Survey Research Methods (2007)
http://w4.ub.uni-konstanz.de/srm
Vol. 1, No. 1, pp. 29-43

© European Survey Research Association

Estimation of the effects of measurement characteristics on the quality of survey questions

Willem E. Saris and Irmtraud Gallhofer
ESADE, Universitat Ramon Llull

When designing questionnaires, many choices have to be made. Because the consequences of these choices for the quality of the questions are largely unknown, it has often been said that designing a questionnaire is an art. To make it a more scientific activity we need to know more about the consequences of these choices. In order to further such an approach we have:

1. made an inventory of the choices to be made when designing survey questions and created a code book to transform these question characteristics into the independent variables for explaining quality of survey questions;
2. assembled a large set of studies that use Multi-trait Multi-method experiments to estimate the reliability and validity of questions
3. carried out a meta-analysis that relates these question characteristics to the reliability and validity estimates of the questions.

On the basis of the results of these efforts we have constructed a database. This data base contains at present 1023 measurement instruments based on 87 experiments conducted on random samples from sometimes regional but mostly national samples of 300 to 2000 respondents. The database contains information on studies of reliability and validity of survey questions formulated in three different languages: English, German and Dutch. The purpose of this study was to generate cross national generalizations of the findings published so far drawn from national studies. This analysis provides a quantitative estimate of the effects of the different choices on the reliability, validity and the method effects.

Keywords: Reliability and validity of questions, meta analysis, MTMM

Introduction

The development of a survey item demands that many choices be made. Some of these choices follow directly from the aim of the study, such as the choice of the actual domain of the survey item(s) – e.g., church attendance, neighborhood etc.- and the conceptual domain of the question – e.g., evaluations, norms etc. As these choices are directly related to the aim of the study the researcher doesn't have much freedom of choice. But there are also many choices that will influence the quality of the survey item and are not fixed. These choices have to do with the formulation of the questions, the response scales and additional components such as an introduction, a motivation etc., the position in the questionnaire and the mode of data collection.

The effects of several of these choices on the response distributions have been studied in many ways by many people. The following studies provide typical examples of studies of response effects: Belson (1981), Sudman and Bradburn (1982), Schuman and Presser (1981), Billiet et al. (1986), Molenaar (1986), Presser and Blair (1994), Forsyth et al. (1992), Esposito et al. (1991), (1997), Sudman et al. (1996), Van der Zouwen (2000), Graesser (2000), Graesser et al. (2000), Tourangeau et al. (2000).

In most of these approaches, the research is directed to problems in the understanding of the survey items by the respondent. The hypothesis is that problems in the formulation of the survey item will affect the quality of the responses but the standard criteria for data quality, such as validity, reliability and method effect are not directly evaluated.

Campbell and Fiske (1959) suggested that validity, reliability and method effects can be directly evaluated if more than one method is used to measure the same traits. Their design is called the Multitrait Multimethod or MTMM design. In psychology and psychometrics much attention has been paid to this approach. For a review, we refer to Wothke (1996) and Eid and Diener (2006). In marketing research too, this approach has attracted much attention (Bagozzi and Yi 1991). In survey research, this approach has been applied by Andrews (1984). Andrews (1984) also suggested using meta-analysis of the available MTMM studies to determine the effect on the reliability, validity and method effects of different choices made in the design of survey questions.

His suggestion is relevant because it is not possible to derive general conclusions from single MTMM studies. All variations in methods studied are placed in a specific context, i.e., a specific mode of data collection, specific variables, specific question structures etc. A meta analysis of a large enough series of MTMM studies can allow an estimation of the different effects of the choices made in questionnaire design on the reliability, validity and method effects of survey

Contact information: ESADE, Universitat Ramon Llull, Av. de Pedralbes, 60-62 08034 Barcelona, Spain (w.saris@telefonica.net).

questions. That is what we are planning to do in this paper.

So this study deviates in two points from the above mentioned studies. In the first place we concentrate on the reliability and validity of survey questions and not on the response distributions. Secondly, we do a meta analysis across a large number of MTMM studies to derive general statements about the effects of the choices on the reliability and validity by a multivariate analysis

All MTMM experiments, based on at least regional random samples, performed in the period between 1979 and 1997, known to us, have been collected. These studies come from Andrews (1984) and Rogers, Andrews and Herzog (1992) in the US; Költringer (1995) in Austria, Scherpenzeel and Saris (1997) in the Netherlands and Billiet and Waege (1989, 1997) in Flanders (Belgium). The MTMM experiments were conducted in ongoing survey research. Some questions from the surveys were chosen to be repeated using a different method at the end of the substantive study. This means that the experiments were directed to evaluate single questions and not composite scores as more frequently has been done (Bagozzi and Yi 1991). This limits the number of data sets included in this study. In total, 87 MTMM studies have been found containing 1023 survey items in three languages: English, German and Dutch. A meta-analysis of these 87 studies will be reported. An overview of studies has been presented in the Appendix.

Looking at the coding systems used in the different countries Scherpenzeel (1995) came to the conclusion that the results of these studies could not be compared due to the lack of comparability of the coding systems used. Therefore, all questions of these studies have been coded again, using the same coding system. The choice of the variables to code the questions can be found in Saris and Gallhofer (2007). The codebook used in this study can be obtained from the authors.[1] Here we will present only a short overview of the variables generated by the coding system of the choices made in designing a survey question used in this cross national study. These question characteristics will be used as explanatory variables for the reliability and validity of the questions. After the explanatory variables are introduced the estimation of reliability and validity (the explained variables) using MTMM experiments will briefly be discussed. Then the meta analysis can be discussed and the results will be commented upon.

The explanatory variables: the choices made in the development of a survey item

A survey item consists of *several components*. We suggest that a survey item may contain the following components:

- introduction
- information about the topic or definitions
- instruction to respondent/interviewer
- opinions of others
- requests for an answer
- answer categories

In general, not all these components will occur at the same time. Only a request for an answer must be available. Since the request is not always formulated as a question (see also Tourangeau et al. 2000) but can also be formulated as an instruction or an assertion, we call this component a "request for an answer" and not a question. A request for an answer will always be available. It is unlikely that more than two of these components will accompany the request for an answer. Given the importance of the requests, we will begin with the choices related to this component and, following that, we will discuss the choices related to the other components.

The domain of the request

The first choice to be made has to do with the *Domain of the request*. This choice is of course completely determined by the aim of the study. If one is interested in the evaluation of the government, the domain is the government and one cannot change that. It will be clear that requests for an answer can refer to many domains. Therefore the classification of domains is rather difficult. Coding the requests for an answer we have used an elaborate classification of domains developed and used by the Central Data Archive in Cologne (Germany) to classify survey items. However in our analysis, only a rough classification could be used which is indicated in Table 1.

