

RECSM Working Paper Number 24

2011

The development of the program SQP 2.0 for the prediction of the quality of survey questions

Willem E. Saris

Daniel Oberski

Melanie Revilla

Diana Zavala

Laur Lilleoja

Irmtraud Gallhofer

Tom Gruner

RECSM / UPF

Barcelona

Table of Contents

	Page	
Preface	5	
Chapter 1	Summary of earlier MTMM studies with respect to characteristics of survey questions which influence the quality of single questions	7
	Willem Saris and Irmtraud Gallhofer	
Chapter 2	The SB-MTMM approach developed for the ESS	29
	Willem Saris, Albert Satorra and Germa Coenders	
Chapter 3	The experiments done in the ESS rounds 1-3	39
	Willem Saris, Irmtraud Gallhofer, Diana Zavala and Melanie Revilla	
Chapter 4	The problems and solutions of the analysis of the MTMM experiments	49
	Melanie Revilla and Willem Saris	
Chapter 5	An overview of the quality of the questions	63
	Diana Zavala, Melanie Revilla, Laur Lilleoja and Willem Saris	
Chapter 6	The prediction procedure the quality of the questions based on the present data base of questions	71
	Daniel Oberski, Thomas Gruner and Willem Saris	
Chapter 7	The program SQP version 2 for prediction of quality of questions and its applications	89
	Daniel Oberski, Thomas Gruner and Willem Saris	
Chapter 8	Conclusions and future developments	109
	Willem Saris	
References		113
Appendix	ESS Questions involved in MTMM experiments	117

Preface

Data from survey research contain both random and systematic errors, which are attributable to a range of factors. In attitude surveys, for instance, random error is a consequence of mistakes made by the respondent, interviewer and others in recording the answers. Systematic errors in contrast can arise from ‘faulty’ questions or different reactions of respondents to the chosen methods, thus generating biased answers. In a comparative context, measuring and correcting for errors is exacerbated by the fact that the size of these different error components may vary cross-nationally, resulting in reduced comparability of findings.

The aim of the here reported part of the Joint Research Action (JRA3), developed in the context of the ESS Infrastructure research, is to estimate the size of these different error components and to propose correction procedures so that a higher degree of equivalence can be achieved across data from different countries. Not all aspects of data quality are easy to measure or evaluate. Among the most widely used quality criteria are reliability, validity, extent of item non-response, relative bias and response effects, misunderstanding of questions, and problems in the interaction between interviewer and respondent. A large body of research has been undertaken into the sorts of question which are particularly error-prone in relation to one or more of these criteria, several of which have tested alternative formats and wordings by means of ‘split ballot experiments’ (Schuman & Presser 1981; Krosnick & Fabrigar, forthcoming). Meanwhile, non-experimental studies have investigated the effect of question characteristics on item non-response and bias (Molenaar 1986), and longitudinal studies (with test-retest designs) have evaluated the effects of question design on the reliability of responses (Alwin & Krosnick 1991). ‘Multi-Trait Multi-Method’ (MTMM) studies have in turn evaluated the effects of question design on reliability and validity (Andrews 1984; Költringer, 1995; Scherpenzeel 1995; Scherpenzeel & Saris 1997).

Most MTMM studies have concentrated on the effect of one factor on the distribution of the variable of interest, but a few have employed meta-analysis of MTMM studies to determine the effects of alternative design choices during the development of questions on reliability and validity (Andrews, 1984, Cote and Buckley (1987) and Lance, Dawson, Birkelbach and Hoffman (2020) Költringer, 1995; Scherpenzeel 1995; Scherpenzeel & Saris 1997). Recent meta-analysis covering all available MTMM experiments directed at the quality of single questions (Saris and Gallhofer, 2007) has been used to develop a program for predicting the quality of survey questions, the Survey Quality Predictor (SQP). Using this program (Oberki et al 2004), the question designer codes the choices they have made in developing the survey item, and the program employs these codes to estimate the reliability, validity and ‘total quality’ of that item. This approach has been applied during the questionnaire design process of each Round of the ESS.

