A comparison of the validity of interviewer-based and online-conjoint analyses

Andreas Klein
University of Duisburg-Essen

Katrin Nihalani
University of Duisburg-Essen

Krish S. Krishnan
Indiana University of Pennsylvania

ABSTRACT

Interviewer-based conjoint analyses are usually time consuming and expensive. By changing the data collection process to the Internet, it is easier to contact a larger number of people within a shorter amount of time. In addition, it is a more cost effective method. However, it is questionable if the data is of the same quality, since conjoint analyses regularily need more time to explain the task to the interviewee. This might have a negative impact onto the validity of the gathered data. Therefore, a scientific study was conducted in order to compare the validity of such two data sets. The results show that there are no big differences in the validity between an interviewer-based computer and an online-conjoint analysis. Rather it can be shown that the validity of the data of the online-conjoint analysis is slightly higher.

Keywords: Conjoint Analysis, Online-Conjoint Analysis, Market Research, Online Market Research, Multivariate Data Analysis

INTRODUCTION

Since about 1990 conjoint analyses are widely used as a tool to develop new products, predict consumer's response to alternative pricing strategies or for running market simulations (Voeth, 1999; Green, Krieger & Wind, 2001; Woratschek, 2001; Sattler & Nitschke, 2003; Backhaus, Wilken, Voeth & Sichtmann, 2005). In addition, over time classical conjoint analyses have been developed further to allow the evaluation of a larger number of attributes (adaptive conjoint analysis) and to integrate real choice situations (choice based conjoint analysis). Generally it is stated, that the validity of a conjoint analysis is higher compared to compositional methods (Voeth, 2000). One reason for that perception is that consumers generally do not evaluate elements of a product separately. Instead they evaluate the entire product at once.

However, it is still a problem that conjoint analyses are very time consuming and cost intensive. One of the main reasons why conjoint analyses often base only on a small sample size or a so-called convenience sample (Ernst & Sattler, 2000; Sattler, Hensel-Börner & Krüger, 2001; Sattler & Nitschke, 2003). Although computer assisted conjoint analyses are conducted today (Hauser & Toubia, 2005) it is still problematic that costs for direct contact between interviewer and interviewee are relatively high, e. g. costs for travelling, training, remuneration etc. of the interviewers (Dibb, Rushmer & Stern, 2001; Ilieva, Baron & Healey, 2002; Görts & Behringer, 2003; Welker, Werner & Scholz, 2005).

Due to the easy and cheap access to the Internet and new software solutions, it is quite easy to conduct the data collecting process for a conjoint analysis over the Internet without any direct contact to the interviewer. By using the Internet as a medium to contact people, geographical barriers do not count as much as before and people can be interviewed, who could not have been interviewed before (Bamert & Heidingsfelder, 2001; Couper, Tourangeau & Kenyon, 2004). In addition, participating in an interview on a PC at home causes less stress for an interviewee compared to an interview, which is conducted in a public place. Compared to paper and pencil conjoint analyses another advantage of the online- or computer-based data gathering process is that data, which is gathered digitally, can be transferred to statistical programs more easily and is easier to check than data, gathered through paper-based interviews (Theobald, 2000; Henning-Thurau & Dallwitz-Wegener, 2002; Ilieva et al., 2002). This helps to improve the quality of the input data and thereby the validity of the part-worths (Daiber & Hemsing, 2005).

Nevertheless, it is questionable if the data, which is gathered over the Internet without an interviewer, is of the same quality as data, which is gathered through personal interviews. Especially since a conjoint analysis is a rather complex method compared to a regular questionnaire, it is questionable, if the validity of the data is negatively affected by an online data collecting process (Duffy, Smith, Terhanian, & Bremer, 2005; Grant, Teller & Teller, 2005; Schillewaert & Meulemeester, 2005).

So far and to our best knowledge only one analysis has been accomplished, in which the validity of two choice based conjoint analyses has been compared. In this analysis one data set was the result of a computer-based personal interview and the other data set was the result of an online survey (Sethuraman, Kerin & Cron, 2005). In addition to the different data gathering methods, the stimuli in the online survey were presented in a multimedia way (Vriens, Looschilder, Rosenbergen & Wittink, 1998; Daiber & Hemsing, 2005). The authors concluded that the data gathered online leads to a higher concurrent and predictive validity. However, it has to be criticised that the predictive validity is only based on the assumptions of managers and not on the data set itself. In addition, it cannot be concluded that the difference in the validity is based

solely on the different data collecting methods, since it could be possible that the multimedia presentation of the online survey is the cause for the difference in validity.

Moreover, two other studies are dealing just with the simulation of an online survey (Ernst & Sattler, 2000; Strebinger, Hoffmann, Schweiger, & Otter, 2000). Because of the lack of other studies in that it cannot be stated that either the online data collecting method or the computer-based personal interview leads to a higher validity of a conjoint analysis.