The concepts

A second choice that has to be made in the development of a request for an answer has to do with the *concept* that one would like to measure. The link between different concepts of the social sciences and requests that can be used in survey research has been discussed in Saris and Gallhofer (1998), Gallhofer and Saris (2000) and Saris and Gallhofer (2004), (2007). In these papers it is shown that all well known social science concepts such as feelings, evaluations, norms etc. can be transformed into assertions and assertions can be transformed into requests. Secondly, a fundamental distinction is made between concepts measured by simple requests and concepts that are operationalized by complex assertions or requests. An assertion becomes complex if it is an assertion about an assertion. The designer has the choice of using a simple or a complex assertion. Complex assertions are used as measures of the strength of opinions (Krosnick and Abelson 1991).). Many different simple concepts have been distinguished in the codebook but in the analysis only a limited number could be used because of dependencies with domains and the low frequency of the occurrence of some concepts in the set of questions used in the experiments. For the complete list of concepts we refer to Saris and Gallhofer (2007). The short list used in the analysis can be found in Table 1.

Associated characteristics

With the choice of the domain and the concept, other characteristics are determined. We call them associated char-

[1] The codebook can be found on the website www.sqp.nl

acteristics. In this respect we refer to Social Desirability, Centrality and Time specification. Social desirability requires a subjective judgment of the coder with regard to the desirability of different response alternatives. Centrality or saliency of the topic for the respondent can also not objectively be determined. It has been suggested to consider how many people would not know how to answer the request. The time specification is much simpler; it refers to whether the request concerns the past, present or the future.

Regarding the choices discussed so far, it will be clear that the designer of the questionnaire has little freedom. The choices are mainly determined by the research problem and the purpose of the specific request. For the choices which follow below the designer has much more freedom of choice.

The formulation of the request

In specifying the *formulation of the request* the designer has much more freedom. There are many different ways in which requests for answers can be formulated. The most common way, in many languages, is the specification of a request by inversion of the subject and the (auxiliary) verb. We call this "a simple or direct request". A different approach is to use a statement or stimulus representing the concept the researcher wishes to measure. The request for an answer can then be formulated as an "agree/disagree" request or as an instruction to answer in a specific way. This type of requests formulated by sentences as "Do you agree or disagree that ..." or "Do you think that ..." has been called an indirect request (Saris and Gallhofer 2004).

Sometimes special words are used in requests: "who, which, what, when, where and how". Such requests are called "WH" requests. These WH words can also be paraphrased by using for example "at what moment" instead of "when" etc.

Given the discussed choices we have made the following distinctions:

1. Simple or direct requests
2. Indirect requests such as Agree/disagree requests
3. Other requests using terms like "Who, Which, What, When, Where, How, Why", also called WH requests.

Furthermore, one can ask people to indicate the degree in their opinion or the strength of their agreement by asking "How much ...". If such phrases are used, these requests are coded as requests with *gradation*.

Besides these basic choices, many more choices have to be made in specifying a request in the strict sense. Here we would like to mention

- The use of an *absolute or comparative* statements
- A request with *balanced or unbalanced response alternatives* in the query part
- *Stimulation* to answer included in the request or not
- Emphasis to give the *subjective opinion* or not
- Presence or absence of *extra information* in the request; for example, definitions or explanations
- *Arguments* for the different opinions are included in the request or not

All these choices have to be made and are made in practice whether we realize it or not.

The response scale

The next component about which the designer of a survey item has to make decisions is the response scale. Again there are many possibilities. The most fundamental decision is whether one uses an *open ended* request or a *closed* request. If one has chosen a closed request one still has a choice with respect to the *scale type*:

1. a category scale with 2 categories (yes/no)
2. a category scale with more categories
3. frequency
4. magnitude estimation where the size of the number indicates the opinion
5. line drawing scale where the length of the line indicates the opinion
6. more steps procedure

Besides the basic choice regarding the type of scale, one has to make many more choices which have been presented in Table 1. Some of these choices have to be explained.

First of all the variable "Range". This variable is introduced because of the fact that there is sometimes a difference between theoretically possible range of the scales and the range of the scale used. For example scales can go from "very dissatisfied" to "very satisfied" (bipolar) while in the study the scale goes from "not satisfied" to "very satisfied" (unipolar).

Another coding variable to be explained is " the number of fixed reference points". Here we refer to the fact that people can have a different interpretation of a term like "very satisfied". The position on a scale can be different for different people. Some may see "very satisfied" as the end point of the scale but others not. But if one uses the term "completely satisfied" there can not be any doubt about the position of that term. This is the end point of the scale and that is therefore called a fixed reference point.

All other distinctions are more obvious. For more details we refer to Saris and Gallhofer (2007).

Presence of other parts of the survey item

A survey item can stand alone or can be placed in a battery of similarly formulated survey items. In a battery the request or instruction is normally mentioned only once, before the first stimulus or statement is provided. This raises the question what text belongs to the survey items after the first one; Should we include the request and the answer categories or not? We have decided that the request belongs to the first survey item and not to the latter ones because the text will not be repeated. That means that the items after the first item in a battery will not have a request or instruction, but will consist only of a stimulus or statement and answer categories.

Another distinction relates to the amount of text provided in the request it self. As was mentioned above, a survey item can contain many different components besides the request for an answer and the response categories. On this

point the designer again has a choice, but it is clear that the more parts are included the longer the item becomes. This can have a negative effect on the response and the quality of the response.

We have looked at the following parts to ascertain whether they were present next to the request for an answer:

1. Presence of emphan introduction
2. Presence of *a motivation*
3. Presence of *information regarding the content*
4. Presence of *information regarding a definition*
5. Presence of an *instruction to the respondent*
6. Presence of an *instruction to the interviewer*

Besides the choice of different components for the survey item one can also formulate the item in more or less complex ways. This can be evaluated as follows:

1. The *number of interrogative* sentences
2. The number of *subordinate clauses*
3. The *total number of words* in the survey item
4. The *average number of words* of the sentences
5. The *average number of syllables* per word
6. The total *number of nouns* in the request text
7. The *percentage of abstract nouns* relative to the total number of nouns

Data collection method

Furthermore a choice is made (mostly before any other choice) concerning the mode of data collection. We have operationalized this choice in the following possibilities:

1. *Computer assisted* data collection or not
2. *Interviewers administered* or not
3. *Visual information* used or not

On the basis of these choices the different data collection methods can be characterized.

Position of the item in the questionnaire

Other decisions have to do with the design of the whole questionnaire and the connection between the different requests in the questionnaire. The first point we would like to mention is the choice whether or not to use batteries of similar requests.

The second point has to do with the position of an item in the questionnaire. It is not clear what the optimal position is, but, in any case, not all items can be optimally placed so one has to look for an optimal solution considering all items.

A third point would be the layout of the questionnaire: the routing and the position on the page or screen etc. This aspect has not been taken into account in this research because there is not even enough information about the choices we have to make, although first steps have been taken by Dillmann (2000).

Language used

Given that the data come from three different language areas it is necessary also to introduce as one of the possible explanatory variables the language which is used to formulate the questions. This can of course make a difference in the quality of the responses.

