Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica
OFS BFS UST

Schweizer Haushalt-Panel
Panel suisse de ménages
Swiss Household Panel

Actualités OFS
BFS aktuell
Attualità UST

0 Statistische Grundlagen und Übersichten
Bases statistiques et vues d'ensemble
Basi statistiche e presentazioni generali

October 2001

Mode effects in panel surveys: A comparison of CAPI and CATI*

Annette Scherpenzeel
in collaboration with Philippe Eichenberger

* Co-financed by the Swiss Federal Statistical Office and the Swiss National Science Foundation (NR. 5004-056052)

Further informations:
Swiss Household Panel: Annette Scherpenzeel
Tel. 032 718 36 00
e-mail: swiss.panel@unine.ch website: http://www.unine.ch/psm

Order number: 448-0100

CH-2010 Neuchâtel
Espace de l'Europe 10

Abstract

In a methodological experiment, carried out within the framework of the Swiss Household Panel, two data collection strategies were evaluated: Computer Assisted Personal Interviews (CAPI) and Computer Assisted Telephone Interviews (CATI). The study was designed as a split-ballot in combination with repeated measurements. The participation rates in the two modes were similar and most answer distributions were the same in both modes. Moreover, it is shown that CATI saves costs and time in comparison with CAPI. Using a Multitrait-Multimethod design, estimates of reliability and validity were obtained for all data. Comparing these estimates, the conclusion must be that the choice of CATI versus CAPI has no implications for the quality of the data. Once the choice has been made, it does not make a very large difference either which type of response scale is used. What can really make a difference are the formulation and the topic of the question.

Introduction

In 1998, the "Swiss Household Panel" was founded, supported by the Swiss National Science Foundation, the University of Neuchatel, and the Swiss Federal Statistical Office. For this large-scale panel study, about 8000 persons in 4000 households will be interviewed each year, to assess changes in their living conditions, attitudes and beliefs. Soon after its start it was decided to use telephone interviewing for the data collection, in contrast with most other European household panel studies (BHPS, SOEP, Eurostat) that use face-to-face interviewing. A methodological experiment was then designed in which two data collection strategies were compared: Computer Assisted Telephone Interviews (CATI) and Computer Assisted Personal Interviews (CAPI). For a review of the advantages and disadvantages of computer assisted data collection we refer to Groves and Nicholls (1986) and Saris (1991).

Extensive literature exists on comparisons between face-to-face, mail and telephone interview techniques. Overviews of such studies are given by Nicholls and Groves, 1986; Groves and Nicholls, 1986; de Leeuw and van der Zouwen, 1988; Groves, 1989; Lyberg and Kasprsyk, 1991. However, the results one finds in this mass of literature are not very consistent. Groves and Kahn (1979), for example, found few significant differences in univariate and bivariate response distributions between telephone and face-to-face interviewing modes. In their meta-analyses of a large number of comparative studies, De Leeuw and Van der Zouwen (1988) and De Leeuw (1992) also report only small differences between telephone and face-to-face interviewing. However, other studies report larger mode effects. Silberstein and Scott (1991), for example, found large mode effects in family expenditure research. A study in the field of time-budget research showed considerable differences in the reports of some of the activities, obtained by different methods of data collection (Kalfs, 1993). Scherpenzeel and Saris (1995) found rather high levels of error variance reflecting an interaction between data collection mode and topic of investigation.

This inconsistency is probably caused by differences in the design of the studies. The design of mode comparison studies can differ in the following ways:

1. *The objective of the study;* Recently, Kaase and Saris (1997) have introduced a useful way to describe the different effects involved in comparative studies of data collection techniques: It can be expected that the total difference between two different data collection modes (T) is equal to the difference due to coverage differences (C) plus the difference due to differences in non-response (N) plus the difference due to the marginal mode of data collection (M):

$$T = C + N + M$$

   The non-response differences result from differences in sample design and the fieldwork by survey organisations, such as call-back rules, refusal conversation rules, centrality of the supervising process, interviewer selection and training procedures. In some studies of data collection techniques, the topic of interest is the marginal effect of the medium of communication (M). In such studies, usually conducted by researchers of communication, a controlled experimental design is used where only the mode of administering the questions is varied and all other factors are kept the same. In a second type of study, usually done by practical survey researchers, the objective is the comparison of „real life" procedures for data collection. Hence, the total difference (T) is studied, given all the differences in the way surveys are conducted in the two modes. Each mode is optimised by using the relevant common fieldwork procedures. Therefore, this type of research is sometimes labelled a „Maximum telephone /Maximum personal" comparison (Groves, 1989).

2. *The experimental design of the study;* Split ballot experiments are used to make comparisons between data collection methods on the basis of identical samples from the same population. This approach is simple and shows whether the method used makes a difference. However, it does not indicate which method or question is the best when differences, in for example answer distributions and means, are found (Saris, 1998). Furthermore, it does not give an estimation of the reliability of the questions used within each method. The split ballot

design is the most common design in mode comparison studies. Some examples are Schuman and Presser (1981); Bishop (1988); Catlin and Ingram (1988); Körmendi (1988); Schwartz et al (1991); Arbeitsgemeinschaft LINK / DemoSCOPE, (1997).
A second type of design is the test-retest design. By repeating the same question to the same people this design can provide an estimate of the reliability of a question (Alwin and Krosnick, 1991; Forsman and Schreiner, 1991). However, it does not give any indication of the validity of the question. Test-retest designs are extremely rare in the field of mode comparisons. A single example is the study of Martin et al. (1993).
The third type of design, the Multitrait-Multimethod design (Campbell and Fiske, 1959), is somewhat related to the test-retest approach. It also consists of repeating the same questions to the same people, but in addition, one aspect of the repeated questions is systematically varied. In this way, random variance can be distinguished from systematic method variance and estimates of both reliability as the complement of random error variance and validity as the complement of systematic method variance are obtained. The Multitrait-Multimethod design is relatively new to the field of survey research, but Andrews (1984); Rodgers, Andrews and Herzog (1992); Költringer (1993); Scherpenzeel and Saris (1997) have done some studies using this design. In the studies of Andrews (1984) and Scherpenzeel and Saris (1995, 1997), the Multitrait-Multimethod design has been used specifically to study the effects of modes of data collection.

3. *The criteria on which the modes are compared*; In many studies of data collection techniques, the costs, speed and response rates of the different procedures are evaluated. In addition, the similarity of the (univariate) answer distributions and means is a commonly used criterion. Elaborate overviews of these types of studies are given by Nicholls and Groves (1986); Groves and Nicholls (1986); de Leeuw and van der Zouwen (1988); Groves (1989); Lyberg and Kasprsyk (1991).
Other studies concentrate on the quality of the data in the sense of the measurement error connected with different modes. Sometimes it is possible, for example, to estimate the accuracy of the data by comparing the results of a survey measurement with official records (Körmandi, 1988). Other criteria for data quality that have been used include, for example, the item-missing-data rates (Groves and Kahn, 1979; Jordan et al. 1980; Groves and Mathiowetz, 1984; Sykes and Hoinville, 1985; Catlin and Ingram, 1988; Körmendi, 1988; de Leeuw, 1992), the detail of answers on open questions (Groves and Kahn, 1979; Catlin and Ingram, 1988; Körmendi, 1988; de Leeuw, 1992), the number of socially desirable answers on a particular question (Körmendi, 1988; Sykes and Collins, 1988; de Leeuw, 1992), the prevalence of response effects (Bishop et al., 1988; Schwartz et al., 1991) and the completeness and number of reported activities (Kalfs, 1993).
Very few studies use estimates of reliability as criteria for data quality, obtained with the test-retest approach described above, or estimates of reliability and validity obtained with the Multitrait-Multimethod approach (some examples of both types of studies are given above).
An additional category of study investigates the interaction of modes of data collection with well-known response effects such as socially desirable answering; the tendency to choose „don't know" or „no opinion" categories; the impact of the tone of wording of questions, etc. (Among others: Schuman and Presser, 1981; Groves, 1989; Schwartz et al., 1991; Scherpenzeel and Saris, 1995).

