

Online interviews and data quality: A multitrait-multimethod study

Draft paper to be presented at the MESS Workshop, 22-23 August 2008, Zeist.

Annette Scherpenzeel
a.c.scherpenzeel@uvt.nl
CentERdata, Tilburg University
Koopmans Building room K639
P.O. Box 90153
Warandelaan 2
5000 LE Tilburg
The Netherlands

Abstract

In this study, a split-ballot mtmm design is used to compare the quality of the data collected by web interviews with the quality of the data collected by traditional data collection methods. A probability sample of the Dutch population was contacted and interviewed by one of two traditional computer assisted data collection modes: telephone interviewing or face-to-face interviewing. At the end of these interviews, the respondents were asked to become a member of an online household panel. Respondents who were willing to participate but had no internet access were equipped with a user-friendly computer with internet access. All interviewed respondents who joined the panel were asked to fill in the same interview in a web interview format, a few months after the first telephone or face-to-face measurement. The mtmm analysis gives estimates of both reliability as the complement of random error variance and validity as the complement of systematic method variance for every measure in the study. These estimates show that panel data collected using a self-administered web interview are at least as valid and reliable as data collected using more traditional modes of interviewing, when the sample is kept constant.

Introduction

Opinion- and market research via the internet has increased significantly in recent years, but scientific panel studies are still often carried out using traditional data collection methods. One obstacle to use web-interviews for this type of study is the lack of knowledge about the quality and comparability of the data that is obtained with this new method of interviewing. A few methodological studies have indicated what differences can be expected in coverage, response rates and sampling bias. However, differences in data quality caused by variation in the data collection instruments can also make the data incomparable. It is possible, for example, that respondents in a telephone interview use the end categories of scales more often than respondents in a web-interview. This systematic method variance, which Groves (1989) identified as “observational measurement error”, has to be taken into account when comparing data obtained with different instruments.

In this study, we compare the quality of the data obtained from the same respondents by three different computer assisted modes of interviewing: Computer assisted telephone interviewing (Cati), Computer assisted personal interviewing (Capi) and Computer assisted web interviewing (Cawi). All three modes were used within the framework of a panel study, in which the panel members answer questionnaires at regular intervals. Disregarding differences in coverage, response rates and sampling bias, the main differences between these modes are the interviewer presence, the form or presentation of the questions and the available time for the interview. Whereas the Cawi interview is a self-administered interview, both Capi en Cati are interviewer administered. Still, the lack of visual contact might give the telephone interview a somewhat higher degree of anonymity than the face-to-face interview. Furthermore, the questions, scales and information are read to the respondents in the Cati interview, while in the Cawi they are presented visually. The Capi interview is generally read to the respondents, but interviewers have the possibility to show the screen to the respondent as well. Finally, the modes differ in the amount of time the respondents have or feel they have to answer the questions. Silent pauses in a telephone conversation are generally felt as uncomfortable, both for respondents and interviewers. The Cawi interview is filled in at home, at any time and at the pace the respondents wishes. There is no pressure to keep the interview short or to avoid pauses. In addition, respondents can take a break and continue with the interview at some other time. The time pressure in a Capi interview is probably lower than in a telephone conversation, but most people will not want an interviewer in their home for hours.

Quality criteria

In this study, a combination of a split-ballot and a multitrait-multimethod (mtmm) design is used to compare the data quality obtained with web interviews with that obtained with traditional data collection methods. The design is somewhat related to the test-retest approach since it consists of repetitions of the same questions to the same people. However, the repetitions are carried out in a different mode of interviewing, thus systematically varying one characteristics of the survey measures.

In our study, we measured each of a number of traits with different methods. This design was introduced by Campbell and Fiske (1959). Saris and Andrews (1991) proposed the true score mtmm model for classical mtmm experiments. This model can be formulated as follows (for more detail we refer to Saris and Andrews, 1991 and Scherpenzeel and Saris, 1997):

