, the „true" deviation amount on average 5.3 utility points, whereas the preference functions as estimated with OLS deviate by 11.8 utility points and those estimated with LINMAP by 10.2 utility points.

[16] Based on the normal distributed error term, each stimuli is valued most likely with its true value. The symmetric distribution also causes joint probabilities to be highest in the middle.

[17] Slight modifications are necessary in order to account for rank inconsistencies of stochastic data: Instead of requiring a perfect rank recovery of the solutions within the estimation interval, they only ought to exceed a benchmark value of goodness-of-fit, which here has been chosen to be the fit which would be obtained if using OLS-regression.

[18] In order to limit the necessary computer time for simulations, the applied algorithm either stopped if 500 solutions were found or if 1000 trials were unsuccessful in finding another solution. This procedure, which required substantial calculation time with standard software, turned out to be insufficient to ensure a full approximation of the estimation interval. The true utility function was on average enclosed only in two-third of all estimation intervals.

Appendix: Modeling ranking probabilities

Based on the model of a normal distributed error term centered around the error-free metric stimuli utilities, it is possible to calculate probabilities to observe any rank order under a given preference function and fixed size of the error term. This, however, requires extensive calculations of overlapping probability integrals. Thus we apply an approximation by dividing the range of possible utility values into equal intervals of size 2*δ and mean value z.

First, the expected metric utility value $M\tilde{V}(i)$ are calculated as the error-free sum of the part-worth values of all stimuli i. Hereby, the stimuli are sorted by increasing preference with $M\tilde{V}(i) < M\tilde{V}(j)$ and, correspondingly, rank R(i) < R(j) (lower ranks are defined as less preferred) for any $i < j$. Given a ND(0, σ) error term, the initial unconditional probabilities of observing a stochastic utility value $MV(i)$ either within an interval centered around z or lower than that are calculated as following:

$$p(MV(i) \le z-\delta) = \int_{-\infty}^{z-\delta} NV(M\tilde{V}(i),\sigma)$$

and $\quad p(z-\delta < MV(i) \le z+\delta) = p(MV(i) \le z+\delta) - p(MV(i) \le z-\delta)$

The joint probability of observing a rank order R is calculated using an iterative procedure: Beginning with the least preferred stimulus of the error-free rank order $\tilde{R}$, the probability of observing its ranking position (given all previous ones) as postulated by the rank order R is approximated by the product of all pairwise comparisons of any rank pair R(i), R(j) with j>i:

$$p(i \,/\, R) = \prod_{j=i+1}^{n} p(i,j)$$

with: $\quad p(i,j) = \sum^{z} p(z-\delta < MV(i) \le z+\delta) * (1 - p(MV(j) \le z+\delta)) \quad \cap R(i) < R(j)$

: $\quad p(i,j) = \sum^{z} p(z-\delta < MV(i) \le z+\delta) * p(MV(j) \le z-\delta) \quad \cap R(i) \ge R(j)$

Subsequently, the probability distributions of the remaining stimuli are recalculated based on this information. This leads to new, conditional probabilities which build the basis for the next iteration step:

$$p(MV(i) \leq z+\delta) = p(MV(j) \leq z+\delta) * (1 - p(MV(i) < z+\delta)) \qquad \cap R(i) < R(j)$$

$$p(MV(i) \leq z+\delta) = p(MV(j) \leq z+\delta) * p(MV(i) < z+\delta) \qquad \cap R(i) > R(j)$$

This procedure is repeated from the least preferred to the most preferred stimulus of the error-free rank order $\tilde{R}$. Finally, the probability of observing the entire rank order R is calculated as the product of observing each of the n stimuli in accordance to its position within the rank order R.

$$p(R) = \prod_{i=1}^{n} p(i / R)$$

References

Albers, S. (1984), Fully Nonmetric Estimation of a continuous nonlinear Conjoint Utility Function, International Journal of Research in Marketing, 311-319.

