Validity and Reliability of Online Conjoint Analysis

Torsten Melles

Westfaelische Wilhelms-Universitaet Muenster

Ralf Laumann

Westfaelische Wilhelms-Universitaet Muenster

Heinz Holling

Westfaelische Wilhelms-Universitaet Muenster

ABSTRACT

Using conjoint analysis in online surveys is gaining growing interest in market research. Unfortunately there are only few studies that are dealing with the implementing of conjoint analysis in the World Wide Web (WWW). Little is known about specific problems, validity, and reliability using online measures for conjoint analysis.

We conducted an online conjoint analysis using a fixed design of thirty paired comparisons. A traditional computerized conjoint analysis was conducted in the same way. Several criteria were used to assess reliability and validity of both data collection methods.

The results show that data drawn from an Internet conjoint analysis seem to be somewhat lower in reliability (internal consistency) compared to traditional computerized conjoint analysis. Nevertheless, the reliability seems to be sufficient even in the case of its online form. Regarding predictive validity, both data collection methods lead to comparable results. There is no evidence that the number of thirty paired comparisons might be too high in the case of Internet conjoint analysis. More paired comparisons seem to be favorable taking the moderate internal consistency of responses into account and the additional possibilities of reliability testing.

BACKGROUND AND INTRODUCTION

After three decades using conjoint analysis there is still a growing interest for choosing this method to analyse preferences and predict choices in marketing research and related fields (Cattin and Wittink, 1982; Wittink and Cattin, 1989; Wittink, Vriens and Burhenne, 1994; Melles and Holling, 1998; Voeth, 1999). The contributions of many researchers have led to a diversification of methods for stimulus-construction, scaling, part-worth estimation, data aggregation and collecting judgments from the subjects. The following paper will focus on techniques for collecting judgments that can be used in conjoint analysis and on an empirical examination of the reliability and validity of the newly introduced method of online conjoint analysis, conducted over the World Wide Web (WWW). Little is known about the quality of data generated by an online conjoint analysis. Is the WWW an appropriate place for collecting complex judgments? Is the quality of this method comparable to that of other collection techniques?

Several techniques using different media have been proposed to collect multiattribute judgments. Up to the mid 80s conjoint analysis was nearly exclusively done by paper-and-pencil-tasks in the laboratory or by traditional mail surveys. The introduction of ACA has led to a radical change. Today, the method most often used for conjoint is the computeraided personal interview (CAPI).

The methods used in conjoint analysis can be categorized according to three dimensions (Table 1). This categorization is a simple attempt of aligning the methods. It is neither comprehensive nor are the categories sharply distinct. Collection methods using configural stimuli are not listed as well as mixtures of different procedures like the telephone-mail-telephone (TMT) technique. Each one of the methods displays a specific situation for making judgments. Therefore, it cannot be expected that these methods are equivalent. In a traditional mail survey, for example, the questionnaire is done by paper and pencil without an interviewer present to control or help the subject. In the case of a computeraided interview the stimuli are shown on a screen and an interviewer is present. This facilitates a higher level of control and help. Problems can arise in cases where interviewer biases have to be expected.

Table 1: Methods of collecting multiattributive judgments in conjoint analysis. Visual methods use verbal descriptions or pictures.

		computeraided	non-computeraided
personal	visual	computeraided personal interview (CAPI)	personal paper-and-pencil-task
	acoustic	(personal interview)	
non-personal	visual	disk-by-mail (DBM), online-interview	traditional mail survey
	acoustic	computeraided telephone-interview (CATI)	telephone-interview

Comparisons have been made between traditional mail surveys, telephone interviews, personal paper-and-pencil tasks (full-profile-conjoint) and ACA, the mostly used computeraided personal interview method (e.g. Akaah, 1991; Chrzan and Grisaffe, 1992; Finkbeiner and Platz, 1986; Huber, Wittink, Fiedler, and Miller, 1993). It is very difficult to draw conclusions from these comparisons because of many factors confounded, favouring one method against the other. For instance, ACA uses a specific adaptive design and a specific scaling of judgments and estimation procedure. So differences to part-worths gained from mail-surveys can arise from each of these characteristics or their interaction as well as from the specific data collection method. Apart from this limitation, personal paper-and-pencil task and ACA can be viewed as nearly equivalent in reliability and validity. Traditional mail surveys and telephone interviews can lead to the same level of accuracy. However, this depends on several characteristics of the target population and it is only suitable with a low number of parameters (six attributes or fewer).