The response y on item i can be decomposed into a random component, e_i , and a stable component T_i , which is called the "true score" in classical test theory (Heise and Bohrnstedt, 1970; Lord and Novick, 1968). If the response variable and the variable representing the stable component are standardised, equation (1) results:

$$y_i = h_i T_i + e_i \quad (1)$$

Here h_i represents the strength of the relationship between the stable component, or true score, and the response. The true score can be further decomposed into a component representing the score on the variable of interest, F_j , and a component, M_k ,

due to the method used. After standardisation, this leads to the formulation of equation (2):

$$(2) \quad T_i = b_{ij}F_j + g_{ik}M_k$$

Here b_{ij} represents the strength of the relationship between the latent variable of interest and the true score, and g_{ik} indicates the effect of the method on the true score. All three variables are standardised so the three coefficients mentioned are also standardised. Furthermore, we assume, as is normally done, that the correlations between the disturbance variables and the explanatory variables in each equation and across equations is zero, and that the trait factors are correlated but that the method factors are not correlated with each other nor with the trait factors. If all variables except the disturbance terms are standardised, the coefficients of the model have a special interpretation. In this case, h_i is called the "reliability coefficient" and its square is an estimate of test-retest reliability in the sense of the classical test theory; b_{ij} is called the "true score validity coefficient" because the square of this coefficient is the variance in the true score explained by the variable of interest; g_{ik} is called the "method effect" because the square of this coefficient is the variance in the true score explained by the method used.

For identification purposes the study should include at least three traits and three methods. Equation (2) represents the basic equation of the classical MTMM model and generates the factor structure presented in table 1. The model is illustrated in figure 1.

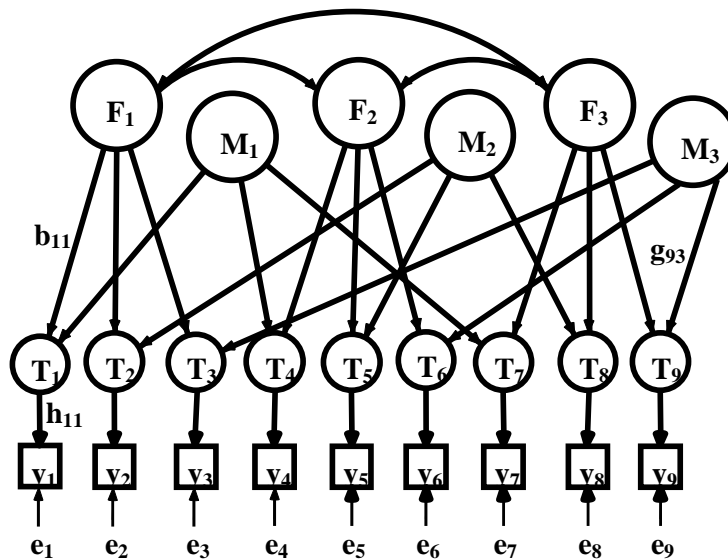


Figure 1. The standard true-score MTMM model for three traits and three methods.

Because the CFA analysis of this factor structure often leads to problems of identification and convergence, many different assumptions and restrictions have been made in the past, each leading to a slightly different model specification and interpretation of the parameters. Saris and Andrews (1991) suggested making the assumption that the method effects are the same for all traits affected by the same method. If this assumption is made, all parameters of this model are identified and no

convergence problems occur in estimation, as has been shown by Corten et al. (2002). The model can be estimated using programs for structural equation models. The major problem of this design is that the respondents have to answer the same questions approximately three times.

Experimental design

The methods in our study are the three interview modes, Cati, Capi and Cawi. For a complete mtmm design, all questions have to be asked in each interview mode. However, we used a combination of a split ballot design and the mtmm design, as shown in table 1. In order to reduce the respondent burden, Saris et al (2004) have proposed the split ballot mtmm experiment in which repetitions of the questions are spread over two identical subsamples from the same population. In the present study, the split ballot design was well-suited to the general design of the panel recruitment during which the data were collected (see description of sample and fieldwork below). When the assignment of individuals to groups has been made at random and there is a large sample in each group, the split ballot mtmm model can be estimated using multiple group estimation, which is available in most software for structural equation modelling. The model can also be estimated by considering the split ballot data as a missing data problem and use Full Information Maximum Likelihood (FIML), which is also available in most software packages nowadays. However, Saris et al (2004) suggest to use the multiple-group approach since the standard errors and test statistics derived from the FIML approach are not robust for nonnormality. For our model, we tested both estimation procedures and obtained identical estimates of validity, reliability and method effects. We use the standardized solution for the coefficients presented, obtained with the FIML and multiple-group estimation in LISREL 8 for Windows (Jöreskog and Sörbom, 2007)