Bateson, J., Reibstein, D., Boulding, W. (1987), Conjoint Analysis Reliability and Validity: A Framework for Future Research, in: Houston, M. (Hrsg.), Review of Marketing, Chicago, IL: American Marketing Association. 451-481.

Ben-Akiva, M., Morikawa, T., Shiroishi, F. (1992), Analysis of the Reliability of Preference Ranking Data, Journal of Business Research, Vol. 24, 149-164.

Box, G., Hunter, W., Hunter, J. St. (1978), Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley&Sons, New York u.a.

Bradley, M., Daly, A. (1994), Use of the Logit Scaling Approach to Test for Rank-Order and Fatique effects in Stated Preference Data, Transportation, Vol. 21, 167-184.

Brockhoff, K. (1972), On Determining Relative Values, Zeitschrift für Operations Research, Vol. 16, 221-232.

Carmone, F.J., Green, P.E. (1981): Model Misspecification in Multiattribute Parameter Estimation, Journal of Marketing Research, Vol. 18, 87-93

Carmone, F.J., Green, P.E., Jain, A.K. (1978): Robustness of Conjoint Analysis: Some Monté Carlo Results, Journal of Marketing Research, Vol. 15, 300-303.

Cattin, P., Bliemel, F. (1978), Metric versus Nonmetric Procedures for Multiattribute Modeling: Some Simulation Results, Decision Science, Vol. 9, 472-480.

Chapman, R. G., Staelin, R. (1982), Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model, Journal of Marketing Research, Vol. 19, 288-301.

Colberg, R. T. (1977), Validation of Conjoint Measurement Methods: A Simulation and Empirical Investigation, University of Washington, Ph.D.

Darmon, R., Rouziès, D. (1989), Assessing Conjoint Analysis Internal Validity: The Effect of Various Continuous Attribute Level Spacings, International Journal of Research in Marketing, Vol. 6, 35-44.

Darmon, R., Rouziès, D. (1991), Internal Validity Assessment of Conjoint Estimated Attribute Importance Weights, Journal of the Academy of Marketing Science, Vol. 19/4, 315-322.

Darmon, R., Rouziès, D. (1994), Reliability and Internal Validity of Conjoint Estimated Utility Functions under Error-Free versus Error-Full Conditions, International Journal of Research in Marketing, Vol. 11, 465-476.

Davidson, D., Suppes, P., in collaboration with Siegel, S. (1957), Decision Making: An Experimental Approach, Stanford, Cal.

Green, P.E., Rao, V. R. (1971), Conjoint Measurement for Quantifying Judgmental Data, Journal of Marketing Research, Vol. 8, 355-363.

Green,P.E., Srinivasan,V. (1978), Conjoint Analysis in Consumer Research: Issues and Outlook, Journal of Consumer Research, Vol. 5, 103-123

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

Green, P.E., Wind, Y. (1973), Multiattribute Decisions in Marketing: A Measurement Approach, Hinsdale, Il: Dryden Press.

Hensher, D. (1994), Stated Preference Analysis of Travel Choices: The State of Practice, Transportation, Vol. 21, 107-133.

Kalish, S., Nelson, P. (1991), A Comparison of Ranking, Rating and Reservation Price Measurement in Conjoint Analysis, Marketing Letters, Vol. 2/4, 327-335.

Karson, M., Mullet, G. (1989), Conjoint Utility Limits as Affected by Conjoint Design and Estimating Program, Marketing Research, Dezember, 27-32.

Kendall, M.G. (1962), Rank Correlation Methods, 3. Auflage, London: Charles Griffen.

Krieger, A.M., Green P.E., Umesh, U.N. (1996), Effect of Level of Disaggregation on Conjoint Cross Validations: Some Comparative Findings, Paper presented in: Marketing Science Conference, Berkeley, May 1997.

Luce, R. D., Suppes, P. (1965), Preference, Utility, and Subjective Probability, in: Luce, R.D., Bush, R.R., Galanter, E. (Hrsg.), Handbook of Mathematical Psychology, Volume 3, New York: John Wiley & Sons, 249-410.