Using the Internet for conjoint analysis receives growing interest, especially in marketing research (Saltzman and MacElroy, 1999). Nevertheless, little is known about problems arising from the application of conjoint analysis over the Internet and the quality of this data. Only few studies are dealing with online conjoint analysis. These exceptions are studies published by Dahan and Srinivasan (1998), Foytik (1999), Gordon and De Lima-Turner (1997), Johnson, Leone, and Fiedler (1999), Meyer (1998), Orme and King (1998).

Meyer (1998) observed that the predictive validity of his online conjoint analysis (using a full-profile rating task) was much better than random generated estimations. But there is no way to compare it with data gained from other collection methods. Orme and King (1998) tested the quality of their data using a holdout task (first choice). They found single concept judgments to perform as good as graded paired comparisons. The stimuli were full-profiles consisting of four attributes. Orme and King (1998) emphasize on the common features of Internet surveys and traditional computerized surveys. Only Foytik (1999) compared the Internet to other data collection methods in conjoint analysis. Drawing from several studies he reports higher internal consistency measured by Cronbach's Alpha and the Guttman Split-Half test of the Internet responses compared to traditional mail responses as well as more accurately predicted holdout choices.

At this point there are some unresolved questions regarding the quality and specific features of Internet conjoint analysis. Some of these questions are:

- How many judgments should / can be made?
- Is the predictive validity and reliability comparable to traditional computerized surveys?
- Which ways can be effective in handling problems of respondents' drop-out during the interview and "bad data"?

METHOD

The research questions were tested by conducting a conjoint analysis of call-by-call-preferences. Resulting from a liberalization of the German telephone market there are several suppliers that offer one single telephone call without binding the consumer. This call-by-call use is possible through dialing a five-digit supplier-specific number before the regular number. A choice is made each time before a call to be made. Call-by-call services vary between weekdays and weekends, in the time of day, and the target of the call. There is no supplier dominating the others in general.

The selection of attributes and levels based on results of earlier studies, expert interviews and a pilot study. The attributes should have been relevant at the moment the decision between different suppliers is made and the levels should have been realistic. Having this criteria in mind, four attributes (price per minute, possibility to get through, interval of price cumulation, extras) were chosen. Two of them had three levels, the other two had two.

Subjects were users of call-by-call-services who visited the internet site *http://www.billiger-telefonieren.de* and decided to participate in the study. This was done by 9226 respondents during the two week period the conjoint survey was available on the website. In order to elicit their true preferences that are related to choice, subjects were asked to evaluate services that were relevant to them and that were adapted to their telephoning habits. If one is calling mainly weekdays between 7 and 9 pm to a distant target in Germany, he was asked to judge the services in front of this situation. So it is possible to distinguish different groups of users, and it is assured that the subjects are able to fulfill the task.

Judgments were made by a graded paired comparison task. As in ACA no full-profiles were used. Due to cognitive constraints the number of attributes was limited to three (e.g. Agarwal, 1989; Huber and Hansen, 1986; Reiners, Jütting, Melles, and Holling, 1996).

Each subject has been given 30 paired comparisons. This number provides a sufficiently accurate estimation of part-worths given a design with fewer than six attributes and three levels on each attribute. Reiners (1996) demonstrated for computeraided personal interviews that even more than 30 paired comparisons can lead to slightly more reliable and valid part-worths. Additionally, results from other studies show that one can include more questions in personal interviews than in non-personal interviews (e.g. Auty, 1995). Due to these differences of online testing to personal interviews - that may cause a lower level of control and a lower level of respondent motivation - and regarding the length of the whole questionnaire, 30 paired comparisons seemed to be an appropriate number. We chose an approximately efficient design by using a random procedure that selected from various designs that one with minimal determinant of the covariance matrix (Det-criterion). The sequence of paired comparisons was randomized as well as the screenside of the concepts and the position of the different attributes because of possible sequence- and position-effects.