The major added complication of cross-cultural surveys is that an estimate of the reliability (random error) and validity (systematic error) of questions is required for each different language. Otherwise the results cannot properly be compared. Indeed, we have tentatively begun such an evaluation programme in the second Round of the ESS, which will provide data on the quality of survey questions in more than 20 countries and languages. We proposed to develop this work as part of JRA3, extending the SQP program (including the necessary databases) to predict data quality in different languages and to include questions that have not so far been studied. An important aim is to develop procedures that improve the comparability of results from different countries. Such a program is still in development, but the groundwork has been done. Its further development will be invaluable not just for the ESS but also for other cross-national surveys in Europe and beyond.

Before the start of the ESS, 87 MTMM experiments in three languages had ever been carried out (Corten et al, 2003). The 300 ESS experiments (around 16 in each of around 25 countries) have now added considerable weight to this work. This work was done by the research group of Willem Saris at ESADE.

In order to estimate the correction factors for measurement errors, we had to conduct a meta-analysis of the findings of the experiments and apply it to ESS data from all participating countries, together with data on question characteristics. Only in this way will we generate a suitable formula for predicting the quality of questions. The analytic work of this task is carried out by research of the Research and Expertise Centre for Survey Methodology (RECSM) at the Universitat Pompeu Fabra.

This report will discuss the following topics. In chapter 1 we discuss the characteristics of questions which have been found in the past to have an effect on the quality of the questions. Chapter 2 introduces the adjustment of the MTMM design for the ESS. In chapter 3 we will describe the experiments which have been done in the ESS rounds 1-3 and indicate which characteristics have been varied in these experiments by purpose and which have been different across countries for other reasons. In chapter 4 we will discuss the problems we encountered in the estimation procedures of the ESS MTMM experiments and we will discuss the solution we have developed for these problems. Chapter 5 discusses the results with respect to the quality of the collected questions in the ESS experiments across the different countries. In Chapter 6 we discuss the prediction procedure with respect to the quality of the questions implemented in the new version of the SQP program. In chapter 7 the program SQP version 2 is introduced and illustrated. In the last chapter we will draw some conclusions from the obtained results and indicate what the next steps should be for future research in this context.

Finally we would like to thank all people who have made this work possible. First of all, we would like to thank the European Commission that has subsidized this research. Secondly, we would like to thank the colleagues of the ESS which have had a lot of patience with us to produce the results reported here. Thirdly we thank the National coordinators in all the countries which have put a lot of efforts in to collect the extra data for our research. We are also very grateful to all respondents performing the extra tasks we have asked from them. A group of people that did important work for us was the group of coders of all the questions in the different languages. Finally we would like to thank ESADE and the UPF for the facilities they have provided us to do this work. We are very grateful for all the cooperation we have received over the last 4 years by all the people mentioned here and the ones we did not mention by mistake.

Barcelona, 29 December 2011

Willem Saris

Chapter 1

Summary of earlier studies with respect to characteristics of survey questions which influence the quality of single questions¹

Willem E. Saris

Irmtraud Gallhofer

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:

- *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;*
- *assembled a large set of studies that use Multi-Trait Multi-method (MTMM) experiments to estimate the reliability and validity of questions.*
- *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.

1.1 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, neighbourhood, 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),

¹ The extended report on which this chapter is based can be found in Saris W.E. and I.N.Gallhofer (2007) Design, evaluation and analysis of questionnaires for survey research. New York, Wiley.

Forsyth et al. (1992), Esposito et al. (1991), (1997), Sudman et al. (1996), Van der Zouwen (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 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 question design on the reliability, validity and method effects of survey questions. That is the research that has been done by Saris and Gallhofer (2007) as will also be reported below.

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; Koltringer (1995) in Austria, Scherpenzeel and Saris (1997) in the Netherlands and Billiet and Waeghe (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). 1. 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 characteristics. 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:

- a) Simple or direct requests
- b) Indirect requests such as Agree/disagree requests
- c) 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*:

- a) a category scale with 2 categories (yes/no)
- b) a category scale with more categories
- c) frequency
- d) magnitude estimation where the size of the number indicates the opinion
- e) line drawing scale where the length of the line indicates the opinion
- f) 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 we mention 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 itself. 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:

- a) Presence of emphatic introduction
- b) Presence of a *motivation*
- c) Presence of *information regarding the content*

- d) Presence of *information regarding a definition*
- e) Presence of an *instruction to the respondent*
- f) 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:

- a) The *number of interrogative sentences*
- b) The number of *subordinate clauses*
- c) The *total number of words* in the survey item
- d) The *average number of words* of the sentences
- e) The *average number of syllables* per word
- f) The *total number of nouns* in the request text
- g) The *percentage of abstract nouns* relative to the total number of nouns

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:

- a) *Computer assisted* data collection of not
- b) *Interviewers administered* or not
- c) *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).