Sample characteristics

Since different samples have been used, a possible explanation for quality differences could also be the composition of the sample used in the study. It has often been suggested that lower educated and older people will produce lower quality data. We have added to this set the gender composition of the sample.

MTMM design

Finally, it can be expected that the design of the MTMM experiment itself has an effect on the quality estimates. It is well known that answers to similar questions which have been asked quickly after each other have higher correlations than answers to questions between which the distance is larger. The size of the correlation will affect the estimate of the quality of the question.

In MTMM experiments requests for the same concepts have to be repeated. Therefore a possible explanation of quality can be the relative distance between the requests for the same trait. Therefore characteristics of the design have also be included. The distance is measured in the number of requests between the repetitions of the same requests.

Estimation of reliability and validity

We have shown above that there are many choices that the designer of survey items is making. Each of these choices can have a positive, negative or no effect on the quality of the collected data.

Campbell and Fiske (1959) suggested the use of Multitrait-Multimethod (MTMM) experiments to evaluate the quality of measurements instruments. It is now a standard procedure to use three traits and three methods, i.e. 9 measures in such experiments. Since the same people answer all the questions, the explanation for the differences in the correlations between measures for the same traits is measurement error.[2] It is assumed that each method has its own random errors and systematic errors, also known as method effects. Formally this can be formulated in the following two sets of equations:

$$Y_{ij} = r_{ij}T_{ij} + e_{ij} \quad for \;\; i = 1-3 \;\; and \;\; j = 1-3 \qquad (1)$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j \quad for \;\; i = 1-3 \;\; and \;\; j = 1-3 \qquad (2)$$

where Y_{ij} is the measured variable (trait i measured by method j), T_{ij} is the stable component of the response Y_{ij} (also called "true scores"), F_i is the trait factor of interest, M_j is the method factor (whose variance represents systematic

[2] It is also possible that the differences are partially due to the fact that the questions are repeated and that respondents think about these questions between the two observations. Discussion of this point is reserved for another paper (Saris, Satorra, Coenders 2004).

method effects common for all traits measured with method j, but varying across individuals) and e_{ij} is the random measurement error term for Y_{ij} (with zero mean and uncorrelated with other error terms, with method the factors and the trait factors).

The r_{ij} coefficients standardized can be interpreted as reliability coefficients (square root of test-retest reliability). When standardized, the m_{ij} coefficients represent method effects. The v_{ij} coefficients standardized are validity coefficients (with v_{ij}^2 representing the validity of the measure). Note that the validity is only reduced by method effects in this model because $v_{ij}^2 = 1 - m_{ij}^2$. Whether the question really measures the concept of interest requires an evaluation of the link between the theoretical concept and the concept behind the question asked but that is another matter that is not the topic of this paper.

There have been many suggestions for further specification of the model formulated in the equations 1 and 2. Some authors leave all correlations between the traits and method factors free but mention many problems (Kenny and Kashy 1992, Marsh and Bailey 1991). Andrews (1984) and Saris (1990) suggested that the traits should be allowed to correlate but they should be uncorrelated with the method factors and the method factors should also be uncorrelated with each other. For a detailed discussion of the different models we refer to Wothke (1996), Coenders and Saris (2000) and Saris and Aalberts (2003). Using the model specification of Saris and Andrews (1991) scarcely any problems arise in the analysis as has been shown by Corten et al. (2002) and the model of Saris and Andrews (1991) turned out to be better fitting to several data sets (Saris and Aalberts 2003).

These MTMM experiments are useful to provide estimates of the quality for specific sets of questions. If one would like to make more general statements about the effects of the different choices on the quality of questions a multivariate analysis is necessary because questions studied have many different characteristics that all can affect the quality simultaneously. Such an analysis is possible on the basis of a meta analysis of a multitude of MTMM experiments.

The meta-analysis

As mentioned above, in total, 87 MTMM studies have been found containing 1023 survey items. All these studies are based on, at least, regional samples of the general population. In the US the Detroit area was used, in Austria and the Netherlands national samples were used, while in Belgium random samples of the Flemish speaking part of the population were taken. All experiments used in this meta-analysis have been based on Pearson's correlations even if the questions were categorical. One could also have used polychoric correlations as summary measures. Both possibilities are equally relevant, but as most substantive analyses are so far done using Pearson's correlation coefficients, we have chosen for this approach in the analysis of the MTMM experiments. For a more elaborate discussion of this point with an illustration of the effects on the results of the meta analysis we refer to Saris, Van Wijk and Scherpenzeel (1998).[3]

The topics in the different experiments are highly diverse. In general, the MTMM experiments are integrated into normal survey research where three or more questions of the survey are used for further experimentation. This approach guarantees that questions are used that are commonly used in survey research. The same is true for the variation in the choices made in the design of survey items. The experiments are designed in such a way that the most commonly used methods (choices) can be evaluated. For more details on the studies, we refer to the Appendix and the above mentioned publications.

All the 87 experiments mentioned in the Appendix have been analyzed with the model specified in the previous section. In most of the experiments the models fitted the data rather well but, if a very bad fit was obtained we allowed for some corrections in the model. These corrections were limited to theoretically acceptable possibilities, for example allowing for different effects of the methods for different traits or correlations between the method effects.

Given the different conditions under which the different experiments were done, the estimates of the quality coefficients varied quite a bit from study to study. In order to be able to make general statements about the effects of the different question characteristics on the quality of these questions, all questions were coded on the characteristics mentioned in the last section and a data file was made with as cases the 1023 questions and as variables the characteristics of the questions and the reliabilities and validities estimated using the MTMM experiments.

Normally, Multiple Classification Analysis (MCA) has been used (Andrews (1984), Scherpenzeel (1995), Költringer 1995) in the meta-analysis but this is not possible with many variables. Therefore, (dummy) regression has been used.

The following equation presents the approach used:

$$C = a + b_{11}D_{11} + b_{21}D_{21} + \cdots + b_{12}D_{12} + b_{22}D_{22} + \cdots + b_3\, Ncat + \cdots + e \tag{3}$$

In this equation, C represents the score on a quality criterion, either the reliability or validity coefficient. The variables Dij represent the dummy variables for the jth nominal variable. All dummy variables have the value zero unless the specific characteristic applies for a question. For all dummy variables, one category is used as the reference category . This category has received the value zero on all dummy variables of that set. Continuous variables, like the number of

[3] The major issue in this case is that the quality estimates can be used for correction for measurement errors. However, if the substantive analysis is done using Pearson's correlations the corrections should also be based on Pearson's correlations; if the substantive analysis is done using polychoric correlations, the corrections should also be based on polychoric correlations. So, in principle one needs two meta analyses: one based on Pearson correlations and one based on polychoric correlations. So far we have only done the analysis on the basis of Pearson's correlations. For a comparison of the results for a smaller set of variables we refer to the paper of Saris, Van Wijk and Scherpenzeel (1998).

categories (Ncat), have not been categorized except when it was necessary to take into account non-linear relationships. The intercept is the reliability or validity of the instruments if all variables have a score of zero. It will be noted that it was impossible to introduce all categories of all variables in the analysis. For such an analysis the number of questions was still too small. Therefore in several cases some categories have been combined with some others most similar categories.