4. *The use of computer assisted methods;* Most studies mentioned earlier compare traditional modes of interviewing or traditional face-to-face interviewing with CATI. Few studies are available that compare two computer assisted techniques: CATI and CAPI. A few examples are the study of the research literure on the costs and data quality of CATI and CAPI by Snijkers (1992); the study of Kalfs (1993) who compared two different computer assisted procedures in the field of time-budget; and the evaluation of CATI and CAPI for income and expenditure research carried out by the Arbeitsgemeinschaft LINK / DemoSCOPE, (1997). The latter study is of special interest, since it was also carried out in Switzerland.

Differences in the design aspects described make it too difficult to draw conclusions from the literature when one wants to set up a specific survey. The study proposed here aims to be of as much general interest and practical use as possible for survey practitioners. The objective of the present study is the comparison of „real life" data collection procedures, that is: the study of the total effect T. The significance of an experimental study of the marginal effect of the medium of communication would be very limited for survey practitioners. The experimental design we used is a combination of the split-ballot design with the Multitrait-Multimethod approach. The criteria for comparison of the data collection strategies provided by the split-ballot design are the classical ones: costs, speed, response rates, answer distributions and summary statistics. The criteria for comparison provided by the Multitrait-Multimethod design are the reliability and validity of the data obtained.

Experimental design

In the first stage of the test, an initial sample was split into three groups: Two experimental groups and one control group (see table 1). All respondents were first contacted

by telephone. Next, the respondents in the control group and in experimental group 1 were interviewed immediately by telephone, whilst the respondents in experimental group 2 were visited by an interviewer. In both interviews, a small selection of questions was asked twice within the same interview. At the end of the interviews, each respondent was asked to participate in a second interview that would take place one month later.

In the second stage, one month after the first interview, the respondents were re-interviewed. The two experimental groups were crossed over: Those interviewed using CATI the first time were interviewed using CAPI the second time, and vice-versa (table 1). It was explained to these respondents that we ask them the same questions in two different ways because we want to compare two data collection strategies. The control group, which had been interviewed using CATI the first time, were interviewed a second time using CATI again. This control group was included to control for change in opinions over the time interval of one month. In the experimental groups such a change in opinion would be confounded with the change in interview method. As in the first interview, a small selection of questions was asked twice within the second interview to all groups of respondents.

Table 1 Design

	CATI	CAPI
CATI	Control group: Two CATI interviews	Experimental group 2: First CAPI, second CATI
CAPI	Experimental group 1: First CATI, second CAPI	

Sample

The experiment was carried out in the Swiss (German speaking) agglomeration of Bern which is made up of 34 communes and is covered by 81 postal codes. The goal was to have about 200 respondents in each of the three groups. An initial simple random sample of 1452 telephone numbers within the agglomeration of Bern was drawn from a file containing all valid Swisscom telephone numbers. Next, the names and addresses belonging to these telephone numbers were searched, using direct access to the Swisscom's electronic dictionary. For about 102 numbers (7%), no names and addresses were found (not listed). During the fieldwork stage, the non-response and refusal rates appeared to be much higher than expected, and it became clear that the initial sample was not large enough to obtain the desired number of interviews. Hence, to continue the fieldwork, a second complementary simple random sample consisting of 727 telephone numbers and addresses was drawn from the same region, using the same procedure.

Since the study was meant to serve as an experimental test preceding a household panel survey, the sample unit was the household. All members of the household who were 15 years or older were to be interviewed.

Questionnaire

A selection of questions from the normal panel questionnaire was used, spread across a variety of themes, such as health, satisfaction, social networks, income, time budget, and politics.

Some questions that are known to induce response effects were deliberately included. We were interested in the interactive effects of mode with the social desirability bias, with memory effects, with response scale effects and with question formulation. The precise question-formulation experimentation and the specific hypotheses concerning the interaction of response effects and mode effects are described in the "Results" paragraph.

An identical questionnaire and the same programming language for the interview were used for CATI and CAPI. In the CATI condition however, the program was used within a centralised system, while in the CAPI condition it was used in a stand-alone version, using laptops and diskettes.

Fieldwork

The commercial survey institute IPSO in Dübendorf, Switzerland, carried out all fieldwork. This institute has been doing both CATI and CAPI surveys for many years. The fieldwork for this study was done between January 25 and April 13, 1999.

A letter announcing the study was sent to all households in the sample, one week before an interviewer contacted them. All households were first contacted by telephone. After the first telephone contact, the households in the control group and experimental group 1 were interviewed directly or an appointment was made for a telephone interview at a later time. For the households in experimental group 2 the first telephone contact served to make appointments for a visit at home. In both modes of datacollection, all members of a household who were 15 years or older were asked to participate.

One month after the first interview, all households that had participated in the first interview were contacted again by telephone. Just like the first time, the respondents were then interviewed directly or appointments were made for a telephone interview at a later time or for a visit at home, depending on the group they were in.

A different pool of interviewers was used for the CATI and the CAPI interviews. In each experimental condition, only interviewers experienced in using that mode were used. The CATI instruction was given orally and centrally, a few hours before the interviewing began. The CAPI instruction was in written form. It was sent to the interviewers one week before the start of the interviews, together with a training diskette. The data registered on the training diskettes after use allowed the supervisors to control whether the interviewers really carried out the training. For both interviewing modes this was the common method of instruction, to which the interviewers were accustomed. During the fieldwork, the CATI interviewers worked in a central laboratory, under constant supervision. The CAPI interviewers worked independently from their own home address, but could telephone a central supervisor if they had any questions or problems.

The Multitrait-Multimethod model

In our study, we measured each of a number of traits with a number of different methods. This design was introduced by Campbell and Fiske (1959) and is called a „multitrait-multimethod" (MTMM) design. It can be formulated as follows (for more detail we refer to Saris and Andrews, 1991 and Scherpenzeel and Saris, 1997):

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

$$y_i = h_{il} T_l + e_i \quad (1)$$

where h_{il} respresents the strength of the relationship between the stable component, or true score, and the response. The true score can further be decomposed into a component representing the score on the variable of interest F_j, a component due to the method used M_k, and a unique component due to the combination of method and trait u_l. However, following Saris and Andrews (1991) and Scherpenzeel and Saris (1997) , we assume the unique component u_l to be zero. After standardization this leads to the formulation of equation (2)

$$T_l = b_{lj} F_j + g_{lk} M_k \quad (2)$$

where b_{lj} represents the strength of the relationship between the latent variable of interest and the true score and g_{lk} indicates the effect of the method on the true score. All variables are standardized, except for the disturbance variables which are normally not standardized. 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 the method factors are uncorrelated with each other and with the trait factors. If all variables except the disturbance terms are standardized, the coefficients h_{il}, b_{lj} and g_{lk} indicate the strength of the relationships between the variables in the model, and these coefficients have been given a special interpretation:

- h_{il} is called the „reliability coefficient". The square of this coefficient is an estimate of the test-retest reliability in the sense of the classical test theory.
- b_{lj} is called the „true score validity coefficient" because the square of this coefficient is the explained variance in the true score due to the variable of interest.
- g_{lk} is called the „method effect" because the square of this coefficient is the explained variance in the true score due to the method used.

The methods in our study were the two different data collection modes. However, it was suspected that the response scales used in the different interviews would also account for a large part of the correlations between the variables. This was expected on the basis of Scherpenzeel and Saris (1995). We were especially interested in the effects of categorical scales versus 10 point scales. An hypothesis derived from information theory by Alwin (1997) is, that rating scales with more response categories transmit a greater amount of information and are therefore inherently more precise in their measurement. Therefore, we can expect that using more response categories ensures greater reliability of measurement independent of interviewing mode. To study these effects,

Table 2 Balanced incomplete block design for three traits, two modes of data collection and two response scales

		Trait 1	Trait 2	Trait 3
CATI	Scale 1	Item 1	Item 2	
	Scale 2	Item 3		Item 4
CAPI	Scale 1	Item 5	Item 6	
	Scale 2	Item 7		Item 8

the same questions have to be asked with different response scales, in addition to the repetition of the same questions within each data collection mode. This would mean that each question about a certain trait would have to be asked four times in total: twice within each interview, with two different response scales. This design increases the costs of data collection and could lead to irritation of the respondents and memory effects in short questionnaires. Therefore, we choose to reduce the number of repeated questions by using a balanced incomplete block design.