Table 1 Experimental design

	Feb-Nov 2007	Nov 2007	Population	Net n
Group 1	Cati	Cawi	Telnr	2038
Group 2	Capi	Cawi	Telnr + notelnr	111 + 642

Sample and fieldwork

We used data from the newly build household panel in the Netherlands, the LISS panel (Longitudinal Internet Studies for the Social sciences). The LISS panel is an online panel based on a probability sample. It includes households that had no internet access before they participated in the panel. Households were contacted in traditional ways, by telephone and by face-to-face contacts, and households with no internet access were equipped with a user-friendly computer with internet access by the research institute. De data for the present study were collected during the recruitment stage of the panel, in which respondents were first contacted by telephone or in face-to-face contact and next started to participate in online interviews.

The reference population for the LISS panel is the Dutch speaking population permanently residing in the Netherlands. The sampling and survey units are independent, private households, thereby excluding institutions and other forms of collective households. Households in which no adult is capable of understanding the Dutch language are not included in the reference population. The sample frame was the nationwide address frame of Statistics Netherlands. In co-operation with Statistics Netherlands a simple random sample of addresses was drawn from this address frame. The sample unit is defined as the address because the intention was to build a household panel including all members of a household living at a given address. An announcement letter was sent to all households in the sample in combination with a brochure explaining the nature of the panel study. A 10 euro note was included with the letter. Next, respondents were contacted by an interviewer in a mixed mode design. Those households for which a landline telephone number was known were contacted by telephone¹. The remaining households were visited by an interviewer and thus contacted face-to-face. This mixed mode design implies that the Cati and Capi interviews were carried out among different subpopulations and the assumption of random assignment to groups does not hold. The subpopulation without or with unknown landline telephone number is probably distinct in many characteristics. We thus decided to contact, in an experimental pilot study, one group of 200 regular telephone-possessing households² in Capi instead of by telephone. At the end of the Cati and Capi interviews, the respondents were asked to become a member of an online household panel. Respondents who were willing to participate but had no internet access were equipped with a user-friendly computer with internet access. All interviewed respondents who joined the panel were asked to fill in the same interview in a web interview (Cawi). The time between the first Cati or Capi interview and the subsequent Cawi interview varied between two and 9 months. The first panel members were recruited in the pilot study, carried out from March 2007 to April 2007. The main sample was recruited from May until November 2007 with an average of about 500 households per month. The Cawi interview was carried out in November 2007 among all panel members. At that time, 6537 households who were recruited by Cati or Capi were participating in the online panel. The Cawi interview was completed by 4825 persons of 16 years or older within these households. Controlling for household identification number, sex and birth date we were able to identify 2924 persons who answered the Cati or Capi interview as well as the Cawi interview. We analysed the mtmm model for this total net sample of 2924 and for a subselection of panel members who were recruited in October and November, to estimate the effects of actual change in opinions and attitudes over the long period of time between the repetitions. In addition, we estimated the model once more excluding the respondents without a known landline number, to estimate the effects of subpopulation differences.

Interview and traits

The 10-minute recruitment interview was designed to give respondents an impression of what the panel interviews would be like, including questions about demographic characteristics, health, social integration, political interest, leisure activities, survey

¹ Only about 70% of the Dutch population still has a known, regular telephone connection, and that percentage is decreasing at a rapid pace.

² 200 as a brut sample n. The net sample n in the final match with the Cawi interview was 111.

attitudes, loneliness and personality. All questions except the demographics were repeated in the Cawi panel questionnaire. We selected three sets of three questions each for the mtmm analyses. Questions were selected on the basis of the correlations between them, because too low trait correlations can lead to –empirical- identification problems in split ballot mtmm models with only two groups (see Saris et al 2004). The first set of traits consisted of: Satisfaction with life; Feelings about oneself; Trust in one’s capacities. The second set was: Interest in news; Number of media in which news is followed; Interest in politics. The traits in the third set were: Voluntary work; Active membership of organisations; Hours sport per week.