Moore, W.L. (1980), Levels of Aggregation in Conjoint Analysis: An Empirical Comparison, Journal of Marketing Research, Vol. 17, 516-523.

Müller-Hagedorn, L., Sewing, E., Toporowski, W. (1993), Zur Validität von Conjoint-Analysen, Zeitschrift für betriebswirtschaftliche Forschung, Heft 2, 123-148.

Mullet, G. M., Karson, M. J. (1986), Percentiles of LINMAP Conjoint Indices of Fit for Various Orthogonal Arrays: A Simulation Study, Journal of Marketing Research, Vol. 23, 286-290.

Sattler, H. (1994), Die Validität von Produkttests, Marketing ZFP, Heft 1, 31-41.

Schoemaker, P., Waid, C.C. (1982), An Experimental Comparison of Different Approaches to Determining Weights in Additive Utility Models, Management Science, Vol. 28(2), 182-196.

Schrader, S. (1990), Zwischenbetrieblicher Informationstransfer - Eine empirische Analyse kooperativen Verhaltens, Duncker, Humblot: Berlin.

Shocker, A.D., Srinivasan, V. (1977), LINMAP (Version II): An FORTRAN IV Computer Program for Analyzing Ordinal Preference (Dominance) Judgements Via Linear Programming Techniques for Conjoint Measurement, Journal of Marketing Research, Vol. 14, 101-103.

Srinivasan, V., Basu, A. (1989), The Metric Quality of Ordered Categorical Data, Marketing Science, Vol. 8 (3), 205-230.

Srinivasan, V., Shocker, A. (1973), Estimating the Weights for Multiple Attributes in a Composite Criterion Using Pairwise Judgments, Psychometrika, Vol. 38 (4), 473-493.

Stallmeier, Ch. (1993), Die Bedeutung der Datenerhebungsmethode und des Untersuchungsdesigns für die Ergebnisstabilität der Conjoint-Analyse, Regensburg.

Steenkamp, J.-B., Wittink, D.R. (1994), The Metric Quality of Full-Profile Judgments and the Number-of-Attribute-Levels Effect in Conjoint Analysis, International Journal of Research in Marketing, Vol. 11, 275-286.

Stimson, D. H. (1969), Utility Measurement in Public Health Decision Making, Management Science, Vol. 16, B-17 - B-30.

Teichert, T. A. (1993), The Success Potential of International R&D Cooperation, Technovation, Vol. 13/8, 519-532.

Teichert, T. A. (1994), Zur Validität der in Conjoint-Analysen ermittelten Nutzenwerte, Zeitschrift für betriebswirtschaftliche Forschung, Vol. 46 (7/8), 610-629

Teichert, T. A. (1996), The Confounding of Effects in Rank-Based Conjoint-Analysis Cooperative P&D and partners' measures of success, working paper #409, Institute for Business Administration, University of Kiel.

Teichert, T. A. (1997, forthcoming), Schätzgenauigkeit von Conjoint-Analysen, Zeitschrift für Betriebswirtschaft.

Umesh, U. N., Mishra, S. (1990), A Monte Carlo Investigation of Conjoint Analysis Index-of-fit: Goodness of Fit, Significance and Power, Psychometrika, Vol. 55/1, 33-44.

Weisenfeld,U. (1989), Die Einflüsse von Verfahrensvariationen und der Art des Kaufentscheidungsprozesses auf die Reliabilität der Ergebnisse bei der Conjoint Analyse, Berlin.

Wittink, D.R., Cattin, P. (1981): Alternative Estimation Methods for Conjoint Analysis: A Monté Carlo Study, Journal of Marketing Research, Vol. 18 (2), 101-106.

Wittink, D.R., Krishnamurthi, L., Reibstein, D.J. (1989), The Effects of Differences in the Number of Attribute Levels on Conjoint Results, Marketing Letters, Vol. 1, 113-123.

Wittink, D.R., Vriens, M., Burhenne, W. (1994), Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections, International Journal of Research in Marketing, Vol. 11, 41-52.