In each of the interviews, the question about trait 2 is asked only once, using response scale 1. The question about trait 3 is also asked only once in each interview using response scale 2. Only the question about trait 1 is repeated within each interview, using both response scales. In this design, each method (mode of data collection or response scale) occurs with equal frequency. In addition, each combination of methods (mode of data collection and response scale) occurs with equal frequency. Although not all possible combinations of methods have been used for each trait, the design allows to test 4 different combinations.

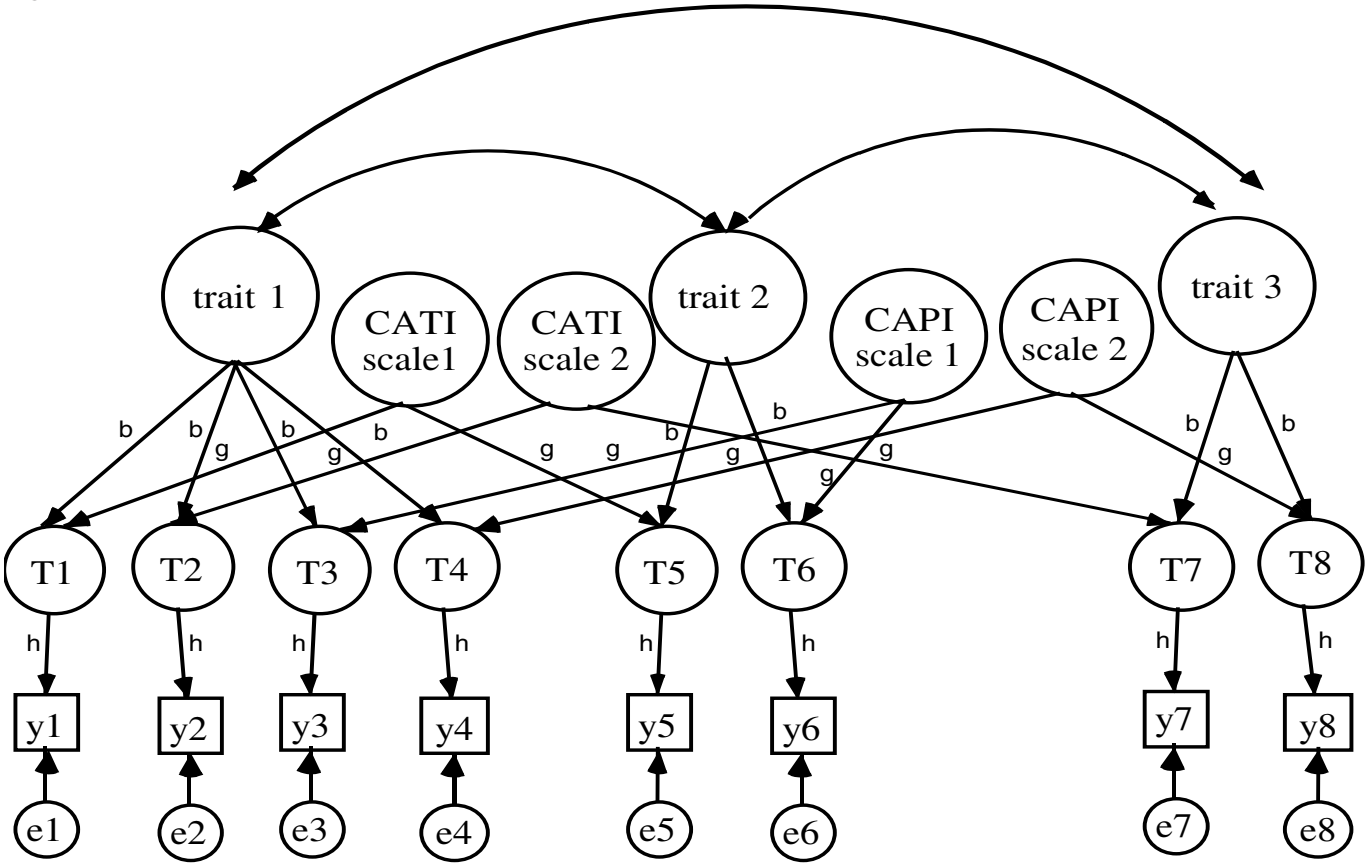
The two types of method factors, those representing the data collection modes and those representing the response scales, can be combined in different ways in the MTMM model. The first way is to specify separate factors for each and to assume additivity of the effects. The second possibility is to specify factors for each combination of data collection mode and response scale and thus to assume interaction effects[1]. Since we expected the response scales to have differential effects within each data collection mode, we chose the model in figure 1. This model was estimated five times, each time with three different traits.

Results

1. Costs and speed

The mean duration of the CATI and CAPI interviews was equal (22 minutes), but the CAPI fieldwork took notably longer than the CATI fieldwork. Whilst it took 6 weeks of interviewing to complete all interviews in CATI, it took 10 weeks of interviewing in CAPI. The telephone screening stage preceding the interviews took 3 weeks in both conditions. The total costs per interview were Sfr 47.– in

Figure 1 The model



[1] In fact, a third possibility would be to include both additive and interactive factors. This, however, constituted a clearly overfactored model: All estimations of this model resulted in improper solutions.

[2] Amounts as given by the survey institute IPSO, personal communication 22-6-1999.

CATI against Sfr 86.– in CAPI[2]. For CATI, these costs include interviewer-payment. For CAPI, the costs include interviewer-payment, travelling costs and field support.

Hence, it can be concluded that CATI saves costs and time in comparison with CAPI.

2. Response

As described in the paragraph about the fieldwork, a second sample was drawn to complete the number of interviews required, and the fieldwork for this secondary sample was somewhat changed. Due to this change in procedure, the contact and response rates for the two samples cannot be compared. We therefore present the rates for the first sample only, which were obtained with the original, complete procedure.

In CATI 6% of all contacted households were not reached, in CAPI this was 7% (table 3). The most frequent causes of not reaching a household were that there was no answer or an answering machine.

The total participation rate of households in the first wave was 31%, that is: in 31 % of the contacted households at least 1 person participated (table 4). Within these households, on average 86% of the contacted individuals participated (table 6). Household participation was slightly different for CATI and CAPI. The household participation rate was about 6 % higher and the refusal rate about 5 % lower in CATI compared to CAPI (table 4). In both conditions, about 10% of the households could not participate because of language problems, age, health or a prolonged stay abroad of the person who answered the first telephone call (table 4).

The general participation rate is very low in comparison with other surveys that have been carried out in Switzerland. This could perhaps have something to do with the request to interview *all* members of the household for our panel study. In normal cross-sectional surveys, a single individual is interviewed per household. It should be noted, however, that only 4 (0.5 %) of those who refused to be interviewed spontaneously mentioned the fact that the request for *all* household members to be interviewed was the reason for their refusal. In addition, we will see in table 9 that single person households were not more likely to participate, as would have been expected if the request to interview all members of the household were a cause of refusal.

The refusal rate in the experiment cannot be increased by the fact that the respondents were asked to participate in two waves of interviewing, because this was not asked at the beginning. Only at the end of the interview were they asked to participate a second time.

Table 3 Household contact rates

	CATI		CAPI	
Not reached	6%		7%	
No Answer		6%		5%
Other problems[1]		0%		2%
Reached households	94%		93%	
Total number of addresses (100%)	674		676	

$\chi^2 = 0.787$, df = 1, p = 0.375.

[1] The category “Other problems” is a summary of the categories: “No telephone connection”; “No private household”; “Modem or fax”; and “Answering machine”.