Results

For each question in the experiment, an estimate of validity, reliability, and method effect was obtained. This resulted in 27 validity coefficients, 27 reliability coefficients and 27 method effect coefficients in total³. It is difficult to infer the differences between the modes over all the traits by just looking at these coefficients and it is not a very systematic way. Therefore, we carried out a secondary analysis with the obtained quality estimates as the dependent variables, to explain their variation by the different methods used. For this secondary analysis, we use the ANOVA procedure. The dependent variables in the meta-analysis are the validity and the reliability estimates obtained in the MTMM-analyses⁴. The mode of interviewing and the topic of the question are entered in the analysis as additive factors (table 2). The eta-squared statistic describes the proportion of total variability attributable to each factor. The effect of each factor level is shown by the marginal mean validity and reliability estimates. The adjusted R squared in the last row of the table indicates the amount of variance explained by all factors together.

Both the mode of interviewing and the topic of the questions have significant effects on validity and reliability. Interviewing by Capi causes a small but significant decrease in validity compared to the other two modes (.91 versus .94 and .93). It also has substantially lower reliability (.67 compared to .76 and .82). The topic of the question has a very strong effect on the validity coefficients. We see that questions about people’s actual activities, such as how much they follow the news; whether they are member of an organisation; or the number of hours they sport, have a higher validity than the personality and satisfaction questions. For reliability, the effect is somewhat weaker and we only see a (significant) difference between satisfaction and the other traits. Of course, some real change in the traits could have occurred during the time between the repeated measurements, especially since this time was rather long for some respondents. In addition, the lower validity and reliability of the Capi mode of interviewing could have been caused by the characteristics of the subpopulation without a known landline number that was mainly interviewed in this mode. We therefore performed another secondary analysis in which we took these factors into account. The results of this analysis, presented in table 3 in Appendix 1, show that the lower validity and reliability of the Capi mode still exists when subpopulation is kept constant. Furthermore, the results make it unlikely that the differences in validity and reliability between the topics are caused by changes over time (assuming that the one to two month interval would show fewer changes).

³ 3 sets of 3 traits x 3 methods

⁴ The method effect are not presented as dependent variables here, since they are the complement of the validity coefficients and show exactly the same variation.

Table 2. Secondary analysis of MTMM estimates: effects of interview mode and topic

Factor	N ¹	Validity coefficient		Reliability coefficient	
		Estimated marginal mean ²	Eta sqd	Estimated marginal mean	Eta sqd
Interview mode			.26*		.59*
Cati	9	.94		.76	
Capi	9	.91		.67	
Cawi	9	.93		.82	
Topic			.87*		.40*
Interest in news and politics	9	.98		.74	
Personality	6	.90		.78	
Life satisfaction	3	.79		.66	
Societal participation	6	.98		.79	
Sport participation	3	.99		.79	
Grand mean (sd)		.93 (.01)		.75 (.01)	
Adjusted R Squared			.84		.58

* = $p < 0.05$.

¹This N is the number of estimates obtained from the different mtmm analyses. It is not the number of respondents on which the mtmm estimates were based.

²Standard errors for the marginal means of the validity coefficient range from .01 (when N is 9) to .02 (when N is 3). Standard errors for the marginal means of the reliability coefficient range from .02 to .03.

Conclusions

Using a multitrait-multimethod design, estimates of reliability as the complement of random error variance and validity as the complement of systematic method variance were obtained for all data. Comparing these estimates, we can conclude that the choice of Cawi as method of interviewing has no negative impact on the data quality, defined as validity and reliability. With regard to validity, Cawi is comparable to Cati and slightly better than Capi. With regard to reliability, Cawi data are clearly superior to data collected using Cati or Capi.

In conclusion, panel data collected using a self-administered Cawi interview are at least as valid and reliable as data collected using more traditional modes of interviewing, when the sample is kept constant.

On the one hand, this result is encouraging for all the survey research that is done via the internet in recent years. On the other hand, the study shows that error variance due to data collection techniques cannot be ignored and can make mixed mode panel data incomparable.

References

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multimethod-multitrait matrix. *Psychological Bulletin*, 56, 833-853.

Corten I., Saris, W.E., Coenders, G., Van der Veld, W., Aalbers, W. and Kornelis, C. (2002). Fit of different models for multitrait-multimethod experiments. *Structural Equation models* 9: 213-233.

Heise, D.R. and Bohrnstedt, G.W. (1970). Validity, invalidity, and reliability. In: Borgatta, E.F. and Bohrnstedt, G.W. (Eds.). *Sociological methodology*. San Francisco, Jossey-Bass.