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.

1.2 Estimation of the reliability, validity and quality

Using this MTMM design and structural equation modelling techniques, the reliability and validity coefficients were obtained for each question, estimating the true score model developed by Saris and Andrews (1991). This is specified as follows:

$$Y_{ij} = r_{ij}T_{ij} + e_{ij} \quad \text{for all } i,j \quad (1.1)$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j \quad \text{for all } i,j \quad (1.2)$$

Where, F_i is the i^{th} trait, M_j the variation in scores due to the j^{th} method, and for the i^{th} trait and j^{th} method, Y_{ij} is the observed variable, r_{ij} is the reliability coefficient, T_{ij} is the true score or systematic component of the response, e_{ij} is the random error associated with the measurement of Y_{ij} , v_{ij} is the validity coefficient, and m_{ij} is the method effect coefficient. The model is completed by some assumptions: the trait factors are correlated with each other; the random errors are *not* correlated with each other, nor with the independent variables in the different equations; the method factors are *not* correlated with each other, nor with the trait factors; the method effects for a specific method M_{j*} are equal for the different traits T_{ij*} (for all i); the method effects for a specific method M_{j*} are equal across the split-ballot groups; as are the correlations between the traits, and the random errors. These assumptions are the ones we start with but when testing the model, if some of them do not hold, they can be realised.

The quality of a measure can be derived from this model. It is the product of the reliability (square of the reliability coefficient) and the validity (square of the validity coefficient), so: $q_{ij}^2 = r_{ij}^2 \cdot v_{ij}^2$. It corresponds to the strength of the relationship between the variable of interest F_i and the observed answer Y_{ij} expressed for the j^{th} method.

1.3 Estimation of the effect of the characteristic of the questions on their quality

In order to integrate the 87 MTMM studies that were carried out in three languages they were reanalyzed, and the survey items were coded according to characteristics listed above. Scherpenzeel (1995) has indicated that without this recoding, the results of the different studies were incommensurable. Therefore, all survey items were coded in exactly the same manner. The code-book is available at the SQP website². The data of the different studies was pooled and an analysis conducted over all available survey items adding a variable “language” to it in order to take into account any effect due to differences in languages.³

Normally, multiple-classification analysis or MCA is applied (Andrews 1984; Scherpenzeel 1995; Költringer 1995) to meta-analysis, but the number of variables that

² Details of the codebook can be found at www.sqp.nl.

³ The analysis shows that the effect of language is additive, meaning that language affects only the absolute level of the quality indicators. If this were true for all languages, it would mean that comparisons of choices could be made for all languages and only the absolute level of the quality criteria could be incorrect.

need to be introduced in the analysis make it impossible. A solution is (dummy) regression. The following equation presents the approach used:

$$C = a + b_{11}D_{11} + b_{21}D_{21} + \dots + b_{12}D_{12} + b_{22}D_{22} + \dots + b_3N_{\text{cat}} + \dots + e \quad (1.3)$$

In this equation, C represents the score on a quality criterion, which is either the reliability or validity coefficient. The variables D_{ij} represent the dummy variables for the j^{th} nominal variable. All dummy variables have a zero value unless a specific characteristic applies to the particular question. For all dummy variables, one category is used as the reference category which has received the value “zero” on all dummy variables within that set. Continuous variables, like the number of categories (N_{cat}), were not categorized, except when it was necessary to take nonlinear relationships into account. The intercept is the reliability or validity of the instruments if all variables have a score of zero. Table 1.1 shows the results of the meta-analysis over the available 1023 survey items. Table 1.1 indicates the effects of different survey design choices on the quality criteria of validity and reliability. The table contains also the standard errors (se) of these coefficients and their significance level (sign). The method effects were not indicated because they can be derived from the validity coefficients.