Table 4 Household refusal rates

	CATI		CAPI	
Participating households[1]	33%		27%	
Refusing households	56%		61%	
Complete refusal[2]		46%		46%
Participation screening[3]		10%		15%
Problems (language, old age, illness)	11%		12%	
Reached households (100%)	636		630	

$\chi^2 = 5.826$, df = 2, p = 0.054.

[1] At least one individual interview was completed in the household.
[2] The first contacted person refused, on behalf of the household, both the screening interview and the individual interview(s).
[3] The first contacted person completed the screening interview but no further individual interviews were completed in the household.

Table 5 Relationship announcing letter and household response rates

First contacted person:	did see letter	did not see letter	Total[1]
Household participated[2]	53%	31%	374
Household refused	47%	69%	422
Total (100 %)	571	225	796

$\chi^2 = 31.730$, df = 1, $p < 0.05$.

[1] About half of the refusing households were willing to answer a few questions, including this question about the letter. As a result, only 422 refusing households are included in this table instead of 747 as would be expected on the basis of table 4.
[2] At least one individual interview was completed in the household.

Table 6 Individual refusal rates within participating households

	CATI	CAPI
Completed interviews	83%	89%
Refusing individuals	12%	11%
Problems (language, old age, illness, abroad)	5%	1%
Total number of contacted individuals[1] (100%)	387	313

$\chi^2 = 9.439$, df = 2, $p < 0.01$.

[1] Within participating households, all household members older than 15 who were not absent for a long period were contacted.

This finding suggests that in panel studies using the household as a sample unit, a refusal rate well above the commonly obtained refusal rates can be expected. The drawing of a larger initial sample and the development of special strategies for refusal conversion might be worthwhile for such studies.

In a debriefing during the fieldwork stage, interviewers and interviewer staff expressed problems in explaining to respondents the exact nature and purpose of the survey. They felt the vagueness of the subject of the study made it difficult for the interviewers to motivate people to participate. Although this is a subjective impression of the interviewers, for which no statistical support is available, in future it might be advisable to describe the contents and purpose of socio-economical panel studies in clearer, concrete, and more understandable terms.

The total percentage of refusing households (table 4) is split up into households that refused both the screening and the individual interviews and households that completed the screening but no individual interviews. In total 12 % of the reached households dropped out only after the screening interview. Hence, the total refusal rate could perhaps be lowered if a special effort is made to motivate people during or at the end of a screening, and by interviewing people immediately thereafter thus avoiding the necessity of a subsequent appointment.

A letter announcing the study had been sent to all households one week before the telephone contact. The first contacted person in each household was asked whether he/she had seen this letter. In table 5, we see that this letter had a very large effect on the participation rate: Seeing the letter increased participation by about 20%.

The participation of individuals within participating households is 6% higher in CAPI than in CATI (table 6), in contrast with the household participation we saw in table 4. This is mainly caused by a higher number of problems in CATI, the refusal rates being almost equal in both modes. This difference in distributions is significant.

The most often mentioned reasons for refusal are “no interest” (CATI: 43%, CAPI: 39% of the refusals) and “no time” (CATI: 17%, CAPI: 18% of the refusals), as is commonly found. Only a very small minority of people (CATI: 2%, CAPI: 1% of the refusals) mention the data collection technique as a reason for refusal.

The gender, age and nationality distributions of the CATI participants and the CAPI participants were similar and did not differ between participants or refusals[3]. The size of the households ranged from 1 to 10 persons and the distributions were the same for the participating and the refusing households. Contrary to our expectations, single person households were not more likely to participate than households with two or more members, nor were large households (6 or more persons) more likely to refuse. This was found in both modes.

[3] The distributions of the refusals included only those people who refused to do the interview but were nevertheless willing to answer a few demographic questions, which is between 30 and 40% of all refusals. This is also true for the distributions of household size.

3. Answer distributions

Differences in response rates are of practical interest only if they actually influence the responses to the questions in the interview. Therefore, we also compared the answer distributions of a selection of questions. This selection included questions that are known to induce socially desirable answering and questions that appeal to memory.

Socially desirable answering

The presence of an interviewer to whom one is required to tell personal or sensitive matters can stimulate socially desirable answers. It is not clear whether the lack of visual contact causes the telephone interviewer to have a higher degree of anonymity than the face-to-face interviewer. In their meta-analysis of the literature about mode effects, de Leeuw and van der Zouwen (1988) found slightly more desirability effects and under-reporting on sensitive topics in telephone interviewing than in face-to-face interviewing. However, more recent studies not yet included in this meta-analysis report no differences on sensitive items between the different modes (Körmendi, 1988; Sykes and Collins, 1988).

In the experiment, questions were included about alcohol consumption and problems of addiction, satisfaction with life, subjective health, problems with social contacts, help from one's partner and television watching, to test the existence of an interaction between mode of interviewing and sensitivity of a question. We will now shortly describe each sensitive question, its underlying hypothesis and the results obtained.

One of the most widely used examples of a question sensitive to socially desirable answering is how often one drinks alcohol. This question was therefore included in the experiment, as an open-ended question coded by the interviewers (codes ranged from „never" to „3 or more times per day"). The answer distributions obtained with this question in CATI and CAPI are the same (table 7). A somewhat related question asked respondents to indicate, by simply saying yes or no, whether they had had any problems with alcohol or other addictions in the past two years. As for the question about alcohol consumption, it was expected that the less anonymous CAPI interview would induce under-reporting. This was not found (table 7). Only a few people report having these problems (2% in both modes) and no significant differences are found in the very skewed answer distributions obtained with CATI and CAPI.

There are good reasons to expect a social desirability bias in the subjective evaluation of one's health. It seems to be a social norm to appear happy and healthy. Saying that one is not feeling well is generally interpreted as a call for attention or help (Veenhoven, 1984). An open-ended question was used to ask people about their health in general, which was then coded into 5 categories, from „very good" to „very bad", by the interviewers. Respondents in the CAPI condition more often described their health as „very good" than respondents in the CATI condition (table 7), confirming the hypothesis that a more anonymous interview induces less positive evaluations of subjective health. However, the difference between the modes of interviewing disappears when the categories „very good" and „good" are collapsed, as is commonly done in national health studies.

Within the same domain of subjective well being, respondents were also asked whether they had any problems with social contacts or with loneliness in the past two years. It was expected again that people would try to appear happier, and thus report fewer problems, in the less anonymous CAPI interview. Again, however, the hypothesis was rejected: the very small percentage of people that report having these problems are the same in both interviewing conditions (table 7).

Another question asked how much help their partners could provide when necessary. This was done to see whether people give a more positive impression of their partner when he or she is present during the interview. In CATI, the respondent can always answer freely because the partner does not usually hear the question. In CAPI, however, the presence of the partner could make a difference. Therefore, we let the CAPI interviewers indicate the presence of the partner during the CAPI interviews. Table 8 shows that the hypothesis is not confirmed: No significant difference is found between the distributions of help from a partner reported in CATI, in CAPI with the partner present during the interview, and in CAPI with the partner absent during the interview.

Some studies indicate that the more anonymous the interview is, the less happiness and satisfaction are reported (Sudman, 1967; Groves, 1989). Therefore, respondents in both conditions were asked to indicate their satisfaction with life in general and with some aspects of their lives by a number between 0 and 10. The results in CAPI, where an interviewer is physically present, were expected to differ from the results in CATI. However, the mean satisfaction scores obtained with CATI and with CAPI do not differ significantly (table 9).

Finally, watching television is sometimes considered socially undesirable behaviour. Kalfs (1993) found that people report less television watching in CATI compared with Computer Assisted Self-interviews (CASI) and with a paper and pencil diary. This difference was especially large among the highly educated. In the present

experiment, the amount of time spent on television watching is equal in both interviewing modes (the mean scores on this question are presented in table 23, together with other time expenditure questions which will be discussed later). In addition, no differences are found when only people with a higher education level are selected.