Table 1 shows the results of the meta-analysis over the available 1023 survey items. Table 1 indicates the effects on the quality criteria validity and reliability of the different choices that can be made in the design of a survey question.[4] Note that in Table 1 all effects have been multiplied by 1000 in order to give a clearer picture of the effects.

Each coefficient indicates the effect on the reliability and validity of an increase of one point in each indicated characteristic while all other characteristics remain the same. For example, all questions concerning consumption, leisure, family, personal relations and race are coded as zero on all domain variables and this set of questions can be seen as the reference category. For these questions the effect on reliability and validity is zero. Questions concerning other issues are coded in several categories. If a question concerns "national politics" the question belongs to the first domain category ($D_{11} = 1$ for this category while all other domain variables $D_{i1} = 0$) and the effects on the reliability and validity are found to be .053 and .045 respectively as can be seen in the table. If a question concerns "life in general" then the 5th category applies (D51=1) and the effects are negative, -.077 and -.016 respectively. From these results it also follows that questions concerning national politics have a reliability coefficient which is .053 + .077 or .130 higher than the questions about life in general. This interpretation holds for all characteristics with a dummy coding such as "concepts", "time reference" etc.

Other characteristics with at least an ordinal scale are treated as metric. For example, "centrality" is coded in 5 categories from 'very central' to 'not central at all'. In this case an increase of one point gives an effect of -.0172 for the reliability and the difference between a very central or salient item and a not at all central item is 5 * -.0172 = -.0875.

Furthermore, there are real numeric characteristics like the "number of categories", "the number of words" etc. In that case, the effect is as always the effect of an increase of one unit, i.e. one word or one category. A special case in this category is the variable "position" because it turns out that while the effect of position on reliability is linear, it is non-linear for validity. To describe the later relationship, the position variable is categorized and the effects should be determined within each category.

Another exception is the "number of categories in the scale". For this variable we have specified an interaction term because the effect was different for categorical questions and frequency measures. So, depending on whether the question is categorical or a frequency question, a different variable has to be used to estimate the effect on the reliability and the validity.

The results of the Meta analysis

Table 1 presents the effects of the different choices described above on the *quality criteria* validity and reliability. The results presented in this table can be summarized as follows.

Domain, concept and associated characteristics

The domain, concepts and associated characteristics are no longer really open to choice, once the research design has been decided upon. Nevertheless, there are significant differences in reliability and validity between the different domains, concepts and associated characteristics:

In contrast with our expectations the quality of questions about political issues is much better than requests from other domains.

Behavioral survey items can be much worse than attitudinal questions, especially items concerning frequency of behavior. Complex items should be avoided if possible.

It appears that reporting about the past is more reliable than reporting about the future and the present.

Formulation of the requests

In formulating the requests the researcher has more freedom. We found here that indirect request like agree/disagree requests are more or less as good as direct requests with respect to reliability and a bit better with respect to validity.

The reliability is better when one uses requests with gradation, although this has also a small, non significant, negative effect on the validity.

Of other request characteristics, only lack of balance and emphasizing subjective opinion have a significant negative effect on the validity. Otherwise there is no significant effect.

Response scale

Use of response scales with gradation in the form of frequency, magnitude estimation or line production and stepwise procedure has a positive effect on the reliability but is often associated with method effects like rounding off. Line production and stepwise procedures incur smaller problems in this respect. The reason for this seems to be that all numeric scales will be affected by rounding off (Tourangeau et al. 2000) while that is not happening with line responses.

With respect to the use of labels it is wise to use at least some labels but not complete sentences. The former will improve the reliability.

The use of a neutral middle category and fixed reference points improves both the reliability and validity significantly.

The correspondence between the numbers and the labels has a significant positive effect on reliability while the asymmetry of the response categories has some positive effect on both quality criteria. This effect may be due to other related variables as yet unknown, since we can not explain it.

[4] The effects on the method effects have not been indicated because they can be derived from the validity coefficients.

Table 1: Results of the Meta-Analysis (coefficients * 1000)

Variables		**effect on reliability**			**effect on validity**		
	# measures	**effect**	**se**	**sign**	**effect**	**se**	**sign**
Domain							
National politics (0-1)	137	52.8	12.3	.000	44.7	10.9	.000
International politics (0-1)	64	29.4	18.1	.104	57.8	15.9	.000
Health (0-1)	82	16.9	13.9	.225	21.6	12.0	.073
Living cond/ backgr (0-1)	223	21.4	8.7	.014	4.6	7.4	.541
Life in general (0-1)	50	-76.8	12.6	.000	-15.9	10.8	.139
Other subj. var., Crimes, (0-1)	235	-66.9	14.2	.000	-1.0	12.4	.935
Work (0-1)	96	12.8	12.0	.287	28.2	10.4	.007
Others i.e.:Consumption(26)/ Leisure (59)/Family (3)/ Personal rel. (45)/Race (3)	136	0.0	–	–	0.0	–	–
Concepts							
Evaluative belief (0-1)	96	6.1	14.	.669	13.8	12.3	.260
Feeling (0-1)	110	-4.2	10.9	.704	-7.5	9.4	.427
Importance (0-1)	96	35.9	15.6	.021	18.6	13.6	.171
Future expectations (0-1)	39	2.6	24.0	.913	-9.0	20.6	.662
Behavior (0-1)	27	-126.2	21.8	.000	-150.5	19.2	.000
Complex concepts	77	-72.3	17.4	.000	-47.2	15.2	.002
Other concepts i.e.Evaluation (447)/ Judgment[a] (123)/ Norms/Rights/Policies (8)	578	0.0	–	–	0.0	–	–
Associated characteristics							
Social des :no/bit/much (0-2)	1023	2.3	6.2	.709	8.0	5.3	.137
Centrality very c -not (1-5)	1023	-17.2	5.2	.001	-8.9	4.4	.046
Time reference:							
Past (0-1)	106	43.9	15.0	.004	-1.6	12.9	.901
Future(0-1)	83	-13.3	16.1	.409	-10.1	13.8	.465
Presence (0-1)	940	0.0	–	–	0.0	–	–
Request formulation: Basic choices							
Indirect: agree/disagree(0-1)	167	4.0	10.9	.713	41.6	9.5	.000
Other types: i.e. Direct request (190) more steps[b] (22) and WH-questions(0)	212	0.0	–	–	0.0	–	–
Use of statements or stimulus (0-1)	317	-23.0	12.4	.065	-12.1	11.1	.275
Use of gradation(0-1)	809	79.6	14.1	.000	-22.8	12.4	.066

[a]The judgments have nearly all the domain score "'other beliefs"'. Therefore these two variables could not be used together in the analysis of this moment. Judgment is omitted. The effect of other beliefs could also be seen as an effect of the concept judgment.
[b]This variable also occurs in the characterization of the response scale. Therefore this category could not be treated here seperately.