Each coefficient indicates the effect of a 1 point increase on each indicated characteristic while keeping all other characteristics constant. For example, all questions concerning “consumption,” “leisure,” “family,” “personal relations” and “race” are coded as zero on all domain variables that can be seen as the reference category. For these questions the effect on reliability and validity is zero. Questions concerning other issues are coded further into several categories. If a question concerns “national politics” it belongs to the first domain category ($D_{11}=1$ for this category, while all other domain variables $D_{i1}=0$) and its effect on reliability and validity will be positive, .0528 and .0447, respectively as can be seen from the table. Note that all the effects in the table are multiplied by 1000. If a question concerns “life in general” then the fifth category applies ($D_{51}=1$) and the effects are negative: -.0768 and -.0159, respectively. From these results it also follows that questions concerning national politics have a reliability coefficient of .0528 + .0768 or .1296 higher than the questions about life in general. This interpretation holds for all characteristics with a dummy coding such as “concepts,” “time reference,” and so on.

Table 1.1: Results of the Meta-Analysis

Variables	Number of measures	Effect on reliability			Effect on validity		
		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 condition/background (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 subjective variables (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	136	0.0	—	—	0.0	—	—
Concepts							
Evaluative belief (0–1)	96	6.1	14.0	.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
Facts:background (18)							
Behavior (9) (0–1)	27	-126.2	21.8	.000	-150.5	19.2	.000
Other simple concepts	578	0.0	—	—	0.0	—	—
Complex concepts	1023	-72.3	17.4	.000	-47.2	15.2	.002
Associated characteristics							
Social desirability: no/ a bit/much (0–2)	1023	2.3	6.2	.709	8.0	5.3	.137
Centrality: very central to not central (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
Present (0–1)	940	0.0	—	—	0.0	—	—
Formulation of Requests: basic choice							
Indirect question							
Agree/disagree (0–1)	167	4.0	10.9	.713	41.6	9.5	.000
Other types: direct request (190), more steps ⁱ (22)	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

Table 1.1 (continued)

Variables	Number of measures	Effect on reliability			Effect on validity		
		Effect	se	sign	effect	se	sign
Formulation of the request : 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 choice							
Yes/no (0—1)	3	-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
Kind of label: short, sentence (0—1)	35	-47.5	16.0	.003	-9.1	13.7	.506
Don't know: present, registered, not present (1—3)	1023	-6.7	4.8	.165	-1.9	4.1	.647
Neutral: present, registered, not present (1—3)	1023	12.6	4.6	.007	8.4	4.0	.038
Range:							
Theoretical range and scale unipolar							
Theoretical range and scale bipolar;							
Theoretical range bipolar but scale unipolar (1—3)	1023	-15.1	9.6	.116	9.2	8.5	.277
Correspondence:							
high—low (1—3)	1023	-16.8	7.5	.025	1.1	6.5	.867
Symmetric labels (0—1)	195	25.5	11.8	.031	22.3	10.4	.033
First answer category: negative, positive (1—2)	358	-7.5	8.7	.387	14.7	7.6	.052
Fixed reference points (0— 3)	1023	14.7	4.3	.001	21.4	3.7	.000
Number of Categories (0—11)	1023	13.5	2.1	.000	-1.9	1.8	.298
Number of frequencies (0—5000)	1023	-.068	.009	.000	-.065	.008	.000

Table 1.1 (continued)

Variables	Number of measures	Effect on reliability			Effect on validity		
		Effect	se	sign	effect	se	sign
Survey item specification: basic choices							
Question present (0–1)	841	27.2	15.2	.074	11.5	13.1	.379
Instruction present (0–1)	103	-43.7	15.4	.005	-4.2	13.3	.753
No question or instruction	79	0.0	—	—	0.0	—	—
Respondent's instruction (0–1)	492	-12.7	7.3	.083	-14.9	6.2	.017
Interviewer's instruction (0–1)	119	-.068	10.5	.995	5.7	9.0	.524
Extra motivation/ information or 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
Survey item specification: other choices							
Complexity of the introduction							
Question in the intro (0–1)	62	-44.6	16.3	.006	-21.3	14.1	.132
Number of subordinate clauses >0	129	29.3	9.8	.003	7.6	8.6	.377
Number of words per sentence >0	510	-1.3	.867	.134	1.4	.75	.063
Mean of words per sentence >0	510	.064	1.1	.954	-.373	.9	.699
Complexity of request							
Number of sentences (0–n)	192	12.7	9.8	.199	-8.3	8.6	.335
Number of subordinate clauses (0–n)	746	13.6	6.8	.048	-17.7	5.9	.003
Number of words (1–51)	1023	.809	.749	.280	-1.3	.644	.041
Mean of words per sentence (1–47)	1023	-2.2	.926	.014	1.1	.807	.161
Number of syllables per word (1–4)	1023	-32.5	9.6	.001	-10.4	8.2	.207
Number of abstract nouns on the total number of nouns (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-administered (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