Summarised, nearly no differences are found between CATI and CAPI in the answers to sensitive questions. No interaction of mode of interviewing and sensitivity to socially desirable answering could be demonstrated. It is possible that the questions chosen are not as sensitive to the social desirability bias as we thought. It is also

Table 7 Answer distributions of four questions sensitive to social desirable answering

	CATI	CAPI	χ^2, df
	How often do you drink alcoholic drinks?		
3x or more a day	1%	1%	
2x a day	2%	1%	
1x a day	11%	12%	
Several x a week	12%	11%	
1 to 2 x a week	27%	26%	
More seldom	36%	38%	
Never	10%	10%	
Don't know / No answer	0%	0%	
Total	451	413	χ^2 = 3.153, df = 7
	Did you have alcohol-problems or problems of addiction?		
Yes	2%	2%	
No	98%	97%	
No answer	0%	0%	
Total	451	413	χ^2 = 1.543, df = 2
	How is your health presently?		
Very good	23%	33%	
Good	60%	51%	
Average	15%	14%	
Bad	1%	1%	
Very bad	0%	0%	
Don't know / No answer	0%	0%	
Total (100%)	451	413	χ^2 = 11.969, df = 5*
	Did you have difficulties with contacts or problems of loneliness?		
Yes	8%	8%	
No	92%	91%	
No answer	0%	0%	
Total	451	413	χ^2 = 0.152, df = 2

* = p < 0.05. ** = p < 0.01.

Table 8 Answer distributions for help from partner

	CATI	CAPI		χ^2, df
		Partner present	Partner not present	
	How much practical help could your partner give you?			
A lot	74%	73%	70%	
Moderate amount	17%	18%	22%	
A little bit	7%	5%	5%	
Nothing / Don't want help	2%	4%	3%	
Don't know / No answer	0%	0%	0%	
Total	320	104	194	χ^2 = 7.962, df = 10

* = p < 0.05. ** = p < 0.01.

possible, however, that CATI is not experienced as more anonymous than CAPI. This means an interviewer with whom one has only verbal contact has the same impact as an interviewer with whom one has verbal *and* visual contact.

Memory effects

The two modes of interviewing differ in the 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 (Groves, 1989; Schwartz et al, 1991). As a result, respondents will not take a long time to think about their answer or to search their memory. For questions about facts, events and past behaviour this can lead to underreporting when events are forgotten, but also to overreporting when events are displaced in time ("forward telescoping": Sudman en Bradburn, 1974; Groves, 1989).

The available time for a CAPI interview is restricted to some degree, because most people will not want an interviewer in their home for hours. Nevertheless, we hypothesize that the time available for a CAPI interview is probably longer than in a telephone conversation, because the social rule is that normal telephone conversations are shorter than face-to-face conversations (Dillman, 1978). Therefore, questions that largely rely on memory may be answered more reliably in CAPI than in CATI.

To test this hypothesis, questions were asked about the number of visits to a physician in the past year, time spent on different daily activities, and behaviour on different voting occasions during the past year. A short description now follows of all questions that rely on memory, their underlying hypothesis and the results obtained.

It was expected that the number of visits to a physician in the past year would be under-reported in CATI in comparison with CAPI, because people would take less time to count the exact number. This expectation is rejected: No significant difference is found in the mean number of visits reported in CATI versus CAPI (table 10). In CATI the answers are somewhat more spread than in CAPI (standard deviation 9.00 versus 6.30, range 120 versus 60). It is possible we found no differences because the forgotten visits in CATI outweighed the visits that were falsely placed in the past year („forward telescoping").

Table 9 Mean score on five questions sensitive to social desirable answering

	CATI Mean score (SD)	CAPI Mean score (SD)	2-tailed t
	How satisfied are you with . . . Give a number between 0 and 10, if 0 means not at all satisfied and 10 means completely satisfied		
with life in general	7.95 (1.65)	8.11 (1.49)	t = 1.495
with health	8.07 (1.76)	8.09 (1.75)	t = 0.214
with social contacts	8.21 (1.72)	8.23 (1.63)	t = 0.139
with financial situation	7.40 (2.23)	7.43 (2.16)	t = 0.161

* = p < 0.05. ** = p < 0.01.

[1] The time spent on Saturdays and Sundays was asked separately. However, since no differences were found here either, the table presentation is limited to ordinary weekdays.

Table 10 Number of visits to a physician and time expenditure

	CATI Mean score (SD)	CAPI Mean score (SD)	2-tailed t
	In the past year, how many times have you visited a physician		
open answer	3.90 (9.00)	3.92 (6.30)	t = 0.042
	How many hours on an ordinary weekday[1] do you spent on . . .		
Housekeeping	2.02 (2.09)	2.03 (2.28)	t = 0.059
Hobbies and leisure activities	1.86 (1.70)	1.85 (1.64)	t = 0.019
Television watching, all respondents	1.50 (1.29)	1.61 (1.43)	t = 1.162
Television watching, higher educated[2]	1.47 (1.51)	1.46 (1.92)	t = 0.034

* = p < 0.05. ** = p < 0.01.

[1] The time spent on Saturdays and Sundays was asked separately for each activity. Since no differences between CATI and CAPI were found here either, the table presentation is limited to ordinary weekdays.
[2] University, college

For the second test of memory effects, respondents were asked to indicate the number of hours they normally spend on each of three activities: housekeeping, television watching and hobbies. They were asked to indicate separately the number of hours spent on weekdays, Saturdays and Sundays. It was expected that time pressure in the CATI interview would lead to less precise estimates of time expenditure. However, the results in table 10 show that the amount of time spent on each activity is equal in both interviewing modes, for ordinary weekdays and for Saturdays and Sundays.

A third test of memory effects was done within the field of voting behaviour. Questions were asked about three referenda that took place during the past year: The „Gene-protection initiative" referendum, the „Truck-traffic tax" referendum, and the referendum on „Construction and financing of public transport plans". Respondents were asked whether they had voted or had not (voting participation) and, if so, whether they had voted for, against, or had abstained. It was hypothesised that the reports obtained with CAPI would resemble the official statistics more than the reports obtained with CATI, because memory effects are stronger in CATI.

The percentage of people who said they had voted does not differ between CATI and CAPI (table 11). However, it is notably different from the real percentage of people that voted. This difference between reported and real voting participation has often been found. A common explanation is that survey samples contain an over-representation of people who are interested in politics and/or are politically active. Another explanation states that the difference represents a social desirability bias, voting participation being seen as the norm. Our hypothesis about memory effects causing differences in reported voting participation between CATI and CAPI is rejected, but it could be that memory failure is equally strong in both modes of interviewing. This would add a third explanation of the difference between reported and real voting participation.

In table 12 we find the voting behaviour of those that said they had voted. For two of the three referenda, the expected differences between CATI and CAPI were found. In the CAPI interview, 8% fewer respondents report having voted *for* the „Gene-protection initiative" referendum than in the CATI interview and 14% more respondents voted against. In addition, more respondents in CATI than in CAPI could not answer the question („don't' know" or „no answer"). On this question, the voting distributions obtained with CAPI resemble the population figures.

For the „Truck-traffic tax" referendum, no significant differences were found between CATI and CAPI, but the answer distributions differed greatly from the population figures.

With regard to the referendum on „Construction and financing of public transport plans", voting in favour was not notably lower in CAPI than in CATI. In contrast, voting against was reported 8% more often in the CAPI interview. This apparent contradiction can be explained by the relatively high percentage of respondents that chose the „don't know" option in the CATI condition (16%, versus 7% in CATI). In this case, the percentage of reported voting *in favour* is quite close to the population figure in both interviewing conditions, but the percentages of reported and real voting *against* the referendum differ considerably.