Therefore a scientific study was conducted in order to examine if one specific data collecting process results in a higher validity of a conjoint analysis. One data set was collected online and the other one was also collected on a computer, but with the help of an interviewer sitting beside the interviewee. The questions of both surveys were identical. The only difference was that one group filled out the questions on a computer and had the chance to ask an interviewer in case of any questions (interviewer-based survey), whereas the other interviewing process was conducted online without the help of an interviewer (online survey). To exclude learning effects the two samples did not consist of the same test persons (Agarwal & Green, 1991; Huber, Wittink, Fiedler & Miller, 1993; Gierl & Höser, 2002). So the only difference between both data collecting processes was the presence of the interviewer in one setting. This allows conclusions about the relevance of the interviewer for the validity of a conjoint analysis in comparison to a self-administered online-conjoint analysis by the interviewee. To ensure a correct behavior of the interviewers they were trained intensively and only helped if a participant asked for help.

VALIDITY CRITERIA

The validity of a conjoint analysis can be examined by different validity criteria. In general, the content validity and the criterion-related validity are used to examine the results of a conjoint analysis (Müller-Hagedorn, Sewing & Toporowski, 1993).

The content validity examines the plausibility, the completeness and the adequateness of an analysis (Albrecht, 2000). Thus among other things it is examined if the results of an analysis correspond with given expectations. One way of analyzing the content validity of a conjoint analysis is calculating the average relative importance of each attribute and to validate if the results meet a priori expectations. In addition, a priori expectations about the relations of the part-worths of different attribute levels can be formed. Verifying on an individual level if the relations of the estimated part-worths correspond with a priori expectations is another way of examining the content validity. A low amount of incorrect part-worth relations indicates a high content validity. However, it has to be noted that expectations about the relations of part-worths cannot be formed for every attribute. Thus the content validity gives only a first impression of the validity of the results.

The criterion-related validity refers to the relation of the predictor and the criterion. The criterion-related validity is typically split into the concurrent validity and the predictive validity (Müller-Hagedorn et al., 1993). The concurrent validity measures how well the estimated values reflect the input data. Thus the concurrent validity measures the internal consistency of the data. The correlation coefficients Pearsons R and Kendalls tau are used to measure the concurrent validity (Backhaus, Erichson, Plinke & Weiber, 2003). In addition, the First Hit Rate can also be used to measure the concurrent validity (Hensel-Börner, 2000). The First Hit Rate is based on First Choice Rule, that examines if the stimulus, which was mostly preferred and therefore set number one by the respondent, is the one with the highest utility.

The predictive validity refers to the ability of the estimated part-worths to predict real buying behavior (Backhaus & Brzoska, 2004). This implies that a researcher has to observe an actual purchasing/non-purchasing decision before he is able to measure the predictive validity. In order to get to know a buying decision, and if no "real" buying decision is available, so-called validation stimuli are used. Those are not used to estimate the part-worths. In addition, the predictive validity can be measured on an individual or on an aggregate level. The problem of measuring the predictive validity on an aggregate level is that one predicting error can be outweighted through another predicting error. The problem of measuring the predictive validity on an individual level is that test persons often give different answers to the same question (Riley, Ehrenberg, Castleberry, Barwise & Barnard, 1997).

One way of measuring the predictive validity on an individual level is to analyze the hit rate (Voeth, 2000). The hit rate examines if the real purchasing decision matches with the predicted purchasing decision. The Hit Rate is the quotient of the sum of buying decisions, which have been predicted correctly, and the sum of the total buying decisions.

$$HitRate = \frac{D_{correct}}{D_{total}}$$

$D_{correct}$ = Sum of correct predicted buying decisions of all test persons
D_{total} = Sum of all buying decisions of all test persons

On an aggregate level the predictive validity can be measured several ways. The different measures can be differentiated from each other by the variables, which are part of the equation. Some measures consist of empirical input data, estimated values and number of validation stimuli. Other measures also include a reference value. The Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) are two measures that quantify the predictive validity on the aggregate level. Both measures consist of the empirical buyer proportion, the estimated buyer proportion and the number of validation stimuli included (Brodie & Bonfrer, 1994; Danaher, 1994; Green & Krieger, 1996; Leeflang, Wittink, Wedel & Naert, 2000; Hanssens, Parsons & Schultz, 2001; Andrews, Ansari & Currim, 2002).

$$RMSE = \sqrt{\frac{\sum_i (D_i - \hat{D}_i)^2}{VS}} \qquad MAE = \frac{\sum_i \left| D_i - \hat{D}_i \right|}{VS}$$

D_i = Buyer proportion of the validation stimulus i
$\hat{D}_i$ = Estimated buyer proportion of the validation stimulus i
VS = Amount of validation stimuli

Lower values of RMSE and MAE indicate a higher predictive validity. Measures which include a reference value are the Relative Absolute Error (RAE) and Theils U (Leeflang et al., 2000; Hanssens et al., 2001). Since the purchase/non-purchase question has two possible answers (yes/no), normally 50 % is taken as a reference value for each validation stimulus. Lower values of RAE and Theils U indicate a higher predictive validity.

$$RAE = \frac{\sum_i \left| D_i - \hat{D}_i \right|}{\sum_i \left| D_i - \tilde{D}_i \right|} \qquad TheilsU = \sqrt{\frac{\sum_i \left(D_i - \hat{D}_i \right)^2}{\sum_i \left(D_i - \tilde{D}_i \right)^2}}$$