Different precautions were taken to prevent “bad data” caused by a high drop-out rate of respondents:

- functional, simple web-design in order to maximize the speed of the data transfer
- providing an attractive incentive after finishing the questionnaire (participation in a lottery)
- explicitly emphasizing on the fact that the whole interview takes 20 minutes to perform, before the respondent finally decided to participate
- emphasizing on the importance of completely filled in questionnaires.

IP-addresses and responses to personal questions were checked in order to prevent double-counting of respondents. Datasets with identical IP-addresses together with the same responses to personal questions were excluded as well as datasets with identical IP-addresses and missing data to personal questions.

The quality of responses and conjoint-data was measured by multiple criteria:

- Estimating part-worths using an OLS-regression provided with R^2 a measure of internal consistency (goodness of fit). This gives a first indication to the reliability of judgments. But it is necessary to emphasize that the interpretation of this measure can be misleading and must be made carefully. Beside several important problems of this measure there are two specific ones that are related to the distribution of responses. A high R^2 can result from “bad data” (e.g. due to response patterns without any variance) and a low R^2 can result from using only the extremes of the graded scale. For the special case of dichotomous responses and a situation where proportions of success are bounded by [.2, .8] Cox and Wermuth (1992) have shown that the maximum possible value of R^2 is .36.
- Stability of the part-worth estimations on the group level has been measured by intercorrelations between part-worths using each single paired comparison as an input. This means that the responses to the first paired comparison have been taken and aggregated in an estimation on the group level. The same has been done with the second paired comparison and so on. This aggregate estimation was possible as a fixed design has been used and the position of each paired comparison across a high number of respondents has been randomized. Between each pair of estimated part-worth-sets Pearson r has been calculated and plotted in an intercorrelation matrix. Assuming homogeneity of preference structures the mean correlation of each row, respectively of

each paired comparison, is a measure of stability of estimated part-worths. Due to a warm-up-effect[1] and descending motivation together with cognitive strain while performing the task, an inverted u-function is expected.

- A Split-Half test has been performed to test the reliability of responses and part-worths on the individual level. Part-worth-estimations using the first fifteen paired comparisons have been correlated with part-worths derived from the last fifteen paired comparisons. To provide a reliability measure for the whole task the correlation coefficient has been corrected by the Spearman-Brown-Formula. This reliability measure must be interpreted carefully and can only be taken as a heuristic because the reduced design is not efficient taking Det-criterion into account.
- We used a holdout task as a further measure of reliability, respectively internal validity. Due to the difference between this criterion and the task in this case it seems to be more a measure of internal validity than a measure of reliability (see Bateson, Reibstein, and Boulding, 1987, for a discussion on these concepts). Estimated part-worths were used to predict rankings of holdout concepts that were drawn from actual call-by-call-offers made by the suppliers. The name of the supplier was not visible to the subjects. The rankings were drawn from the first choice, second choice and so on between the concepts. The maximum number of concepts that needed to be selected was five. The ranking was correlated using Spearman Rho with the predicted rank order of the same concepts for each individual subject.
- A choice task that asked the respondents to select between different suppliers (concepts not visible) was used to measure external validity. This task was in analogy with the holdout task.

We conducted a computeraided personal interview that was similar to the Internet interview in order to compare the reliability and validity of both conjoint analyses. A student sample (N=32) was asked to indicate their preferences regarding suppliers offering a long distant telephone call at 7 pm weekdays.

Results

The percentage of dropped out respondents over the interview gives a first impression of the quality of measurement. This number is encouragingly low (<15%). The percentage of missing data raised from the first paired comparison task to the last at about 7%.