In conclusion, the hypothesis of stronger memory effects in CATI than in CAPI seems at least partly confirmed with regard to voting behaviour. An alternative explanation with regard to the „Gene-protection initiative" referendum could be the existence of a social desirability bias. This topic was very polarised at the time, those supporting the initiative often being depicted as fundamentalists. In this sense, it could have been more difficult to admit one voted in favour in CAPI, when face-to-face with the interviewer. However, this would have caused a greater bias in CAPI than in CATI, while we found exactly the opposite. In addition, the memory effect hypothesis is supported by the high percentages of „don't know" answers in CATI in respect of the "Construction and financing of public transport plans" referendum. It seems acceptable that people had forgotten how they voted on this complicated topic, which was not covered much by the media. As expected, memory failure is more

Table 11 Voting participation: percentage of people who say they voted on three referenda

	Population[1]	CATI	CAPI
«Gene-protection initiative», 7.6.98	39%	67%	69%
«Truck-traffic tax», 27.9.98	54%	74%	70%
«Construction and financing of public transport plans», 29.11.98	38%	51%	56%
Total[2]		417	382

[1] The population percentages represent the real voting participation in the canton of Bern, which is not completely equal to the agglomeration of Bern as it was defined for our sample.
[2] Only persons entitled to vote

Table 12 Distributions of real and reported voting behaviour on three referenda

	Population[1]	CATI	CAPI	χ^2 cati/capi, df
	How did you vote on the «Gene-protection initiative»?			
Voted for	42%	48%	40%	
Voted against	57%	39%	53%	
Abstained	1%	3%	2%	
Don't know		9%	5%	
No answer		2%	0%	
Total[2]		279	262	χ^2 = 12.050, df = 4*
	How did you vote on the «Truck-traffic tax»?			
Voted for	57%	75%	73%	
Voted against	42%	20%	21%	
Abstained	1%	1%	3%	
Don't know		3%	4%	
No answer		1%	0%	
Total[2]		306	267	χ^2 = 2.934, df = 4
	How did you vote on the «Construction and financing of public transport plans» ?			
Voted for	66%	69%	71%	
Voted against	33%	9%	17%	
Abstained	1%	4%	5%	
Don't know		16%	7%	
No answer		1%	0%	
Total[2]		214	215	χ^2 = 13.853, df = 4**

* = $p < 0.05$. ** = $p < 0.01$.

[1] The population percentages represent the voting behaviour in the canton of Bern, which is not exactly equal to the agglomeration of Bern as it was defined for our sample.
[2] Only persons that said they had voted on this referendum.

pronounced in CATI than in CAPI, where time pressure is less and the interviewer can more easily help to remember and explain the difficult topic.

4. Validity and reliability

Estimation of the multitrait-multimethod models

Some estimation problems were encountered when the five multitrait-multimethod models were estimated. Improper solutions with negative error variances and/or with negative method variance were found for most models. In addition, the model often showed a bad fit. Inspecting the MTMM correlation matrices, we found some very high correlation's between repetitions of the same questions within the same interview. Although these repeated questions had different response scales, we found correlation's of, for example, .860. These correlation's suggested that the respondents had remembered their previous answers, despite the different response scale. The time between the repetitions of the same questions within the same interview had indeed been somewhat short (it was, on average 20 minutes). In the model, the measures correlating so highly were specified to load on different method factors, and this could have caused the estimation problems. The solution we choose was to model these memory effects by introducing correlated errors between measures of the same trait *within one interview.* This modification of the model solved most estimation problems and significantly improved the fit. For one model[4], we continued to find very small, negative variances of one or more method factors. Since these variances were very small and non-significant, we constrained them to be zero. This means the method factor in question is dropped from the model, and corresponding validity coefficients are assumed to be 1. In fact, in each of these cases, we compared the chi-squares of the model with and the model without the method factor[5] and found that the method factor in question did not contribute significantly to the fit of the model. This is consistent with the finding that improper solutions in structural equation models, and especially in MTMM models, are often due to overfactoring (Rindskopf, 1984).

[4] The model with as traits the agree-disagree statements, dealing with the topic „Control in life".
[5] The difference in fit between two nested models may be tested statistically by comparing the chi-square of a model with the chi-square of a model one step earlier in the hierarchy (Widaman, 1985).

Time and order effects

The respondents in the control group were interviewed twice with the same interviewing method (CATI). In this group, differences in the answers of the first and the second wave could only be caused by a change in opinion over time or by random error. To test for the presence of a time effect, the following series of nested models was specified: (0) model with trait factors only, (1a) model with trait factors and a time factor, (1b) model with trait factors and method factors representing the response scales used, (2) model with trait factors, a time factor and method factors representing the response scales used. All model estimations were done with the AMOS program for linear structural equations (Arbuckle, 1997). The tests showed that in most cases, the best model was a model with trait factors and response scale effects, but without time effects. Hence, we can conclude that there was no significant change over time in opinions or attitudes in the experiment. Any nonrandom differences in the answers of first and the second wave in the experimental groups must be caused by the change in interview method.

As was described in the paragraph "Experimental design", the order of the two different interview methods was varied over two experimental groups. Respondents in group 1 were interviewed first by telephone and secondly by face-to-face interview, respondents in group 2 vice-versa (table 1). To test whether the order of the interviews had a significant effect on the parameter estimates, we carried out multi-group analyses with equality constraints on the parameters across the groups. The models with equal parameters in the two experimental groups were rejected, which means that there is an effect due to which interview technique was used first. We assume, for the rest of the analyses, that this order effect is counterbalanced by adding together the two experimental groups.

Meta-analysis

For each question in the experiment, an estimate of validity, reliability, and method effect (consisting of the interactive effects of interview mode and response scale) was obtained. This resulted in 42 validity coefficients, 42 reliability coefficients and 42 method effect coefficients in total. Some effects can be inferred by just looking at these coefficients, but as Scherpenzeel and Saris (1997) have shown, this is not a very systematic way of assessing the effects. Therefore, we carried out a secondary „meta"-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 (SPSS for Windows, 1999).

The dependent variables in the meta-analysis are the validity and the reliability estimates obtained in the

Table 13 Effects of interview mode, response scale and topic on validity and reliability

		Validity coefficient			Reliability coefficient	
	Mean (sd)		.96 (.05)		.76 (.14)	
Factor		**N**	**Estimat. marginal mean**	**Partial Eta sqr**	**Estimat. marginal mean**	**Partial Eta sqr**
Mode	CATI	21	.96		.75	
	CAPI	21	.96	.00	.77	.01
Scale	10 point	16	.98		.75	
	Categories (4 or 5)	22	.95		.75	
	Frequency (number)	4	.95	.14	.80	.01
Topic	Control in life	6	.95		.59	
	Mood	2	.88		.74	
	Politics	6	.98		.81	
	Health	6	1		.80	
	Alcohol use	4	1		.84	
	Social network	6	.97		.77	
	Satisfaction	12	.92	.41**	.80	.31
R^2				.45		.35

* = p < 0.05. ** = p < 0.01.

[1] Standard errors for the marginal means of the validity coefficient range from .01 (when N is larger than about 20) to .03 (when N is smaller than about 5). Standard errors for the marginal means of the reliability coefficient range from .03 to .10.

MTMM-analyses[6]. The mode of interviewing, the response scale and the topic are entered in the analysis as additive factors (table 13). 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 R squared in the last row of the table indicates the amount of variance explained by all factors together.

In this analysis, the mode of interviewing has no effect at all. The response scale also has very small effects only. The 10 point scale tends to cause a very slight decrease in validity compared to the other two scales. For reliability, the real number frequency scale seems to be a little bit better than the other two scales. Both trends are, however, rather weak and not significant. Hence, our a-priori hypothesis, that rating scales with more response categories ensure greater reliability of measurement, is not supported. The topic of the question has the largest effects on the quality of the data. However, the effect of the topic is not very interesting as it is. We can see that a question asking about one's mood has the lowest mean validity of all questions in the study, questions about health and alcohol use the highest[7]. Reliability is lowest for questions about the control in life and highest for questions about alcohol use. We can hardly advise researchers to ask only questions about alcohol use to obtain the highest data quality possible. It is much more interesting to know why one topic evokes more valid and/or more reliable answers than another. In the next analysis, we have tried to untangle the effect of topic by introducing other, more explanatory dimensions.

We have characterised all the questions included in the study on the following dimensions: (1) Sensitivity: whether or not the topic is likely to evoke social desirable answers[8]. (2) Type of information: whether the respondents is asked to indicate a frequency (frequency of having headaches, for example) or an intensity (degree of satisfaction, for example). (3) Formulation: whether a direct, "forced-choice" question form is used (e.g. „Do you feel you have control over the things that happen in your life?) or whether the respondent is asked how much he or she agrees or disagrees with a given statement (e.g. „I have little control on the things that happen in my life").