$\tilde{D}_i$ = Reference value for the validation stimuli i

STUDY DESIGN

To compare the validity of the discussed two conjoint analyses a product has to be chosen, that can easily be evaluated by the test persons. Therefore, a MP3 player was selected, which was characterized by the following attributes and their respective levels: size (stick or walkman), data capacity (1 GB, 5 GB or 10 GB), price (89.99 €, 139.99 € or 189.99 €) and display (black/white or coloured). This attributes have been state of the art at the time the survey was conducted. For a traditional conjoint analysis it is suggested to limit the number of presented stimuli to 20 (Voeth, 2000). Therefore, the amount of attributes did not exceed four and the number of attribute levels was limited to three. There are no suggestions concerning the amount of stimuli for an online-conjoint analysis or an interviewer-based computer survey with a full-profile-approach (Lines & Denstadli, 2004). However, it is evident that the complexity and the challenge increase with the number of stimuli.

In order to be able to compare the validity of both conjoint analyses, it had to be assured that the study design of the surveys was identical. Thus the interviewer-based computer survey and the online survey consisted of the same questions. The survey itself was divided into two parts. The attributes and their respective attribute levels have been presented in the first part of the questionnaire. The respondents were asked to rank the nine stimuli, which have been derived through an orthogonal main-effect design. Since ranking nine different stimuli is a fairly complicated task, it had to be assured that the test persons could visually move the stimuli (see Figure 1) on their computer screen before ranking them finally.

Figure 1: Stimuli of the Orthogonal Main-effect Design

Offer A	Offer B	Offer C	Offer D	Offer E
Size: **Walkman** 	Size: **Stick** 	Size: **Stick** 	Size: **Stick** 	Size: **Walkman** 
Data Capacity: **10 GB**	Data Capacity: **1 GB**	Data Capacity: **10 GB**	Data Capacity: **5 GB**	Data Capacity: **1 GB**
Price: **189,99 €**	Price: **189,99 €**	Price: **139,99 €**	Price: **139,99 €**	Price: **139,99 €**
Display: **Black/White**	Display: **Coloured**	Display: **Coloured**	Display: **Black/White**	Display: **Black/White**

Offer F	Offer G	Offer H	Offer I
Size: **Stick** 	Size: **Stick** 	Size: **Stick** 	Size: **Walkman** 
Data Capacity: **10 GB**	Data Capacity: **1 GB**	Data Capacity: **5 GB**	Data Capacity: **5 GB**
Price: **89,99 €**	Price: **89,99 €**	Price: **189,99 €**	Price: **189,99 €**
Display: **Black/White**	Display: **Black/White**	Display: **Black/White**	Display: **Coloured**

After having ranked the nine products, whereas "1" indicated the best offer and "9" the worst offer, the respondents were asked to specify up to which rank they would purchase the product. This procedure is known as placing of a "limit card". The limit card separates the product cards into two parts, whereas one part consists of products a test person would be willing to buy and the other part consists of products a respondent would not be willing to buy (Voeth & Hahn, 1998; Voeth, 2000; Sattler & Nitschke, 2003). The position of the limit card is also needed to measure the predictive validity. If the predictive validity is calculated by using a choice model (e. g. First Choice or Bradly-Terry-Luce) the result heavily depends on the chosen model (Hartmann & Sattler, 2004).

The validation stimuli have been special combinations of different attribute levels. However, they did not consist of extreme attribute levels, since the buying decision is generally easier to predict for such stimuli (Albrecht, 2000). In addition, they were not part of the presented nine stimuli. So the respondents did not need to rank the validation stimuli as they would have to do with holdout stimuli, but they had to declare a purchase or non-purchase decision. The reason for using validation stimuli instead of holdout stimuli is that validation stimuli generally simulate a buying decision better than holdout cards (Sattler et al., 2001).

The second part of the questionnaire consisted of three questions. The respondents were asked to express their opinion about the degree of difficulty of ranking the stimuli on a six point rating scale, whereas "1" referred to "completely agree" and "6" referred to "completely disagree". These results help interpreting the validity analysis (Hensel-Börner & Sattler, 2000; Hartmann & Sattler, 2004). Finally, the test persons were asked to answer a few statistical questions about their age, gender etc.

RESULTS

Basic Characteristics

Overall, 220 data sets could be used from the online survey and 317 data sets from the interviewer-based computer survey, which have been derived from a group of students at a German University. The two data sets were compared in regards to the age and gender of the participants, since it is critically discussed that the characteristics of the respondents influence the results of a survey (Tscheulin & Blaimont, 1993; Sattler et al., 2001; Sattler & Nitschke, 2003). From that viewpoint the formation of both of the samples is comparable concerning the average age and the gender of the participants. Nevertheless, it is stated, that a big amount of student participants (convenience sample) influences the transferability of a market research study. However, this does not cause any problems in this study, since the main goal of the study is to analyze differences in the validity of an online survey and an interviewer-based computer survey.

As mentioned above the test persons had to rate the difficulty of ranking the stimuli on a 6 point rating scale whereas "1" referred to "completely agree" and "6" referred to "completely disagree".