Lord, F. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA, Addison-Wesley.

Groves, R.M. (1989). *Survey errors and survey costs*. New York, Wiley and Sons.

Saris, W.E., Satorra, A. and Coenders, G. (2004) A new approach to evaluating the quality of measurement instruments: The split ballot MTMM design. In *Sociological Methodology* 2004, 311-347.

Saris, W.E. and Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modelling approach. In: Biemer, P.P., Groves, R.M., Lyberg, L.E. Mathiowetz, N. and Sudman, S. (Eds.). *Measurement errors in surveys*. New York, Wiley and Sons.

Scherpenzeel, A.C. and Saris, W.E. 1997. The validity and reliability of survey questions: a meta-analysis of Multitrait-Multimethod studies. *Sociological Methods and Research* 25, 341-383.

Appendix 1. Controlling for change over time and characteristics of subpopulation

In this analysis, we entered the same validity and reliability estimates as in the analysis presented in table 2, but also validity and reliability estimates obtained from additional analyses with the same mtmm model on 1) a subselection of respondents who were interviewed in Cati or Capi within one tot two months before the Cawi interview and 2) a subselection of respondents with known landline number. This secondary analysis was based on 72 estimates⁵ instead of the 27 of table 2, because the mtmm estimates were obtained from repeated analyses on partly overlapping samples.

It was not possible to construct mutually exclusive subsamples because the time between the measurements was not experimentally manipulated but posterior included as a control variable. Due to that, it was not completely crossed with the factor subpopulation, which was manipulated as experimental factor in the pilot study from March to April. The overlapping samples might have caused some dependencies in the estimates obtained from the mtmm models and hence the effect sizes in the secondary analysis in table 3 might not be accurate. However, the goal of this analysis was not to obtain an accurate estimate of the effects of time and subpopulation, but to test whether the differences between the interview modes in table 2 were not caused by differences in subpopulation or by change over time.

The results show that neither a shorter time between the measurements nor restricting the analysis to the landline-possessing subpopulation resulted in significant differences in validity and reliability coefficients. The main interest of this analysis is the interaction between mode and the two control factors. The table shows that the interactive effect of mode and time between the measurements was not significant, nor the interactive effect of mode and subpopulation.

⁵ 27 measures x 3 samples (total sample; subsample from Okt/Nov; subsample with known landline number) gives 81 estimates. Since one of the mtmm models did not reach a proper solution for one subsample we had $81 - 9 = 72$ estimates.

Table 3. Secondary analysis of MTMM estimates: effects of interview mode and topic controlled for the effects of time and subpopulation

Factor	N ¹	Validity coefficient		Reliability coefficient	
		Estimated marginal mean	Eta sqd	Estimated marginal mean	Eta sqd
Interview mode			.27*		.28*
Cati	24	.94		.77	
Capi	24	.88		.71	
Cawi	24	.93		.82	
Topic			.84*		.34*
Interest in news and politics	27	.98		.76	
Personality	18	.88		.79	
Life satisfaction	9	.77		.68	
Societal participation	12	.96		.81	
Sport participation	6	.99		.80	
Time between interviews			.03 ^{ns}		.02 ^{ns}
1 to 8 months	54				
Less than 2 months ²	18				
Subpopulation			.03 ^{ns}		.01 ^{ns}
All	45				
Known landline nr ²	18				
Mode * Time			.02 ^{ns}		.02 ^{ns}
Cati 1 to 8 months	18				
Cati < 2 months	6				
Capi 1 to 8 months	18				
Capi < 2 months	6				
Cawi 1 to 8 months	18				
Cawi < 2 months	6				
Mode * Subpopulation			.04 ^{ns}		.02 ^{ns}
Cati All	15				
Cati Known landline nr	9				
Capi All	15				
Capi Known landline nr	9				
Cawi All	15				
Cawi Known landline nr	9				
Grand mean (sd)		.91 (.01)		.77 (.01)	
Adjusted R Squared			.82		.50

* = p < 0.05.

¹This N is the number of estimates obtained from the different mtmm analyses. It is not the number of respondents on which the mtmm estimates were based.

²The first category of this factor is the total sample. The second category is a subselection of the total sample.