We expected the interviewing mode to interact with these three question characteristics: Firstly, in the paragraph „answer distributions" we described that less socially desirable answering might be found in CATI, because the telephone interviewer is assumed to have a higher degree of anonymity than the face-to-face interviewer. In the multitrait-multimethod analyses, we expect that a smaller social desirability bias in CATI will result in higher data quality. Secondly, we expext questions about the frequency of past events and behaviour to be answered more reliably in CAPI than in CATI, because of the higher time pressure in CATI (see also the paragraph „answer distributions"). Thirdly, the „agreeing-response bias" refers to a presumed tendency for respondents to agree with attitude statements presented to them (Schuman and Presser, 1981). One of the theoretical explanations of this agreeing-response bias relates it to the interviewer-respondent interaction. The inclination of some respondents to (uncritically) agree with any statement read by the interviewer is interpreted as a form of social deference (Schuman and Presser, 1981) or, as we prefer to see it, as a form of socially desirable answering. Taken as a form of socially desirable answering, the bias is supposed to be larger in CAPI than in CATI. Since the bias generally increases correlated error, we expected to find a larger difference between the validity coefficients of the agree-disagree statements and the direct questions in CAPI.

To test these hypotheses, we included the additive effect of each of these factors in the ANOVA model as well as the interaction effect of each factor with the mode of interviewing (table 14).

The effects of the factors interviewing mode and response scale in table 14 are again small and non-significant. The sensitivity of the topic has a significant effect on validity but no effect on reliability. Validity is a little bit lower for sensitive topics. The type of information that is asked for has a relatively strong influence on the validity[9]. The validity coefficients of questions about the frequency of certain events, activities or behaviour are on average .05 higher than the validity coefficients of questions about the intensity of feelings, opinions or attitudes. The reliability is not influenced by this factor. The

[6] The method effect estimates (consisting of the interactive effect of mode and scale) are not presented as dependent variables here, since they are the complement of the validity coefficients and show exactly the same variation.

[7] As described in the paragraph „estimation of the multitrait-multimethod models" the non-significant negative variance of one method factor in one of the models was constrained to be zero. The corresponding validity coefficients are then assumed to be 1, which explains why some of the marginal means in this table are 1.

[8] Rated by the research staff members.

[9] The factor „type of information" is to some degree collinear with the response scale factor, due to the frequency number scale. The factors are not perfectly collinear however, because some frequencies were asked also with a category scale (ranging, for example, from „very often" to „never").

formulation of the question, on the contrary, has a rather strong effect on the reliability and no effect on the validity. The reliability is much higher for the direct question form than for the form in which a statement has to be judged in agree/disagree form. However, it has to be noted here that there is a perfect overlap between the statement formulation in this analysis and the topic "control in life" in the previous analysis (table 13): The statement form was used exclusively for this topic. Consequently, we do not know whether the negative effect on reliability is an effect of this specific topic or of the formulation of the question. We can conclude that the most important of the three new dimensions for validity is the type of information that is asked for and the most important for reliability could be the formulation of the question.

The interactions of the factors Mode and Scale, Mode and Sensitivity, and Mode and Information have no effects on the data quality (table 14). The eta-squared statistics

Table 14 Effects of interview mode, response scale, topic sensitivity, type of information and formulation on validity and reliability

			Validity coefficient			Reliability coefficient	
Mean (sd)				.96 (.05)		.76 (.14)	
Factor			**N**	**Estimat. marginal mean[1]**	**Partial Eta sqr**	**Estimat. marginal mean[1]**	**Partial Eta sqr**
Mode	CATI		21	.99		.69	
	CAPI		21	.95	.11	.72	.00
Scale	10 point		16	.98		.68	
	Categories (4 or 5)		22	.96		.69	
	Frequency (number)		4	.96	.11	.74	.02
Sensitivity	Non-sensitive topic		24	.99		.70	
	Sensitive topic		18	.95	.23**	.71	.00
Information	Intensity		32	.94		.70	
	Frequency		10	.99	.24**	.71	.00
Formulation	Direct question		36	.98		.80	
	Agree with statement		6	.96	.02	.61	.24**
Mode x Scale	CATI	10 point	8	1		.67	
		Categories	11	.98		.69	
		Frequency	2	.99		.73	
	CAPI	10 point	8	.96		.70	
		Categories	11	.94		.70	
		Frequency	2	.94	.00	.76	.00
Mode x Sens.	CATI	Non-sens.	12	1		.68	
		Sensitive	9	.97		.71	
	CAPI	Non-sens.	12	.97		.72	
		Sensitive	9	.93	.00	.72	.00
Mode x Infor.	CATI	Intensity	6	.96		.68	
		Frequency	5	1		.71	
	CAPI	Intensity	16	.92		.72	
		Frequency	5	.97	.01	.72	.00
Mode x Form.	CATI	Question	18	.97		.80	
		Statement	6	1		.59	
	CAPI	Question	18	.98		.81	
		Statement	6	.91	.23**	.63	.00
R^2					.54		.34

* = p < 0.05. ** = p < 0.01.

[1] Standard errors for the marginal means of the validity coefficient range from .01 (when N is larger than about 20) to .03 (when N is smaller than about 5). Standard errors for the marginal means of the reliability coefficient range from .03 to .11.

show that the interactions do not better explain the variation in validity and reliability coefficients than the separate, additive factors. Hence, response scale effects are no more pronounced in CATI than in CAPI, face-to-face contact of the respondent and the interviewer does not induce more social desirable answering than telephone contact, and the higher time pressure in CATI does not result in more memory effects.

The only significant interaction effect is the effect of mode and the formulation of the question on validity.

The interaction factor has a considerably stronger effect on validity than the two separate, additive factors. It shows that in CAPI, the direct question formulation is clearly better than the agree-disagree form. In CATI there is less difference than in CAPI, and in the opposite direction. This finding confirms our hypothesis[10].

Conclusions

The general participation rate is very low in this study, which is possibly related to the request to interview *all* members of the household. It should be noted, however, that single person households are not more likely to participate than households with two or more members. The household participation rate is slightly higher in CATI compared to CAPI, which is in agreement with an earlier comparative study of CATI and CAPI in Switzerland (Arbeitsgemeinschaft LINK / DemoSCOPE, 1997). Problems of language, old age or illness in CATI lead more often to individual non-participation than in CAPI. On the other hand, the fieldwork duration of CAPI is notably longer, and the costs per interview are much higher.

It is not sufficient to consider only differences in response rates and other technical data (such as costs and speed) between CATI and CAPI . Such differences are of practical interest only if they actually influence the substantive data obtained, that is: the responses to the questions in the interview. Therefore, we also compared the answer distributions of a selection of questions and the quality of the data obtained.

Nearly no differences were found between CATI and CAPI in the answer distributions for sensitive topics. In contrast, reports of behaviour on different voting occasions during the past year are different in CATI as compared to CAPI. The high percentages of „don't know" answers in CATI in particular support the hypothesis that memory fails more in CATI than in CAPI, where time pressure is less and the interviewer can more easily offer memory prompts and explain difficult topics.

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 CATI versus CAPI has no implications for the data quality, defined as validity and reliability. The type of response scale used also makes very little difference. What can have an impact, however, for the quality of the data are the topic and the formulation of the question. Sensitive topics negatively influence the validity of the data in both modes. In addition, the validity of questions about the frequency of certain events, activities or behaviour is clearly higher than the validity of questions about the intensity of feelings, opinions or attitudes. This finding is confirmed by earlier MTMM studies: Rodgers et al. (1992) and Költringer (1993) also found higher data quality for frequency questions than for attitudinal questions. The best question formulation in both modes is a direct, "forced-choice" question form. Question forms in which the respondent is asked whether he or she agrees or disagrees with a given statement generate lower data. As described, we do not know for sure whether this effect on reliability is really the effect of the formulation or the specific effect of the topic "control in life", since an overlap existed between the statement formulation and this topic. However, the lower reliability of agree-disagree statements was also reported by Scherpenzeel and Saris (1997).