Figure 2: Evaluation of the Difficulty of Ranking the Stimuli

Item	Interviewer-based Survey	Online Survey
The Conjoint Analysis was easy to handle. *	Ø 2.40 (s.d. = 1.283)	Ø 2.68 (s.d. = 1.265)
The sorting of the stimuli was exhausting. ***	Ø 4.40 (s.d. = 1.721)	Ø 5.01 (s.d. = 1.541)
The number of the cards was not too high to be able to evaluate the products. n. s.	Ø 2.30 (s.d. = 1.290)	Ø 2.34 (s.d. = 1.458)

Figure 2 shows the arithmetic average (Ø) and the standard deviation (s.d.) of the answers. Since the respondents replied to the first and third question ("The conjoint analysis was easy to handle." and "The number of the cards was not too high to be able to evaluate the products") with a relatively "low number", it can be assumed that the presented conjoint analysis was cognitively not too difficult. However, the difference in the response (0.04) of the third question is not significant, whereas the relatively low difference between the interviewer-based computer and the online survey (0.28) for the first question is significant on the level $p \leq 0.05$. Moreover, participants of the interviewer-based computer survey find it more difficult to sort the cards on the PC screen than the participants of the online survey. A significance test shows that the difference is highly significant ($p \leq 0.001$). Hence, it can be stated that the participants of the interviewer-based computer survey feel the task to sort the cards more difficult than the participants of the online survey, even though there was an interviewer to help in case of any questions. This effect could be due to social effects or to a distraction by the interviewer (Duffy et al., 2005).

Content Validity

Among other things the content validity measures the plausibility of a study. A detailed evaluation of the content validity is to compare the relation of the estimated part-worths. There is no hypothesis possible about the relation of the part-worths of different sizes of the MP3 player

and no hypothesis about the part-worths of different colours. However, it is reasonable to expect that a higher price of a MP3 player will result in a lower part-worth than a lower price. Furthermore, it can be assumed that the part-worth of the data capacity will increase with an increasing amount of data capacity. Following these assumptions the relations of the part-worths can be examined on an individual level. The higher the amount of incorrect relations, the lower is the content validity. The results of the examination are shown by Figure 3.

Figure 3: Percentage of Correctly Predicted Part-worth Relations

Sample	Price Sample	Price Random Number Model	Data Capacity Sample	Data Capacity Random Number Model
Interviewer-based Survey	90.91 %	42.73 %	95.00 %	58.18 %
Online Survey	96.53 %	49.21 %	98.42 %	43.85 %

First of all it can be observed that the amount of correct predicted part-worth relations of the two criteria of the online survey (96.53 % – Price; 98.42 % – Data Capacity) are slightly higher than the two criteria of the interviewer-based computer survey (90.91 % – Price; 95.00 % – Data Capacity).

However, to be able to draw conclusions about the quality of the data collecting methods, the results have to be compared with results of other studies. To the best of our knowledge no comparable studies have been conducted before or are at least not well documented. Therefore, it is advisable to compare the results of the study with results, which are based on pseudo-random numbers (Huber et al., 1993; Hartmann, 2004). Therefore, two alternate samples with pseudo-random numbers and the identical number of data sets have been estimated. These pseudo-random numbers have been used as input data for two alternative conjoint analyses (interviewer-based computer random model and online survey random model). As it can be seen in Figure 3, both original samples generate much better results than the random number models. A two-sample test for the difference of two cumulated distributions shows if the percentages differentiate significantly from each other or not. This test was applied to compare the values of the samples with the values of the random number models (Bleymüller, Gehlert & Gülicher, 2004). As expected, the two-sample tests for the difference of two cumulated distributions indicate that all four differences of the percentages of correct predicted part-worth relations of the random number models and of the samples are highly significant on the level $p \leq 0.001$. Thus it can be concluded that the results derived from the interviewer-based computer and the online survey have a high content validity.

While comparing both differences of the percentages of correct predicted part-worth relations of the online survey and the interviewer-based computer survey it can be stated that those are very significant for the price ($p \leq 0.01$) and significant for the memory ($p \leq 0.05$). So it can be concluded that the influence of the interviewer has a negative effect in regards to the content validity.

Concurrent Validity

The concurrent validity examines if the measurement of the data is consistent and to what extent the empirical input data matches the estimated data (Backhaus et al., 2003). The results can be seen in Figure 4.

Figure 4: Average Values of Pearson's R and Kendalls tau

Sample	Pearson's R Sample	Pearson's R Random Number Model	Kendall's tau Sample	Kendall's tau Random Number Model
Interviewer-based Survey	Ø 0.9772	Ø 0.8511	Ø 0.9523	Ø 0.7238
Online Survey	Ø 0.9859	Ø 0.8673	Ø 0.9738	Ø 0.7523

It can be stated that the average values for Pearson's R and Kendall's tau are all very similar to each other. Besides that the values for Pearson's R and Kendall's tau are clearly above 0.9, which indicates a very strong correlation and thus a very high concurrent validity (Clarke, 1993; Fahrmeir, Künstler, Pigeot & Tutz, 2001). That implies that a statement about the superiority of one data collecting method is not possible at first hand.

Again, it is suggested to compare the results with results of a random number model (Weisenfeld, 1987). Figure 4 illustrates that the values of Pearson's R and Kendall's tau of the random number model are much lower than the respective values of the survey samples. A Kolmogoroff-Smirnov-two-sample-test was applied to determine if the differences between the values of the survey samples and the values of the random number models are significant (Brosius, 2004). The tests show that the differences are highly significant ($p \leq 0.001$). Hence it can be stated that the high values of Pearson's R and Kendall's tau are not that high by accident.