Table 1.1 (continued)

Variables	Number of measures	Effect on reliability			Effect on validity		
		Effect	se	sign	effect	se	sign
Position in questionnaire							
In battery (0–1)	225	-10.3	12.3	.403	28.9	10.7	.007
position of question	1023	.304	.064	.000			
position 25 (1–25)	396				1.5	.402	.000
position 100 (26–100)	458				.420	.137	.002
position 200 (101–200)	129				.267	.062	.000
position 300(>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							
Percentage of low educated (3–54)	993	-.911	.596	.127	1.1	.511	.027
Percentage of high age (1–49)	1023	-.410	.560	.464	-.753	.488	.123
Percentage of males (39–72)	1023	-.030	.690	.966	.405	.596	.497
MTMM design							
Design: one or 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
Number of traits (1–10)	1023	-.370	2.0	.855	-1.7	1.7	.320
Number of methods (1–4)	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		
Correction for single item distance							
		-42.3			-62.25		
Starting point for single item							
		782.9			977.15		

Other characteristics using at minimum an ordinal scale are treated as metric. For example, “centrality” is coded in five categories from “very central” to “not central at all.” In this case an increase of one point gives an effect of $-.0172$ on reliability and the difference between a very central or salient item and a not at all central item is $5 \times -.0172 = -.0875$.

Furthermore, there are real numeric characteristics like the “number of interrogative sentences,” “the number of Words.” In that case, the effect is an increase of one unit per word or interrogative sentence.

A special case in this category is the variable “position” because it turns out that while the effect of “position” on reliability is linear, for validity it is non-linear. To describe the latter relationship, the “position” variable is categorized, and the effects are determined within their respective categories.

Another exception is the “number of categories in the scale.” For this variable we have specified an interaction term, because the effects were different for categorical questions versus frequency measures. Therefore, depending on whether the question is a categorical or a frequency question, a different variable is specified to estimate the effect on the reliability and the validity.

1.4 Results of the meta-analysis

Below we discuss the most important results presented in Table 1.1.

Domain, concept, and associated characteristics

- The research design determines the domain, concepts, and associated characteristics. Nevertheless, there are significant differences in reliability and validity for items from different domains, measuring different concepts or with different associated characteristics.
- Behavioral survey items tended to have a more negative effect than attitudinal questions, especially items concerning the “frequency of behavior.” Although only a few items of this type were analyzed; therefore, the standard error of the effect is relatively large.
- Complex items should be avoided where ever possible, given their negative effect.
- It appears that reporting about the past is more reliable than reporting about the future or the present.

Formulation of the requests

In formulating the requests, the researcher has more freedom of design. We found that

- Indirect requests such as agree/disagree options perform similarly to direct requests on reliability and a bit better with respect to validity.
- The use of statements or stimuli has a small negative effect on reliability and validity; therefore, it is better to avoid them.
- On the other hand, the reliability improves with gradation requests, although they have a small negative effect on validity.
- A lack of balance in the formulation of the request has a significant negative effect on validity.
- Emphasizing subjective opinion has a significant negative effect on validity.

Response scale

- Use of response scales with gradation in the form of frequency, magnitude estimation or line production and the stepwise procedure has a positive effect on reliability, but is often associated with strong method effects such as rounding off errors, which reduces validity.

- Line production and stepwise procedures incur a relatively smaller method effect.
- Reliability is improved when labels instead of complete sentences are used.
- Not providing a neutral middle category improves both reliability and validity significantly.
- The use of fixed reference points has a quite large positive effect on reliability and validity. This approach is especially recommended for long scales with 7 or more categories.
- The effect of range is rather limited, which may be due to the selected categories.
- Making the numbers correspond with the labels has a significant positive effect on reliability.
- Symmetry within response categories has some positive effect on reliability and validity.
- The number of categories has an opposite effect for category and frequency scales. In the case of a category scale (2-points – 15-points and more steps procedures), reliability can be increased by more than 100 points by going from a 2-point to an 11- point scale.
- In the case of a frequency scale, reliability and validity experience a large decrease if the range of the scale is too wide (i.e., if very high frequencies are possible).
- For magnitude estimation and line production, this effect does not apply. The number of categories seems to be integrated in the effect of the method itself.