Table 1: (continued)

Variables	# measures	effect on reliability			effect on validity		
		effect	se	sign	effect	se	sign
Request formulation: Other choices							
absolute-comparative (0-1)	98	12.7	16.3	.436	-8.4	14.5	.564
unbalanced(0-1)	411	-3.2	11.2	.772	-22.3	9.7	.022
stimulance (0-1)	92	-11.1	13.3	.406	-11.7	11.5	.308
subjective opinion(0-1)	86	-5.9	19.9	.767	-34.3	17.2	.047
knowledge given(1-4)	358	-12.7	8.8	.145	-6.3	7.5	.401
opinion given(0-1)	101	.653	14.5	.964	-10.3	13.1	.429
Response scale: Basic choices							
Yes/no (0-1)	37	-22.2	19.5	.254	-1.9	17.1	.911
Frequencies	23	120.8	24.8	.000	-95.9	21.5	.000
Magnitudes	169	116.2	20.8	.000	-115.5	18.3	.000
Lines	201	118.1	20.9	.000	-32.7	18.2	.073
More steps	26	48.7	27.3	.075	24.5	23.5	.297
Categories	630	0.0	–	–	0.0	–	–
Response scale: Other choices							
Labels: no/some/all (1-3)	1023	33.0	10.0	.001	-4.5	8.8	.605
Klabels: sentence (0-1)	35	-47.5	16.0	.003	-9.1	13.7	.506
Dnk :present/reg/no(1-3)	1023	-6.7	4.8	.165	-1.9	4.1	.647
Neutral (present/reg/no (1-3)	1023	12.6	4.6	.007	8.4	4.0	.038
Range (uni/bi-bi/bi-uni (1-3)	1023	-15.1	9.6	.116	9.2	8.5	.277
Corresp:high -low (1-3)	1023	-16.8	7.5	.025	1.1	6.5	.867
Asymmetric labels(0-1)	195	25.5	11.8	.031	22.3	10.4	.033
First neg/pos (1-2)	358	-7.5	8.7	.387	14.7	7.6	.052
Fixed references (0- ?)	1023	14.7	4.3	.001	21.4	3.7	.000
Ncateg (ncat*categ)[c](0-11)	1023	13.5	2.1	.000	-1.9	1.8	.298
Nfreq (ncat*freq)3 (0-5000)	1023	-.068	.009	.000	-.065	.008	.000
Item specification: Basic choices							
Direct question present (0-1)	841	27.2	15.2	.074	11.5	13.1	.379
Q-instruction present (0-1)	103	-43.7	15.4	.005	-4.2	13.3	.753
No request or instruction	79	0.0	–	–	0.0	–	–
Resp. instruction (0-1)	492	-12.7	7.3	.083	-14.9	6.2	.017
Interv. instruction (0-1)	119	-.068	10.5	.995	5.7	9.0	.524
Definitions (0-3) >0	304	7.1	6.7	.296	-.3	5.7	.959
Introduction (0-1)	515	5.7	12.1	.637	-10.5	10.3	.312
Item specification: Other choices[d]							
In introduction							
Interrogative sentence(0-1)	62	-44.6	16.3	.006	-21.3	14.1	.132
# Subordinate clauses >0	129	29.3	9.8	.003	7.6	8.6	.377
# Words >0	510	-1.3	.867	.134	1.4	.75	.063
# Mean words/Sentence >0	510	.064	1.1	.954	-.373	.9	.699

[c]Ncateg and nfreq are introduced because it was seen that the number of categories did not have a linear relationship with the dependent variables. For the category scale the effect was positive, for the frequency scale negative and for the others there was no effect at all. Therefore two interaction terms have been introduced.

[d]The variables Syllables/Word and proportion of abstract words have been coded for the introduction and the request but for the introduction these variables correlated very highly with each other and with the variable "'Intro"'. Therefore it was decided that these variables will not be used for the introduction.

Table 1: (continued)

Variables	# measures	effect on reliability			effect on validity		
		effect	se	sign	effect	se	sign
In request							
# Interrogative sentence =0	192	12.7	9.8	.199	-8.3	8.6	.335
# Subordinate clauses =0	746	13.6	6.8	.048	-17.7	5.9	.003
# Words (1-51)	1023	.809	.749	.280	-1.3	.644	.041
Mean words/Sentence (1-47)	1023	-2.2	.926	.014	1.1	.807	.161
Syllabi/Word (1-4)	1023	-32.5	9.6	.001	-10.4	8.2	.207
Abstract word/Word (0-1)	1023	2.9	27.7	.917	-13.9	23.7	.558
Mode of data collection							
Computer assisted(0-1)	626	-3.8	12.6	.760	-38.3	10.7	.000
Interviewer adm(0-1)	344	-50.8	22.9	.027	-104.1	19.5	.000
Oral(0-1)	219	10.4	12.2	.397	25.3	10.3	.014
Position in questionnaire[e]							
In battery(0-1)	225	-10.3	12.3	.403	28.9	10.7	.007
# Request (1-381)	1023	.304	.064	.000			
Pos25 (1-25)	396				1.5	.402	.000
Pos100 (26-100)	458				.420	.137	.002
Pos200 (101-200)	129				.267	.062	.000
Pos300 (>200)	12				.098	.100	.333
Language used in questionnaire							
Dutch (0-1)	731	-20.3	22.8	.373	-76.0	19.8	.000
English(0-1)	174	-72.0	26.6	.007	-2.9	22.9	.899
German (0-1)	118	0.0	–	–	0.0	–	–
Sample characteristics							
% Low education (3-54)	993	-.911	.596	.127	1.1	.511	.027
% High age (1-49)	1023	-.410	.560	.464	-.753	.488	.123
% Males (39-72)	1023	-.030	.690	.966	.405	.596	.497
MTMM design							
Design one/more time points (0-1)	713	4.36	16.3	.790	-36.9	14.3	.010
Distance between repeated methods (1-250)	1023	-.169	.094	.072	-.249	.081	.002
# Traits (1-10)	1023	-.370	2.0	.855	-1.7	1.7	.320
# Methods (1-14)	1023	.959	2.6	.715	-2.3	2.2	.314
Intercept:		825.2	69.5	.000	1039.4	60.4	.000
Explained variance (adjusted)		.47			.61		

[e]The reliability seems to be linear related with position but validity is not. Therefore a split variable is used in the analysis of the validity and not reliability.

With respect to the effect of the number of categories two cases should be distinguished:

1. In the case of a category scale, reliability can be increased by more than .10 going from a 2 point to an 11 point scale.
2. In the case of a frequency scale, reliability and validity are seriously damaged if the range of the scale is too large, i.e., if very high frequencies are possible.

Specification of the survey item as a whole

Compared with items later in a battery with no request or instruction, the first item is more reliable if a normal request is asked and less reliable if an instruction is used.