A Kolmogoroff-Smirnov-two-sample-test can also be applied to compare the results of the online survey with the results of the interviewer-based computer survey. The results show that the divergences of the distribution of the individual values are highly significant for Pearson's R ($p \leq 0.001$) and significant for Kendall's Tau ($p \leq 0.05$). Alternatively a Mann-Whitney test can be accomplished. To conduct this test only ordinal numbers are needed. The result of this test underlines the previous findings for the samples and the random number models.

A comparison of the match of the stimulus, which was set on the first rank by the respondents, with its corresponding utility (First Hit Rate) shows that the online survey results in a higher match (92.11 %) than the interviewer-based computer survey (84.55 %). So in only 7.89 % of the cases of the online survey the stimulus with the highest utility was not set on the first place whereas in the interviewer-based computer survey it was in 14.45 % of the cases. However, to be able to tell something about the quality of the First Hit Rate, a two-sample test for the difference of two cumulated distributions has to be applied to the surveys and random number models. In both cases the test shows a highly significant ($p \leq 0.001$) deviation between the samples and the random number models (First Hit Rate online: 50.47 %; First Hit Rate interviewer-based computer: 54.09 %). This indicates that the quality of the data of the sample tests is very good. In a next step both samples are compared to each other. The result is a very significant difference ($p \leq 0.01$) in both cases. This implies that the online survey also generates better values in regards to the First Hit Rate than the interviewer-based computer survey.

So it can be overall concluded that the online-conjoint analysis leads to better values in regards to the concurrent validity which was measured by Pearsons R and Kendalls tau as well as First Hit Rate. Thus, again the interviewer has a negative effect in regards to the validity.

Predictive Validity

The predictive validity measures the ability of the estimated values to predict real purchasing decisions. Since the respondents had to specify if they would purchase a certain hypothetical product (validation stimulus), the purchase decision of the consumer is known (Voeth, 2000). To predict the purchase decision, the part-worths of the attribute levels of the hypothetical product are added in order to compute the utility of each validation stimulus. The purchase decisions of each consumer can be predicted for each validation stimulus, by asking the test persons to set a limit card (Backhaus, Hillig & Wilken, 2007). By definition a utility higher than zero indicates a positive purchase decision and a utility lower than zero indicates that the consumer is not willing to buy the product (Voeth and Hahn, 1998). The predicted purchase decisions are compared with the stated purchase decision of each test person. After having compared the predicted decisions with the stated choices, the predictive validity can be measured on an individual as well as an aggregate level.

The Hit Rate measures the predictive validity on an individual level (Voeth, 2000). The estimation comprises of all correct predicted purchase decisions in the numerator as well as of all purchase decisions in the denominator. Since the test persons were asked to express their purchase decision for two validation stimuli the Hit Rate is calculated for each stimulus individually and for both stimuli together as shown by Figure 5.

Figure 5: Hit Rates of the Validation Stimuli

Hit Rate	Validation Stimulus I Sample	Validation Stimulus II Sample	Total Hit Rate Sample	Validation Stimulus I Random Number Model	Validation Stimulus II Random Number Model	Total Hit Rate Random Number Model
Interviewer-based Survey	74.09 %	64.09 %	69.09 %	48.18 %	52.27 %	50.23 %
Online Survey	84.86 %	61.83 %	73.34 %	45.43 %	47.63 %	46.53 %

It is to be noted that the amount of correct predicted purchase decisions is much lower for the second validation stimulus than for the first validation stimulus. This might be caused by inconsistent response behavior of the test persons, whose answering behavior for repeated questions is not the same as their answering behavior for earlier questions or in contrast with their ranking decision. Such a change in the answering behavior is supported by empirical studies (Riley et al., 1997) and has therfore been confirmed within this study. Because of the fairly low amount of correct predicted purchase decisions of the second stimulus the Total Hit Rate of the interviewer-based computer survey is below 70 %. In comparison the Total Hit Rate of the online survey is slightly better (73.34 %). However, altogether the data quality of both surveys is good (Srinivasan and Park, 1997).

To evaluate the quality of the Total Hit Rates, they are compared with the Total Hit Rates of the above used random number models. Therefore several two-sample tests for the difference of two cumulated distributions are conducted to analyze the differences between the Hit Rates of the samples and the random number models. The difference of both samples and the random number models are highly significant ($p \leq 0.001$). This indicates good input data of the surveys.

The two-sample test for the difference of two cumulated distributions of the different Total Hit Rates of the online survey and the interviewer-based computer survey indicate no significant difference. Thus it cannot be concluded that there is a difference in the predictive validity.

Rather both data collecting methods lead to the same data quality. However, there are differences, if the first and second validation stimuli are examined separately. The difference between the online and the interviewer-based computer survey is very significant ($p \leq 0.01$) for the first validation stimulus but not significant for the second validation stimulus.

In addition to the Hit Rate, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) are calculated to examine the predictive validity of the two data collecting methods on an aggregate level. The results are shown by Figure 6. It is observable that all values of RMSE and MAE are very close together.