Specification of the survey item as a whole

- The first item is more reliable if a normal request is asked and less reliable if an instruction is used, in comparison to subsequent items in a battery.
- Items in a battery without a request for an answer (almost all items except the first one) are better than items with an instruction but worse than items with a normal request for an answer. This may be due to the complexity of the procedure, which requires extra instruction, and not because of the effect of the instruction. The same may hold true for our discussion of the next effect.
- Respondents' instructions have a significant negative effect on reliability and validity. The item may be so difficult that it requires an explanation, and therefore the effect may be caused by the item and not the instruction.
- Interviewer instructions, extra motivational remarks, definitions, and an introduction seem to have no significant effect on reliability or validity.
- Formulating general questions in the introduction, which are followed by the real request, should be avoided because they have a negative effect on both reliability and validity.
- On the other hand, they seem to have a positive effect on reliability if more explanation is given in subordinate clauses of the introduction.
- This effect holds true for the request itself, having also a positive effect on validity.
- However, there is a limit to the number of words in the request, as if it becomes too long, it has a negative effect on validity.

The two indices for complexity of requests, the number of words per sentence (sentence length), and the number of syllables per word (word length), have a significant negative effect on reliability⁴.

⁴ The variables “syllables/word” and “proportion of abstract words” have been collected for the introduction and the question itself; however, in the introduction these variables correlated very highly with each other and with the variable “intro” and it was decided that these variables cannot be used together with the variable “introduction.”

Mode of data collection

The mode of data collection can be analyzed by each basic method or by a general description.

- The CAI is as reliable as the non-CAI; however, it is less valid.
- A much stronger negative effect can be observed from interviewer-administered questionnaires than the other methods.
- Oral questionnaires have a small but significant positive effect on the validity.

Position in the questionnaire

- The effect of the position of a request within a questionnaire is rather different for either reliability or validity.
- It seems that respondents continuously learn about how to fill in the questionnaire, causing the reliability of the response to increase linearly with its position. Over the range studied, the effect can be more than 100 points.
- On the other hand, the effect on validity is .037 point for the first 25 requests, followed by an effect of .031 for the 25th request until the 100th, and for the 100th — 200th this effect is .026 while after the 200th request there is no further significant increase.

Basic choices for which correction is necessary

Some choices cannot be explicitly made such as language or the characteristics of a population. These choices can nevertheless have an influence on the quality criteria. In addition, the methodological experiments that form the basis for this meta-analysis also have some influence that has to be estimated and controlled for when the other effects are estimated.

- Unfortunately, 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 education level has a significant effect on the validity of responses. Samples with a high number of lower educated people may score in validity .050 lower than samples with few poorly educated people.
- The MTMM design used also has a significant effect on the data quality. As the distance in time between the items for the same trait increases, the reliability declines. For the largest distance found the reliability decreased by .042.
- The distance between the traits has an even larger effect on validity; for the largest distance found, the validity decreased by .062.

In a normal survey MTMM experiments are not present and one measure is available for each trait. Therefore, for predicting the quality of survey items, a correction for the fact that a survey item appears only once within the questionnaire has to be made. This correction is specified at the bottom of Table 12.1. We have corrected for the distance of the “previous measure of the same trait,” where the intercept is adjusted by subtracting .0423 for reliability and .06225 for validity.

1.5 Special topics

In this section, we will focus on the effects of certain choices that warrant further detail.

The choice of direct requests or agree/disagree requests

Agree/disagree requests score better on validity (.041) than do direct requests. However, agree/disagree requests are most commonly used in batteries, and we have found that compared with items presented later in a battery (with no question or instruction), a direct question is more reliable (.0272) while an instruction is less reliable (-.0437). Hence a difference in reliability between the two procedures of .0709 is compensated by .041 in validity. This difference is in favor of direct questions. Differences in reliability between these two types of questions also have been found in other studies (Saris and Galhofer 2006). However, it is somewhat surprising to find that agree/disagree procedures score higher on validity. It is anticipated that acquiescence would lead to the opposite effect (Krosnick and Fabrigar 1997); therefore this issue needs to be investigated further.