Instructions for the respondent have a significant negative effect on the reliability and validity. The item is probably so difficult that an explanation is needed. It may therefore be an effect of the item and not of the instruction.

Interviewer instructions, extra motivational remarks, definitions and an introduction seem to have no significant effect on reliability or validity.

Formulating of general questions in the introduction, followed by the real request, should be avoided because this has a rather strong negative effect on reliability and validity. On the other hand, it seems to have a positive effect on reliability if, in the introduction, an explanation is given in subordinate clauses. Such extra information in the request for an answer has also a positive effect on the reliability but this effect is cancelled out by a negative effect on the validity.

There is clearly a limit to this possibility of providing more information because the indices for complexity of the requests, the number of words per sentence (sentence length) and the number of syllables per word (word length) have a significant negative effect on the reliability of the responses.

Mode of data collection

The mode of data collection can be specified by some basic methods or by description in general terms. We have chosen in the analysis for the latter option. We then see that Computer Assisted Interviewing (CAI) is as reliable as non-CAI but a bit less valid. A much stronger negative effect can be observed for interviewer-administered questionnaires. Oral questionnaires have a small but significant positive effect on the validity. The effects of the interviewer-administered questionnaires require two comments.

1. The first is that we can not control for the interaction with the complexity of the question. It could be that in the face to face questionnaires more complex questions have been asked than in the self-administered questionnaires.
2. Secondly, it should be mentioned that in the choice of the mode of data collection other quality criteria should also be considered; for example, unit non-response and item non-response.

We will come back to this issue in the conclusion.

Position in the questionnaire

The effect of the position in a questionnaire is rather different for reliability than for validity. It seems that because the respondents continue to learn how to fill in the questionnaire, the reliability increases linearly with the position in the questionnaire. Over the range studied the effect can be more than .10. Validity, on the other hand, increases rather rapidly in the beginning: .037 over the first 25 requests, then from 25 till 100 still .031, from 100 till 200 only .026 and after the 200 there is no further significant increase any more.

Although the increases in validity decrease, we have found that the validity and the reliability increase over the whole range of the questionnaire (up to 250 requests for answers).

Basic choices for which correction is necessary

Finally there are choices that are not explicitly made, like the choice for a language and the characteristics of the population. These choices can nevertheless have an influence on the quality criteria. In addition the methodological experiments which form the basis for this meta-analysis also have some influence which have been estimated and controlled when other effects are estimated. The estimated effects of these factors can be summarized as follows:

Compared with questionnaires in German, questionnaires in English are significantly less reliable, while Dutch questionnaires are significantly less valid.

Of the three characteristics of the samples studied only the level of education has a significant effect on the validity of responses. Samples with many poorly educated people can have a validity which is more than .050 lower than samples with few poorly educated people.

The MTMM design also has a significant effect on the data quality: if the distance, measured in the number of questions between the items for the same trait, becomes larger, reliability declines. For the largest distance (alone) the reliability can be .042 lower. Distance has an even larger effect on validity. For the largest distance (alone) the validity can be 0.62 lower.

Conclusions and limitations

We will start by giving special attention to effects that are results of combinations of choices. The first has to do with the choice of the number of categories and the second with the choice of a mode of data collection. We start with the effect of the number of categories.

There is still no consensus about the effect of an increase in the number of categories in the scale. Cox (1980) and Krosnick and Fabrigar (forthcoming) defend the position that one should not use more than 7 categories while Andrews (1984), Költringer (1995) and Alwin (1997) defend the other position that one gets always better results if one uses more categories. Our results suggest that frequency scales, magnitude scales and line scales are all on average more reliable than category scales on average. But frequency and magnitude scales especially pay a price for this reliability in the form of a much lower validity. This problem is due to the use of labels that allow people to use their own scale, causing what Saris (1988) has called “variation in response functions”. Saris suggested as a solution of this problem, the use of fixed reference points. This solution is confirmed here because better validity and reliability will be obtained if fixed reference points are used.

Category scales can also be improved with regard to reliability by using more categories (so far up to 11 categories were studied) without harming the validity. An alternative is to use a two step procedure which improves both the reliability and the validity. But the disadvantage of this approach is that these scales will often generate 3 modal categories: one in the middle and one at both sides (Klingemann 1997). Such distributions are very unusual for normal category scales. These scales can not be compared.

Category scales can also be improved by use of labels as long as they are not full sentences.

All in all, this analysis indicates quite strongly that one should use as many categories as possible in a category scale (more than 7) but with clear short labels.

Table 2: Effect of data collection mode on reliability

Administered by	CAI	not CAI
Interviewer	-.0538	-.050
Respondent	-.0038	.000

Table 3: Effect of data collection mode on validity

Administered by	CAI	not CAI
Interviewer	-.1423	-.104
Respondent	-.0383	.000

If it is possible, it would be even better (nearly always) to use line production or analog scales as they are also called but with fixed reference points.

The second set of effects to be discussed has to do with the selection of the mode of data collection. The commonly used data collection methods can be constructed by combining the mode of administration (interviewer vs. respondent self-admistration) with computer-aiding or not. The effects of these combinations on reliability and validity are presented in Table 2 and Table 3.

This presentation suggests the following order in quality with regard to validity and reliability:

1. MAIL
2. CASI
3. PAPI/TEL
4. CATI/CAPI

The differences between MAIL and CASI are minimal, on the other hand, differences between these two and the PAPI/TEL or CAPI /CATI are large. It should be mentioned that other quality criteria in the mode of data collection choice should also be considered such as unit non-response and item non-response. In general, mail surveys have lower response rates although the use of the total design method can reduce the problem (Dillman 1978, 2000). Therefore, the results suggest that a trade-off between quality, with respect to reliability and validity, and item non-response has to be made.

The results in Table 1 indicated that it happens that the effects on reliability and validity sometimes are in the opposite direction. For some variables we have indicated why these differences can be expected. In other cases this is unclear. An overall measure of quality which follows directly from the specified model and takes this problem into account, is the product of reliability and validity coefficients.

To estimate the quality of an item one has to make a prediction of the reliability and validity of a question. This, however, would mean a lot of work, because the questions first have to be coded and then the effects of Table 1 have to be applied. For a complete questionnaire this would involve a great deal of work although it is not impossible as was shown in the evaluation of the questionnaire of the European Social Survey in 2002. In order to facilitate this possibility, work has been done on the development of computer programs for these predictions. This program has been discussed in a paper of Saris et al. (2004) and can be down loaded from the web.[5]

This analysis presents an intermediate result. So far 87 studies have been reanalyzed with in total 1023 survey items. This seems to be a lot but it is not enough to evaluate all variables in detail. Important limitations are that:

1. only the main categories of the domain variable have been taken into account;
2. requests concerning consumption, leisure, family and immigrants could not be included in the analysis;
3. the concepts norms, rights and policies have been given too little attention;
4. the request types of open-ended requests and WH requests have not yet been studied;
5. mail and telephone interviews were not sufficiently available to be analyzed separately;
6. there is an overrepresentation of requests formulated in the Dutch language;
7. Also only a limited number of interactions and non-linearities could be introduced.