Figure 6: Aggregated Error Measures

Sample	RMSE	MAE	RAE	Theils U
Interviewer-based Survey	0.2138	0.2136	1.7736	1.3532
Online Survey	0.2308	0.2098	0.8365	0.8499

The remarks to RMSE and MAE are to be criticized in that way that an evaluation and an interpretation of the error measures concerning their absolute height is difficult. Because of that it is suggested to compare the RMSE and MAE of a sample with the RMSE and MAE of a random number model (Hartmann, 2004). However, it is highly problematic to compare the RMSE and the MAE of a sample with the RMSE and the MAE of a random number model, since the value of RMSE and MAE of a random number model depend on the stated purchase decisions of the sample. The predicted buyer proportion of a random number model will always be around 50 %. If the stated purchase decisions are also 50 % the RMSE and the MAE of the random number model will be close to zero or exactly zero. Hence the RMSE and the MAE of the random number model highly depend on the amount of stated purchase decisions of the sample and therefore it is not advisable to compare the RMSE and the MAE of a random number model with the RMSE and the MAE of a sample. That is the reason why RMSE and MAE are not calculated with random numbers.

Apart from RMSE and MAE the Relative Absolute Error (RAE) and Theils U are two error measures for the predictive validity on an aggregate level. The estimation of RAE and Theils U include a reference value for the predicted purchase decision of each validation stimuli (Leeflang et al., 2000; Hanssens et al., 2001). The reference value for each validation stimulus is set to 50 %. It is observable that the error measures of the online survey are lower than the error measures of the interviewer-based computer survey. Partially the measures of the interviewer-based computer survey are twice as high as the measures of the online survey. This result again supports the previous findings that the interviewer has a negative effect on the validity and therefore reducing the predictive validity of the sample.

CONCLUSIONS AND FURTHER RESEARCH

The goal of this paper was to examine the influence of an interviewer on different validity criteria. By analyzing the different validity measures of an interviewer-based computer survey and an online survey it became obvious that the content validity of the online survey is slightly higher.

The concurrent validity is pretty high for both methods. However, because of the significant lower values of Pearsons R and Kendalls tau in regards to the concurrent validity an inter-

viewer-based computer survey should not be conducted. This finding is also underlined by the First Hit Rate. By applying a two-sample test for the difference of two cumulated distributions, it could be shown that the deviation of the two samples generated by an interviewer-based computer survey and an online survey is very significant. This implies that online methods generate more valid results concerning concurrent validity.

One result of analyzing the predictive validity is that the second stimulus could not predicted as well with the underlying utility model as the first stimulus, which resulted to a lower Total Hit Rate. Nevertheless, no significant difference could be found for the Total Hit Rate between both of the samples, since the absolute numbers of the total Hit Rates are very similar. This implies that in regards to the Hit Rate both methods are equally good to gather data for a conjoint analysis. However, one difference is in regards to the error measures for the predictive validity. While the values of RMSE and MAE are comparable, the values of RAE and Theils U differ from each other quite a lot. If a reference value of 50 % (RAE and Theils U) is used, again the data of the online survey leads to better results.

The results of this study show, that the presence of an interviewer influences the data negatively. One reason for this finding could be that the participants of the interviewer-based computer survey are feeling observed and controlled by the interviewer and hence are in a greater stress situation than the participants of the online survey (Zou, 1999; Theobald, 2000; Duffy et al., 2005). These social effects and the distraction of the interviewer lower the validity of the study. So although a conjoint analysis is a rather complex method and gathering the data for the analysis is not as simple as gathering data for a regular questionnaire, it can be concluded that the lack of an interviewer does not necessarily result into a lower validity. Rather it could be shown that the results of an online-conjoint analysis are of a higher validity than the results of a conjoint analysis, which was based on a data sample that was gathered on a PC and with the help of an interviewer. However, since the validity of both samples are rather high, it cannot be concluded that gathering data on a PC and with the help of an interviewer should not be done. Rather more research should be done to get more information in regards to the influence of an interviewer.

Generally the advantages of online surveys are widely discussed in the literature and are holding true for an online-conjoint analysis. In fact the analysis shows that the choice of a special data collecting method should not depend on statistical criteria but on the purpose of the examination. If only a small sample size is needed for conducting a conjoint analysis (e. g. the market segment for a new product is very small and a researcher is only interested in the data of just a few important people) a survey with the help of an interviewer can be conducted. However, if the goal of a conjoint analysis is to conduct market simulations there is no reason of not collecting the data online, since a larger sample size can be achieved more easily. If the representativeness of a sample is assured a bigger sample size might even lead to better results of marked simulations etc.

In regards to future research and in regards to compare both methods in more detail, the amount of attributes should be increased. This would affect the orthogonal main-effect design and the amount of stimuli. Although the participants of the online survey evaluated the interview process to be less exhausting than the participants of the interviewer-based computer survey, it can be assumed that a higher amount of stimuli increases the complexity and the difficulty and therefore it can be argued that most probably the validity will decrease (Lines & Denstadli, 2004). The literature argues for example that the validity decreases dramatically with the amount of 20 stimuli (Büschken, 1994). In addition, it is assumed that a stimulus has about four to five

attributes (Perrey, 1998; Voeth, 2000). In this case the interviewer could possibly affect the results in a positive way.