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 on quality. Cox (1980), and Krosnick and Fabrigar (1997) defend the position that one should not use more than seven categories while Andrews (1984), Költringer (1995), and Alwin (1997) argue to the contrary that more categories lead to better results. Our analysis suggests that frequency scales, magnitude scales, and line scales are generally more reliable than category scales. However, frequency and magnitude scales especially pay the price for reliability by sacrificing validity. This phenomenon has two reasons. The first is that people round off their numeric values in a specific way. Some use numbers divisible by 25, others are more precise and use numbers divisible by 10, and others use even numbers divisible by 5. Such differences in behavior cause method effects. The other possible explanation is what Saris (1988) has called “variation in response functions.” When respondents are allowed to specify their own response scales this will lead to method effects and as a consequence to lower validity coefficients. The solution suggested by Saris (1988) is confirmed by this analysis because better validity and reliability is obtained if the scales are made comparable through use of fixed reference points (see Chapter 7).

The reliability of category scales can also be improved by using more categories (so far up to 11 categories were studied) without decreasing validity. An alternative is to use a two-step procedure that improves both reliability and validity. Category scales can also be improved using labels for most categories as long as they are not in full sentence format. In summary, this analysis strongly suggests to use as many categories as possible in a category scale (more than seven) that are short and clearly labeled. Line production or magnitude estimation with fixed reference points are the optimal choice in most cases and should be used whenever possible.

Effects of the mode of data collection

On the basis of the choices specified in Table 1.1, the commonly used data collection methods can be constructed by combining different characteristics. Their results and the effects of their combinations on reliability and validity are presented in Table 1.2.

Table 1.2: Effects of modes of data collection on data quality, based on the combined effect of computer-assisted data collection and interviewer-administered data collection

	CAI	Not CAI	
Interviewer-administered		CATI/CAPI	PAPI/TEL
Reliability coefficient		-.0538	-.050
Validity coefficient		-.1423	-.104
Self-administered CASI		Mail	
Reliability coefficient		-.0038	.000
Validity coefficient		-.0383	.000

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

- a) Mail
- b) CASI
- c) PAPI/Telephone
- d) CATI/CAPI

The differences between Mail and CASI are minimal, on the other hand, differences between these two and the PAPI/Telephone 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 nonresponse and item nonresponse. 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 tradeoff between quality, with respect to reliability and validity, and item nonresponse has to be made.

1.6 Conclusions, limitations, and the future

Our results show that within and between questionnaires there is a wide variation in reliability and validity. In particular the following choices have a large effect on reliability and/or validity:

- The use of direct questions has a large positive effect on reliability and a smaller negative effect on validity when compared with batteries containing statements.
- The use of gradation has a large positive effect on reliability and a smaller negative effect on validity.
- The use of frequencies or magnitude estimation has a large positive effect on reliability and an almost equally large negative effect on validity.
- The use of lines as response modality has a large positive effect on reliability and a much smaller negative effect on validity.
- The more categories a response scale has, the greater the positive effect on reliability is. However, it also has a much smaller negative effect on validity.
- Allowing for high frequencies has both a large negative effect on reliability and validity.
- The use of interviewers has both a large negative effect on reliability and validity.

This analysis is an intermediate result; so far 87 studies have been reanalyzed with a total of 1023 survey items, which is not enough to evaluate all variables in detail. (The database is a work in progress that will be extended in the future with survey items that are at present underrepresented.) Important limitations to consider are listed below:

- Only the main categories of the domain variable have been taken into account.
- Requests concerning consumption, leisure, family, and immigrants could not be included in the analysis.
- The concepts of norms, rights, and policies have been given too little attention.
- The request types of open-ended requests and WH requests have not yet been studied.
- Mail and Telephone interviews were not sufficiently available to be analyzed separately.
- There is an overrepresentation of requests formulated in the Dutch language.
- Only a limited number of interactions and nonlinearities could be introduced.

Nevertheless, taking these limitations into account, the analysis can remarkably explain 47% of the reliability variance and 61% of the validity. In this respect, it is also relevant to refer to the standard errors of the regression coefficients which are relatively small, indicating that the correlations between the variables used in the regression as independent variables are relatively small.