The database will be extended in the future with survey items that are at present underrepresented. For the moment, we can only present the results of the analysis of our present set of survey items. To evaluate the importance of an effect one can look at the amount of items it is based on and the standard error of the estimate.

Nevertheless, taking these limitations into account, it is remarkable that the analysis has shown that the choices discussed above can explain almost 50% of the variance in reliabilities and 60% of the validity which has been estimated for the 1023 survey items. In this respect, it is also relevant to refer to the standard errors of the regression coefficients which are relatively small. This indicates that the correlations between the variables used in the regression as explanatory variables are relatively small.

If one takes into account that all estimates of the quality criteria contain errors while in the coding of the survey item characteristics, certainly, errors are also made, it is remarkable that such a high explained variance could be obtained.

This does not mean that we are satisfied with this result. Certainly, further research is needed as we have indicated above. But we also think that Table 1 is the best summary we can make at this moment of the effects of the choices made designing a questionnaire on reliabiltiy and. We therefore recommend the use of these results provisionally until the new results have been incorporated in the meta-analysis and a new Table 1 can be produced.

References

Alwin, D. (1997). Feeling thermometers versus 7-point scales: Which are better ? *Sociological Methods and Research*, *25*, 318-341.

[5]The program SQP can be found on the website www.sqp.nl

Andrews, F. (1984). Construct validity and error components of survey measures. a structural modeling approach. *Public Opinion Quarterly*, *48*, 409-442.

Bagozzi, R. P., & Yi, Y. (1991). Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research*, *17*, 426-439.

Belson, W. (1981). *The design and understanding of survey questions*. London: Gower.

Billiet, J., Loosveldt, G., & Waterplas, L. (1986). *Het survey-interview onderzocht: effecten van het ontwerp en gebruik van vragenlijsten op de kwaliteit van de antwoorden(research on surveys: effects of the design and use of questionnaires on the quality of the responses)*. Leuven: Sociologisch Onderzoeksinstituut KU Leuven.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait multimethod matrices. *Psychological Bulletin*, *56*, 81-105.

Coenders, G., & Saris, W. (2000). Testing nested additive, multiplicative and general multitrait- multimethod models. *Structural Equation Modeling*, *7*, 219-250.

Corten, I., Saris, W., Coenders, G., Veld, W., Aalberts, C., & Kornelis, C. (2002). Fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, *9*, 213-232.

Cox, E. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing research*, *17*, 407- 422.

Dillman, D. (2000). *Mail and internet surveys: The tailored method.* New York: Wiley.

Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology*. Washington,DC: American Psychological Association.

Esposito, J., Campanelli, P., Rothgeb, J., & Polivka, A. (1991). *Determining which questions are best: Methodologies for evaluating survey questions.*

Esposito, J., & Rothgeb, J. (1997). Evaluating survey data: Making the transition from pretesting to quality assesment. In L. Lyberg et al. (Eds.), *Survey measurement and process quality* (p. 541-571). New York: Wiley.

Forsyth, B., Lessler, J., & Hubbard, M. (1992). Cognitive evaluation of the questionnaire. In C. Tanur & R. Tourangeau (Eds.), *Cognition and survey research* (p. 183-198). New York: Wiley.

Gallhofer, I., & Saris, W. (2000). Formulierung und Klassifikation von Fragen. *ZUMA Nachrichten*, *46*, 43-72.

Graesser, A., Wiemer-Hastings, K., Kreuz, R., & Wiemer-Hastings, P. (2000). Quaid: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, and Computers*, *32*, 254-262.

Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (2000). *The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices.*

Kenny, D., & Kashy, D. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165-172.

Klingemann, H. (1997). The left-right self-placement question in face to face and telephone surveys. In W. Saris & M. Kaase (Eds.), *Eurobarometer: Measurement instruments for opinions in europe* (Vol. 2, p. 113-125). Mannheim: Zuma Nachrichten Spezial.

Költringer, R. (1995). Measurement quality in austrian personal interview surveys. In W. Saris & A. Münnich (Eds.), *The multitrait-multimethod approach to evaluate measurement instruments* (p. 207-225). Budapest: Eötvös University Press.

Krosnick, J., & Abelson, R. (1991). The case for measuring attitude strength in surveys. In J. Tanur (Ed.), *Questions about questions. inquiries into the cognitive bases of surveys* (p. 177-203). New York: Russel Sage Foundation.

Krosnick, J., & Fabrigar, L. (forthcoming). *Designing questionnaires to measure attitudes*.

Marsh, H., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, *15*, 47-70.

Molenaar, N. (1986). *Formuleringseffecten in survey-interviews*. Amsterdam: VU uitgeverij.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? In Marsden (Ed.), *Sociological methodology* (p. 73-104). Oxford: Basil Blackwell.

Rogers, W., Andrews, F., & Herzog, A. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, *8*, 251-275.

Saris, W. (1988). *Variation in response functions: a source of measurement error in attitude research*. Amsterdam: Sociometric Research Foundation.

Saris, W. (1990). The choice of a model for evaluation of measurement instruments. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (p. 118-133). Amsterdam: North Holland.

Saris, W., & Aalberts, C. (2003). Different explanantions for correlated errors in mtmm studies. *Structural Equation modeling*, *10*, 193-214.

Saris, W., & Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (p. 575-599). New York: John Wiley & Sons.

Saris, W., & Gallhofer, I. (1998). Classificatie van survey-vragen. *Tijdschrift voor Communicatie wetenschap*, 96-122.

Saris, W., & Gallhofer, I. (2004). Operationalization of social science concepts by intuition. *Quality and Quantity*, *38*, 235-258.

Saris, W., & Gallhofer, I. (2007). *Design, evaluation and analysis of questionnaires or survey research*. New York: Wiley.

Saris, W., & Krosnick, J. (2007). Comparing the quality of agree/disagree questions and balanced forced choice questions via a split ballot mtmm experiment.

Saris, W., Satorra, A., & Coenders, G. (2004). A new approach for evaluating quality of measurement instruments: Split ballot mtmm design. In R. Stoltenberg (Ed.), *Sociological methodology* (p. 311-347). Blackwell.

Saris, W., Wijk, T., & Scherpenzeel, A. (1998). Validity and reliability of subjective social indicators: the effect of different measures of association. *Social Indicators Research*, *45*, 173-199.

Scherpenzeel, A. (1995). *A question of quality. evaluating survey questions by multitrait-multimethod studies*. Leidschendam (the Netherlands): KPN Research.

Scherpenzeel, A., & Saris, W. (1997). The validity and reliability of survey questions: a meta-analysis of mtmm studies. *Sociological Methods and Research*, *25*, 341-383.

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

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey Bass.