Stability of aggregated part-worths was tested for each paired comparison. Assuming homogeneity of respondents' preferences, the mean correlation between one set of part-worths with the other sets can be accepted as a measure of stability of the first set. In general, the intercorrelations are nearly perfect indicating a high degree of stability of aggregate preference estimations. As we expected, there is a minimal indication of a warm-up-effect. Stability is rising during the first five paired comparisons. After five paired comparisons its degree persists at a high level even after thirty paired comparisons. So there is no evident effect of motivational or cognitive constraints that would reduce the stability of aggregated part-worths during a paired comparison task with thirty pairs.

In order to test Split-Half reliability, part-worth-estimations using the first fifteen paired comparisons were correlated with part-worths derived from the last fifteen paired comparisons. The Spearman-Brown-corrected mean correlation (after Fisher-Z-transforming

[1] Several studies have shown that respondents need some trials to adapt to the task.

the correlations) was .94 indicating reliable part-worth estimations on the individual level. In contrast, the median R^2 after thirty paired comparisons was insufficiently low ($Md_{R^2} = .31$), indicating a low level of internal consistency. This was also true for the computeraided personal interview, though the median R^2 was slightly higher ($Md_{R^2} = .44$). Split-half reliability was also higher in the case of the personal interviews (.97). There is much room for speculation when looking for a reason R^2 being low since part-worth estimations seem to be reliable. A first step to bring light to the dark is to take a look at the distribution of responses. Doing so there are two striking features: The first is that respondents tend to use extreme categories on the graded scale when judging call-by-call services. There is no difference between the computeraided personal interview and Internet interview. This response behavior might be due to a missing trade-off between the attributes that is typical for decisions without a high involvement (e.g. buying decisions that take little cognitive effort). Strictly speaking this decision behavior is not compatible with the additive rule that is used in conjoint analysis. Since there is no compensation between advantages and disadvantages on different attributes part-worths provide only ordinal information. This has to be kept in mind when to predict choices as in running market simulations. The second feature is a response bias in the Internet interview. The distribution was biased to the right side of the scale. The question if these two features have led to a low R^2 can only be answered finally with running Monte Carlo simulations but it was not possible in the scope of this study.

There is no difference between the methods (online and CAPI) regarding predictive validity measured by the holdout task (Table 2). Both provide nearly accurate predictions. This does not seem to be the case, when predictive validity was measured by a choice between different suppliers. The personal conjoint analysis does better in predicting the ranks than the Internet conjoint analysis. But this result is misleading. Taking into account the differences between the Internet sample and the personal interview sample leads to a conclusion that the methods are equivalent. If only subjects that are younger than 30 years, have a high school diploma (Abitur), and were asked to indicate their preferences for suppliers offering a long distant telephone call at 7 pm weekdays were selected, the coefficient was slightly higher in the case of the Internet conjoint analysis. This small difference might be due to the higher interest of the subjects participating in the Internet study of telephone services.

Table 2: Validity of conjoint analysis conducted through a computeraided personal interview (CAPI) and the Internet.

	CAPI	Internet-Interview	
internal validity (holdout task)	.968 (N=32)	.970 (N=7813)	.977 (N=941)
external validity (choice task)	.539 (N=30)	.412 (N=5663)	.552 (N=691)

Remark: The coefficients in the second column are based on all respondents that participated in the Internet survey. The third column is based on a sample of that survey that is equivalent to the computeraided personal interview.

Conclusions and Discussion

The overall conclusion that can be drawn from the results is the general suitability of Internet conjoint analysis to measure preferences in a reliable and valid manner. This statement has to be qualified in several categories. There is a lot of “bad data” resulting from double-counted respondents and response patterns caused by respondents that decide to take a look at the interview but do not participate seriously. Though taking much effort cleaning the data, the reliability of individual level estimates seems to be somewhat lower than in personal interviews. This may be due to the anonymous situation and a lower level of cognitive efforts spent on the task. As in the cases of using conjoint analysis in traditional mail surveys and telephone interviews, the suitability of the Internet conjoint analysis depends on characteristics of the respondents and on the design of the task respectively the number of parameters that must be estimated. The higher the number of parameters, the higher the number of responses that are required for a detailed analysis at the individual level. This again might decrease the motivation to fulfill the task and is a further critical factor for receiving reliable results from an Internet conjoint analysis. Nevertheless, in this particular case reliability was still high even after 30 paired comparisons. Apart from design characteristics reliability and validity of conjoint analysis vary within the characteristics of respondents (Tscheulin and Blaimont, 1993). Up to now, there is no evidence whether this effect might be moderated or not by the data collection method. This could be a further limitation to a broad application of specific data collection methods like telephone or the Internet. Assuming the Internet to be an appropriate medium regarding respondents and design, the following precautions should be taken to assure a high degree of reliability and validity:

- use as many criteria as possible to test the reliability and validity of the conjoint analysis
- use incentives to motivate respondents in giving reliable responses (e.g. giving a feedback of goodness-of-fit)
- encourage respondents to give a feedback and use as much feedback as possible
- IP-addresses and personal data should be controlled for double-counting whenever the participation is based on a self selection of respondents
- analysis on the individual level should precede aggregate analysis to identify “bad data”

Beside the problems of reliability and validity the choice of a data collection method for conjoint analysis is still a practical one. The suitability of Internet conjoint analysis depends on constraints regarding time, money and experience. Preparing an Internet survey needs more time and money than preparing a comparable personal interview. On the other hand, it serves as an advantage in collecting data. There are no interviewers and no notebooks needed to interview the subjects in the case of Internet surveys. Instead of, they will mostly be recruited through the WWW, which can sometimes be easy but as well be expensive and time consuming, if it fails. The opportunities and dangers of the most common options for contacting Internet users are shortly discussed by Foytik (1999). Preparing data for analysis is a much more complicated job in Internet conjoint analysis than in computeraided personal interviews. There is many more “bad data” and identifying it is very time consuming and requires some theoretically guided reflection.

Regardless of the data collection method, we recommend to take a look at the distribution of responses across graded paired comparisons and across each respondent. This is helpful to identify response patterns and simplifying tactics. If, for example, respondents adopt a lexicographic rule, it is not appropriate to assume a compensatory additive one. Ordinary Least Square (OLS) is very robust against such violations but predicting decisions could be better done by a “lexicographic choice model” than BTL or First Choice (both models assume that the choice is made by reflecting all attributes of the objects). Assuming a lexicographic choice means that the object with the maximum utility of the most important attribute will be chosen. If two or more objects have the same utility, the decision is made by taking the next important attribute into account and so on. Moreover, such a rule is more able of covering psychological processes involved in buying decisions that take little cognitive effort. This is often being neglected by researchers using conjoint analysis and running market simulations. Trade-offs are assumed in buying yoghurts, dog food, biscuits, jam, cheese, shampoos, coffee, and so on. Since some early articles from Acito and his coworkers reflections on decision processes when applicating conjoint analysis seem to be lost. Predicting and understanding decision making behavior at the market place as well as judgments in conjoint tasks requires some return to the basics. Data collection methods that enhance simplifying tactics cannot be valid in predicting choices that are made through a complex trade-off. Otherwise they might be useful in predicting choices that rely on the same simplifying rules and serve no disadvantage against alternative methods. We found respondents using simplifying tactics (extreme scale categories) in the Internet survey as well as in computeraided personal interviews. The question whether these tactics are more often used in Internet interviews and limiting the suitability of this medium is an issue for further research.

REFERENCES

Agarwal, M.K. (1989). How many pairs should we use in adaptive conjoint analysis? An empirical analysis. In American Marketing Association (Ed.), *AMA Winter Educators' Conference Proceedings* (pp. 7-11). Chicago: American Marketing Association.

Akaah, I.P. (1991). Predictive performance of self-explicated, traditional conjoint, and hybrid conjoint models under alternative data collection modes. *Journal of the Academy of Marketing Science, 19* (4), 309-314.

Auty, S. (1995). Using conjoint analysis in industrial marketing. The role of judgement. *Industrial Marketing Management, 24* (3), 191-206.