Future research should be done in regards to the influence of the participants on the validity. The participants of the examined convenience sample (age and educational background) has most probably a better understanding in regards to a conjoint analysis (Tscheulin & Blaimont, 1993; Sattler et al., 2001; Sattler & Nitschke, 2003). However, the authors are of the opinion that the fact that a lot of students participated in the survey affects the analysis of the influence of an interviewer on the validity only marginally. Beyond that it could also be examined whether a multimedia presentation of the stimuli affects the validity (Ernst & Sattler, 2000).

REFERENCES

Agarwal, M. K., & Green, P. E. (1991). Adaptive Conjoint Analysis versus self-explicated models: Some empirical results. *International Journal of Research in Marketing*, 8, 141-146.

Albrecht, J. (2000). *Präferenzstrukturmessung*. Frankfurt a. M.: Peter Lang.

Andrews, R. L., Ansari, A., & Currim, I. S. (2002). Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery. *Journal of Marketing Research*, 39, 87-98.

Backhaus, K., & Brzoska, L. (2004). Conjointanalytische Präferenzmessungen zur Prognose von Preisreaktionen. *Die Betriebswirtschaft*, 64, 39-57.

Backhaus, K., Hillig, T., & Wilken, R. (2007). Predicting Purchase Decisions with Different Conjoint Analysis Methods. *International Journal of Market Research*, 49(3), 341-364.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden* (10th ed.). Berlin et al.: Springer.

Backhaus, K., Wilken, R., Voeth, M., & Sichtmann, C. (2005). An Empirical Comparison of Methods to Measure Willingness to Pay by Examining the Hypothetical Bias. *International Journal of Market Research*, 47(5), 543-562.

Bamert, T., & Heidingsfelder, M. (2001). Designeffekte in Online-Umfragen. In: A. Theobald, M. Dreyer, & T. Starsetzki (Eds.). *Online-Marktforschung* (pp. 165-178). Wiesbaden: Gabler.

Bleymüller, J., Gehlert, G., & Gülicher, H. (2004). *Statistik für Wirtschaftswissenschaftler* (14th ed.). München: Vahlen.

Büschken, J. (1994). Conjoint-Analyse – Methodische Grundlagen und Anwendungen in der Marktforschungspraxis, in: T. Tomczak, & S. Reinecke (Eds.), *Marktforschung* (pp. 72-89), St. Gallen: Thexis.

Brodie, R. J., & Bonfrer, A. (1994). Conditions When Market Share Models are Useful for Forecasting. *International Journal of Forecasting*, 10, 277-285.

Brosius, F. (2004). *SPSS 12*. Bonn: Mitp.

Clarke, D. G. (1993). *Marketing Analysis and Decision Making* (2nd ed.)., Redwood City.

Couper, M., Tourangeau, R., & Kenyon, K. (2004). Picture This! Exploring Visual Effects in Web Surveys. *Public Opinion Research*, 68, 255-266.

Daiber, A., & Hemsing, W. (2005). Online Conjoint: Eine bewährte Methode im neuen Gewand. *Planung & Analyse*, 32(1), 47-52.

Danaher, P. J. (1994). Comparing Naive with Econometric Market Share Models When Competitors' Actions Are Forecast. *International Journal of Forecasting*, 10, 287-294.

Dibb, S., Rushmer, A., & Stern, P. (2001). New Survey Medium: Collecting Marketing Data with E-mail and the World Wide Web. *Journal of Targeting, Measurement and Analysis for Marketing*, 10, 17-25.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing Data from Online and Face-to-face Surveys. *International Journal of Market Research*, 47, 615-639.

Ernst, O., & Sattler, H. (2000). Multimediale versus traditionelle Conjoint-Analysen: Ein empirischer Vergleich alternativer Produktpräsentationsformen. *Marketing – Zeitschrift für Forschung und Praxis*, 22, 161-172.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2001). *Statistik*. Berlin et al.: Springer.

Gierl, H., & Höser, H. (2002). Der Reihenfolgeeffekt auf Präferenzen. *Zeitschrift für betriebswirtschaftliche Forschung*, 54, 3-17.

Görts, T., & Behringer, T. (2003). Online Conjoint – Chancen und Grenzen, in: A. Theobald, M. Dreyer, & T. Starsetzki (Eds.). *Online-Marktforschung* (2nd ed.) (pp. 283-196). Wiesbaden: Gabler.

Grant, D., Teller, C., & Teller, W. (2005). "Hidden" Opportunities and Benefits in Using Web-based Business-to-business Survey. *International Journal of Market Research*, 47, 641-666.

Green, P. E., & Krieger, Abba M. (1996). Individualized Hybrid Models for Conjoint Analysis. *Management Science*, 42, 850-867.

Green, P. E., Krieger, Abba M., & Wind, Y. J. (2001). Thirty Years of Conjoint Analysis – Reflections and Prospects. *Interfaces*, 31, 56-73.

Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2001). *Market Response Models* (2nd ed.). Boston et al.: Kluwer Academic Publishers.

Hartmann, A. (2004). *Kaufentscheidungsprognose auf Basis von Befragungen*. Wiesbaden: Gabler.