We expected the interviewing mode to interact with the question characteristics. However, most of the interaction hypotheses were completely rejected: the response scale effects were similar in both modes, face-to-face contact does not induce more social desirable answering than telephone contact, and memory effects are no more pronounced in CATI than in CAPI.

It is possible that the questions chosen are not as sensitive to the social desirability bias and to memory effects as we thought. It is also possible, however, that CATI is not experienced as more anonymous than CAPI and that time pressure is not felt stronger in CATI than in CAPI. This would mean that CATI and CAPI are more similar than we supposed in important interview-characteristics. One possible explanation is that the greater anonymity of the telephone interview is counteracted by a certain „unease" of respondents to discuss sensitive subjects by telephone (Leeuw and van der Zouwen, 1988). Being able to see the interviewer might increase trust. Groves and Kahn (1979) found, for example, that the major reasons of respondents to prefer face-to-face interviews were „know who you are talking to" and „liking to see who one is dealing with". Further research should show whether this explanation is correct.

[10] As was described earlier, it has to be noted that the agree-disagree statement form was used exclusively for the topic „control in life".

The only significant interaction effect we found was the interaction effect of mode and the formulation of the question on validity. In CAPI, the direct question formulation has a clearly higher validity than the agree-disagree form. The reliability was in both modes higher for the direct question formulation. This confirmed our a priori expectation that the "agreeing-response bias", as a form of socially desirable answering, would cause more correlated error in CAPI than in CATI. However, it contradicts the empirical finding above, that sensitive questions decrease validity in equal measure in both modes. A possible way to bring both findings in accordance is to assume that a more trustful atmosphere between respondent and interviewer in the face-to-face interview motivates respondents to give true (valid) answers on *direct* questions but also stimulates them to agree with *statements* of the interviewer, maybe to increase the trust even more. The "agreeing-response bias" then is not so much a form of "socially" desirable answering, but a sort of "situationally" desirable answering. In any case, this explanation strengthens the advantage of the direct question form over the agree-disagree statement form. In conclusion, CATI saves time and costs in comparison with CAPI and should not be regarded a second choice among data collection techniques. The response rates and data obtained by CATI are at least as good as those obtained by CAPI. The response rates on the household level are even somewhat better with CATI, although this difference is compensated by a slightly lower individual response rate. Very similar in both modes of interviewing are most answer distributions, as are the overall validity and reliability of the data. The only reason to prefer CAPI over CATI is when questions that require a good memory are asked. However, strategies exist that help to overcome this shortcoming, such as introducing memory aids and stimulating respondents to take their time to think.

References

Alwin, D.F. (1997). Feeling thermometers versus 7-point scales: Which are better? Sociological Methods and Research 25, 318-340.

Alwin, D.F. and Krosnick, J.A. (1991). The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociological Methods and Research*, 20, 139-181.

Andrews, F.M. (1984). Construct validity and error components of survey measures: a structural modelling approach. *Public opinion quarterly*, 48, 409-422.

Arbeitsgemeinschaft LINK / DemoSCOPE, (1997), Evaluation des Pretests EVE. Bundesamt für Statistik, Einkommens- und Verbrauchserhebung 1998. Luzern, LINK.

Arbuckle, J.L. Amos 3.62. Copyright 1994-1997 SmallWaters Corporation, Chicago.

Bishop, G.F., Hippler, H.J., Schwarz, N., Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (Eds.). *Telephone survey methodology*, New York, Wiley and Sons.

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

Catlin, G. and Ingram, S. (1988). The effects of CATI on costs and data quality: a comparison of CATI and paper methods in centralised interviewing. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (Eds.). *Telephone survey methodology*, New York, Wiley and Sons.

de Leeuw, E.D. (1992). *Data quality in mail, telephone and face to face surveys*. Amsterdam, TT-publicaties.

de Leeuw, E.D. and van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (Eds.). *Telephone survey methodology*, New York, Wiley and Sons.

Dillman, D.A. (1978). *Mail and telephone surveys*. New York, Wiley and Sons.

Forsman, G. and Schreiner, I. (1991). The design and analysis of reinterviews: An overview. 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.

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

Groves, R.M. and Kahn, R.L. (1979). *Surveys by telephone*. New York, Academic Press.

Groves, R.M. and Mathiowetz, N.A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.

Groves, R.M. and Nicholls, W.L., II (1986). The status of computer-assisted telephone interviewing: Part II – Data quality issues. *Journal of Official Statistics*, 2, 117-134.

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.

Jordan, L.A., Marcus, A.C. and Reeder, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. *Public opinion quarterly*, 44, 210-222.

Kaase, M. and Saris, W.E. (1997). The Eurobarometer – A tool for comparative survey research. In: Saris, W.E. and Kaase, M. (Eds.). *Eurobarometer measurement instruments for opinions in Europe*. Mannheim, ZUMA.

Kalfs, N. (1993). *Hour by hour: effects of the data collection mode in time use research*. Amsterdam, NIMMO.

Költringer, R. (1993). Messqualität in der sozialwissenschaftlichen Umfrageforschung. Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Wien.

Körmendi, E. (1988). The quality of income information in telephone and face to face surveys. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (Eds.). *Telephone survey methodology*, New York, Wiley and Sons.

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

Lyberg, L.E. and Kasprsyk, D. (1991). Data collection methods and measurement errors: an overview. 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.

Martin, J., O'Muircheartaigh, C. and Curtice, J. (1993). The use of CAPI for attitude surveys: An experimental comparison with traditional methods. *Journal of Official Statistics*, 3, 641-661.

Nicholls, W.L., II, and Groves, R.M. (1986). The status of computer-assisted telephone interviewing: Part I – Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.

Rindskopf, D. (1984). Structural equation models: empirical identification, Heywood cases, and related problems. Sociological Methods and Research, 13, 109-119.

Rodgers, W.L., Andrews, F.M. and Herzog, A.R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, 8, 251-275.

Saris W.E. (1991). Computer-assisted interviewing. Newbury Park, Sage Publications.

Saris, W.E. (1998). The split ballot Multitrait-Multimethod experiment: An alternative way to evaluate the quality of questions. Forthcoming.

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. (1995). Effects of data collection technique on the quality of survey data: an evaluation of interviewer- and self-administered computer assisted data collection techniques. In: Scherpenzeel, A.C. *A Question of Quality. Evaluating survey questions by multitrait-multimethod studies*. Dissertation. Royal PTT Nederland NV, KPN Research.

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.

Schuman, H and S. Presser (1981). *Questions and answers in attitude surveys: experiments on question form, wording and context*. Academic Press, New York.

Schwartz, N., Strack, F., Hippler, H.J. and Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied cognitive psychology*, 5, 193-212.

Silberstein, A.S. and Scott, S. (1991). Expenditure diary surveys and their associated errors. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and sudman, S. (Eds.), *Measurement errors in surveys*. New York: John wiley, pp303-327.

Snijkers, G.J.M.E. (1992). Computer gestuurd enquêteren: telefonisch of persoonlijk?. (Computer assisted interviewing: telephone of face to face?). *Kwantitatieve Methoden*, 39, 53-69.

SPSS for Windows, (1999). Release 10.0.5 (27 Nov. 1999), SPSS Inc.

Sudman, S. (1967) *Reducing the cost of surveys*. Chigaco, Aldine.

Sudman, S., Bradburn, N. (1974). *Response effects in surveys*. Aldine, Chicago.

Sykes, W. and Collins, M. (1988). Effects of mode of interview: experiments in the UK. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (Eds.). *Telephone survey methodology*, New York, Wiley and Sons.

Sykes, W. and Hoinville, G. (1985). *Telephone interviewing on a survey of social attitudes: a comparison with face-to-face procedures*. London, Social and Community Planning Research.

Veenhoven, R. (1984). *Conditions of happiness*. Dordrecht, Reidel Publishing Company.

Widaman, K.F. (1985). Hierarchically tested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.