Bateson, J.E.G., Reibstein, D.J., and Boulding, W. (1987). Conjoint analysis reliability and validity: A framework for future research. In M.J. Houston (Ed.), *Review of Marketing* (pp. 451-481). Chicago: American Marketing Association.

Cattin, P. and Wittink, D.R. (1982). Commercial use of conjoint analysis: A survey. *Journal of Marketing, 46 (Summer)*, 44-53.

Chrzan, K. and Grisaffe, D.B. (1992). A comparison of telephone conjoint analysis with full-profile conjoint analysis and adaptive conjoint analysis. In M. Metegrano (Ed.), *1992 Sawtooth Software Conference Proceedings* (pp. 225-242). Sun Valley, ID: Sawtooth Software.

Cox, D.R. and Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *The American Statistician, 46* (1), 1-4.

Dahan, E. and Srinivasan, V. (1998). *The predictive power of Internet-based product concept testing using visual depiction and animation*. Working paper, Stanford University, CA.

Finkbeiner, C.T. and Platz, P.J. (1986, October). *Computerized versus paper and pencil methods: A comparison study*. Paper presented at the Association for Consumer Research Conference. Toronto.

Foytik, M. (1999). Conjoint on the web - lessons learned. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 9-22). Sequim, WA: Sawtooth Software.

Gordon, M.E. and De Lima-Turner, K. (1997). Consumer attitudes toward Internet advertising: A social contract perspective. *International Marketing Review, 14* (5), 362-375.

Huber, J. and Hansen, D. (1986). Testing the impact of dimensional complexity and affective differences of paired concepts in adaptive conjoint analysis. In M. Wallendorf and P. Anderson (Eds.), *Advances in consumer research* (No. 14, pp. 159-163). Provo, UT: Association for Consumer Research.

Huber, J., Wittink, D.R., Fiedler, J.A., and Miller, R. (1993). The effectiveness of alternative preference elicitation procedures in predicting choice. *Journal of Marketing Research, 30*, 105-114.

Johnson, J.S., Leone, T., and Fiedler, J. (1999). Conjoint analysis on the Internet. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 145-148). Sequim, WA: Sawtooth Software.

Melles, T. und Holling, H. (1998). *Einsatz der Conjoint-Analyse in Deutschland. Eine Befragung von Anwendern*. Unveröffentlichtes Manuskript, Westfälische Wilhelms-Universität Münster.

Meyer, L. (1998). *Predictive accuracy of conjoint analysis by means of World Wide Web survey* [Online]. Available: http://www.lucameyer.com/kul/menu.htm.

Orme, B.K. and King, W.C. (1998). *Conducting full-profile conjoint analysis over the Internet*. Working paper, Sawtooth Software.

Reiners, W. (1996). *Multiattributive Präferenzstrukturmodellierung durch die Conjoint-Analyse: Diskussion der Verfahrensmöglichkeiten und Optimierung von Paarvergleichsaufgaben bei der adaptiven Conjoint Analyse*. Münster: Lit.

Reiners, W., Jütting, A., Melles, T. und Holling, H. (1996). *Optimierung von Paarvergleichsaufgaben der adaptiven Conjoint-Analyse*. Forschungsreferat zum 40. Kongreß der Deutschen Gesellschaft für Psychologie.

Saltzman, A. and MacElroy, W.H. (1999). *Disk-based mail surveys: A longitudinal study of practices and results*. In *Proceedings of the Sawtooth Software Conference* (No. 7, pp. 43-53). Sequim, WA: Sawtooth Software.

Tscheulin, D.K. and Blaimont, C. (1993). Die Abhängigkeit der Prognosegüte von Conjoint-Studien von demographischen Probanden-Charakteristika. *Zeitschrift für Betriebswirtschaft, 63* (8), 839-847.

Voeth, M. (1999). 25 Jahre conjointanalytische Forschung in Deutschland. *Zeitschrift für Betriebswirtschaft – Ergänzungsheft, 2*, 153-176.

Wittink, D.R. and Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing, 53*, 91-96.

Wittink, D.R., Vriens, M., and Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing, 11*, 41-52.