Hartmann, A., & Sattler, H. (2004). Wie robust sind Methoden zur Präferenzmessung? *Zeitschrift für betriebswirtschaftliche Forschung*, 56, 3-22.

Hauser, J. R., & Toubia, O. (2005). The Impact of Utility Balance and Endogeneity in Conjoint Analysis. *Marketing Science*, 24, 498-507.

Henning-Thurau, T., & Dallwitz-Wegener, D. (2002). Online-Befragungen – Ein neues Instrument für die Marktforschung. *Wirtschaftswissenschaftliches Studium*, 31(6), 309-314.

Hensel-Börner, S. (2000). *Validität computergestützter hybrider Conjoint-Analysen*, Wiesbaden. Gabler.

Hensel-Börner, S., & Sattler, H. (2000). Ein empirischer Vergleich zwischen der Customized Computerized Conjoint Analysis (CCC), der Adaptiven Conjoint Analysis (ACA) und Self Explicated-Verfahren. *Zeitschrift für Betriebswirtschaft*, 70, 705-727.

Huber, J., Wittink, D. R., Fiedler, J. A., & Miller, R. (1993). The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice. *Journal of Marketing Research*, 30, 105-114.

Ilieva, J., Baron, S., & Healey, N. M. (2002). Online Surveys in Marketing Research: Pros and Cons. *International Journal of Market Research*, 44, 361-376.

Leeflang P. S. H., Wittink, D. R., Wedel, M., & Naert, P. A. (2000). *Building Models for Marketing Decisions*. Boston et al.: Kluwer Academic Publishers.

Lines, R., & Denstadli, J. M. (2004). Information Overload in Conjoint Experiments. *International Journal of Market Research*, 46(3), 297-310.

Müller-Hagedorn, L., Sewing, E., & Toporowski, W. (1993). Zur Validität von Conjoint-Analysen. *Zeitschrift für betriebswirtschaftliche Forschung*, 45, 123-148.

Perrey, J. (1998). *Nutzenorientierte Marktsegmentierung*. Wiesbaden: Gabler.

Riley, D., Ehrenberg, A. S. C., Castleberry, S. B., Barwise, T. P., & Barnard, N. R. (1997). The Variability of Attidutinal Repeat. *International Journal of Research in Marketing*, 14, 437-450.

Sattler, H., & Nitschke, T. (2003). Ein empirischer Vergleich von Instrumenten zur Erhebung von Zahlungsbereitschaften. *Zeitschrift für betriebswirtschaftliche Forschung*, 55, 364-381.

Sattler, H., Hensel-Börner, S., & Krüger, B. (2001). Die Abhängigkeit der Validität von Conjoint-Studien von demographischen Probanden-Charakteristika. *Zeitschrift für Betriebswirtschaft*, 71, 771-787.

Schillewaert, N., & Meulemeester, P. (2005). Comparing Response Distributions of Offline and Online Data Collection Methods. *International Journal of Market Research*, 47, 163-178.

Sethuraman, R., Kerin, R. A., & Cron, W. L. (2005). A Field Study Comparing Online and Offline Data Collection Methods for Identifying Product Attribute Preferences Using Conjoint Analysis. *Journal of Business Research*, 58, 602-610.

Srinivasan, S. V., & Park, C. S. (1997). Surprising Robustness of Self-Explicated Approach to Customer Preference Structure Measurement. *Journal of Marketing Research*, 34, 286-291.

Strebinger, A., Hoffmann, S., Schweiger, G., & Otter, T. (2000). Zur Realitätsnähe der Conjointanalyse. *Marketing – Zeitschrift für Forschung und Praxis*, 22, 55-74.

Theobald, A. (2000). *Das World Wide Web als Befragungsinstrument*. Wiesbaden: Gabler.

Tscheulin, D., & Blaimont, C. (1993). Die Abhängigkeit der Prognosegüte von Conjoint-Studien von demographischen Probanden-Charakteristika. *Zeitschrift für Betriebswirtschaft*, 63, 839-846.

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

Voeth, M. (2000). *Nutzenmessung in der Kaufverhaltensforschung*. Wiesbaden.

Voeth, M., & Hahn, C. (1998). Limit Conjoint-Analyse. *Marketing – Zeitschrift für Forschung und Praxis*, 20, 119-132.

Vriens, M., Looschilder, G. H., Rosenbergen, E., & Wittink, D. R. (1998). Verbal versus Realistic Pictorial Representations in Conjoint Analysis with Design Attributes. *Journal of Product Innovation Management*, 15, 455-467.

Weisenfeld, U. (1987). Signifikanztest für die Anpassungsgüte in Conjoint-Analysen. *Marketing – Zeitschrift für Forschung und Praxis*, 9, 267-270.

Welker, M., Werner, A., & Scholz, J. (2005). *Online Research*. Heidelberg: DPunkt.

Woratschek, H. (2001). Preisbildung im Dienstleistungsbereich auf der Basis von Marktinformationen. In: M. Bruhn, & H. Meffert (Eds.). *Handbuch Dienstleistungsmanagement* (pp. 607-626). Wiesbaden: Gabler.

Zou, B. (1999). *Multimedia in der Marktforschung*, Wiesbaden: Gabler.