Sudman, S., Brandburn, N., & Schwartz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Wothke, W. (1996). Models for multitrait-multimethod matrix analysis. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling. issues and techniques* (p. 7-56). Mahwah, N.J.: Lawrence Erlbaum.

Zouwen, J. (2000). An assessment of the difficulty of questions used in the issp questionnaires, the clarity of their wording and the comparability of the responses. *ZA-Informationen*, *45*, 96-114.

Appendix: Overview of the experiments used in the analyses in 2001

Country	number	year	design	datacollection	organisation	topic
NL	101	92	3x2x2	mail/telep	STP	seriousness of crimes
NL	102	91	4x2x2	telep	STP	pol efficacy (europe)
NL	103	92	3x2x2	mail/telep	NIMMO	europe
NL	104	92	4x2x2	tel	NIMMO	satisfaction
NL	105	91	4x2x2	mail	NIMMO	satisfaction
NL	106	92	4x2x2	mail	NIMMO	satisfaction
NL	107	92	4x2x2	mail/telep	NIMMO/STP	satisfaction
NL	108	89	4x3	telep	NIPO	satisfaction
NL	109	91	4x2x2	telep	STP	satisfaction
NL	110	91	3x2x2	telep	STP	satisfaction
NL	111	92	3x2x2	mail/telep	STP	values
NL	112	91	3x2x2	telep	STP	values:comfort/self /respect/status
NL	113	91	3x2x2	telep	STP	values:family/ambition/Independence
NL	114	91	3x2x2	telep	STP	values:comfort/self/respect/status
NL	115	91	3x2x2	telep	STP	values:family/ambition/Independence
NL	116	91	3x2x2	telep	STP	values:comfort/self/respect/status
NL	117	91	3x2x2	telep	STP	values:family/ambition/Independence
NL	118	91	3x2x2	telep	STP	values:comfort/self /respect/status
NL	119	91	3x2x2	telep	STP	values:family/ambition/Independence
NL	120	91	3x2x2	telep	STP	seriousness of crimes
NL	124	91	3x2x2	telep	STP	seriousness of crimes
NL	121	91	3x2x2	telep	STP	seriousness of crimes
NL	122	91	3x2x2	telep	STP	seriousness of crimes
NL	124	91	3x2x2	telep	STP	seriousness of crimes
NL	125	91	3x2x2	telep	STP	seriousness of crimes
NL	–	90	–	telep	STP	EU membership
NL	126	91	4x2x2	telep	STP	EU membership
NL	127	91	3x3	telep	STP	crimes 1,2,3
NL	128	91	3x3	telep	STP	crimes4,5,6
NL	129	91	3x3	telep	STP	crimes 7,8,9
NL	–	88		telep	NIPO	tv/olympische spelen
NL	130	88	3x3	telep	NIPO	vakbeweging
NL	131	88	3x3	telep	NIPO	vakbeweging
NL	132	88	3x3	telep	NIPO	vakbeweging
NL	133	88	3x3	telep	NIPO	vakbeweging
NL	135	92	3x2x2	telep	STP	satisfaction
NL	136	92	3x2x2	telep	STP	satisfaction
NL	137	92	3x2x2	telep	STP	satisfaction
NL	138	92	3x2x2	telep	STP	satisfaction
NL	139	92	3x2x2	telep	STP	work condition
NL	140	92	3x2x2	telep	STP	work condition
NL	141	92	3x2x2	telep	STP	work condition
NL	142	92	3x2x2	telep	STP	work condition
NL	143	92	3x2x2	telep	STP	living condition
NL	144	92	3x2x2	telep	STP	living condition
NL	145	92	3x2x2	telep	STP	living condition
NL	146	92	3x2x2	telep	STP	living condition
NL	–	88	3x3	telep	stp	tv watching
NL	147	88	3x3	telep	stp	eval. tv programs
NL	148	88	3x3	telep	stp	use of the tv
NL	149	88	3x3	telep	stp	reading
NL	150	88	3x3	telep	stp	eval. policies
NL	151	88	3x3	telep	stp	estimate ages
NL	152	88	3x3	telep	stp	political participation

Country	number	year	design	datacollection	organisation	topic
NL	153	88	3x3	telep	stp	estimation of incomes
NL	154	96	4x2x2	telep	stp	trust
NL	155	96	4x2x2	telep	stp	f-scale
NL	156	96	3x2x2	telep	stp	threat
NL	157	96	4x2x2	telep	stp	outgroup
NL	158	96	4x2x2	telep	stp	ingroup
NL	159	96	4x2x2	telep	stp	trust
NL	–	96		telep	stp	ethno/wave 2
NL	–	96		telep	stp	ethno/wave 3
NL	–	98	sbmt	telephone	nimmo	voting
Belg	801	89	5x3	ftf	KUL	satisfaction
Belg	802	97	3x3	ftf/mail	KUL	threat
Belg	803	97	3x3	ftf/mail	KUL	outgroup
Belg	804	97	4x3	ftf/mail	KUL	ingroup
Austria	1	92	4x3	ftf	IFES	party pol
Austria	2	92	4x3	ftf	IFES	econ. expectations
Austria	3	92	4x3	ftf	IFES	postmaterialism
Austria	–	92	4x3	ftf	IFES	pschy problems
Austria	4	92	4x3	ftf	IFES	social control
Austria	5	92	4x4	ftf	IFES	party pol
Austria	6	92	4x3	ftf	IFES	social control
Austria	7	92	4x3	ftf	IFES	EU evaluation
Austria	8	92	3x3	ftf	IFES	life satisfaction
Austria	9	92	3x3	ftf	IFES	political parties
Austria	10	92	4x3	ftf	IFES	conf in institutions
USA	1	79	4x3	ftf	ISR	finances,business,Health,news (l year USA
USA	4	79	4x3	ftf	ISR	same as 2
USA	5	81	3x3	ftf	ISR	finance, business, health,Last year USA
USA	7	81	4x3	ftf	ISR	satisfaction life etc
USA	8	86	2x2x3	ftf	ISR	health/income/3 m
USA	9	86	3x2x2	ftf	ISR	savings/transport/safety
USA	10	86	3x2x3	ftf	ISR	restless/depressed/relaxed
USA	11	86	3x2x3	ftf	ISR	exited/restless/energy
USA	12	86	4x2x2	ftf	ISR	health/income
USA	13	86	5x2x2	ftf	ISR	health/house/income/Friends/life in general

Designs are described by a*b*c where a = number of traits, b = number of methods, c = number of waves
Data collection modes are: mail, telep = telepanel (early version of a Web survey), tel = telephone, ftf = face to face
Research organisations are:
Nimmo = a previous research organization of the University of Amsterdam,
STP = a former independent research organization using the telepanel method (now Centre data of the University of Tilburg),
NIPO = a commercial research organization in the Netherlands,
KUL = Catholic University of Leuven (Belgium),
IFES = a commercial research organization in Austria,
ISR = the Institute for Survey Research of the University of Michigan (USA).