If one considers that all estimates of the quality criteria contain errors while in the coding of the survey item characteristics errors are also made, the high explained variance is very promising.

The authors of this meta analysis concluded “This does not mean that we are satisfied with this result. Certainly, further research is needed, as we have indicated above, but for the moment Table 1.1 is the best summary of our knowledge about the effects of the questionnaire design choices on reliability and validity.”

In the next chapter we will indicate how this work was continued using the possibilities provided by the European Social Survey to include MTMM experiments in the biannual rounds of data collection.

Appendix 1: Overview of the experiments used in the analyses in 2001

Country	number	year	design	mode	data collection organization	topic
NL	101	92	3×2x2	Mail/Telep	STP	Seriousness of crimes
NL	102	91	4x2x2	Telep	STP	political 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/Olympic games
NL	130	88	3x3	Telep	NIPO	Trade-unions
NL	131	88	3x3	Telep	NIPO	Trade-unions
NL	132	88	3x3	Telep	NIPO	Trade-unions

Appendix 1 continued

Country	number	year	design	mode	data collection organization	topic
NL	133	88	3x3	Telepanel	NIPO	Trade-unions
NL	135	92	3x2x2	Telepanel	STP	Satisfaction
NL	136	92	3x2x2	Telepanel	STP	Satisfaction
NL	137	92	3x2x2	Telepanel	STP	Satisfaction
NL	138	92	3x2x2	Telepanel	STP	Satisfaction
NL	139	92	3x2x2	Telepanel	STP	Work condition
NL	140	92	3x2x2	Telepanel	STP	Work condition
NL	141	92	3x2x2	Telepanel	STP	Work condition
NL	142	92	3x2x2	Telepanel	STP	Work condition
NL	143	92	3x2x2	Telepanel	STP	Living condition
NL	144	92	3x2x2	Telepanel	STP	Living condition
NL	145	92	3x2x2	Telepanel	STP	Living condition
NL	146	92	3x2x2	Telepanel	STP	Living condition
NL	—	1988	3x3	Telepanel	STP	TV watching
NL	147	1988	3x3	Telepanel	STP	Evaluation TV programs
NL	148	1988	3x3	Telepanel	STP	Use of the tTV
NL	149	1988	3x3	Telepanel	STP	Reading
NL	150	1988	3x3	Telepanel	STP	Evaluation policies
NL	151	1988	3x3	Telepanel	STP	Estimate ages
NL	152	1988	3x3	Telepanel	STP	Political participation
NL	153	1988	3x3	Telepanel	STP	Estimation of income
NL	154	1996	4x2x2	Telepanel	STP	Trust
NL	155	1996	4x2x2	Telepanel	STP	F-scale
NL	156	1996	3x2x2	Telepanel	STP	Threat
NL	157	1996	4x2x2	Telepanel	STP	Outgroup
NL	158	1996	4x2x2	Telepanel	STP	Ingroup
NL	159	1996	4x2x2	Telepanel	STP	Trust
NL	—	1996		Telepanel	STP	Ethno/wave 2
NL	—	1996		Telepanel	STP	Ethno/wave 3
NL	—	1998	sbmt	Telephone	Nimmo	Voting
Belg	801	1989	5x3	Ftf	KUL	Satisfaction
Belg	802	1997	3x3	Ftf/Mail	KUL	Threat
Belg	803	1997	3x3	Ftf/Mail	KUL	Outgroup
Belg	804	1997	4x3	Ftf/Mail	KUL	Ingroup

Appendix 1 continued

Country	number	year	design	mode	data collection organization	topic
Austria	1	92	4x3	Ftf	IFES	Party politics
Austria	2	92	4x3	Ftf	IFES	Economic expectations
Austria	3	92	4x3	Ftf	IFES	Postmaterialism
Austria	—	92	4x3	Ftf	IFES	Psychological problems
Austria	4	92	4x3	Ftf	IFES	Social control
Austria	5	92	4x4	Ftf	IFES	Party politics
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	Confidence in institutions
USA	1	1979	4x3	Ftf	ISR	Finances,Business,
Health,News						
(1 year USA	2	1979	4x3	Ftf	ISR	Finances,Business,
						Health,News
(n year) USA	3	1979	4x3	Ftf	ISR	Same as 1
USA	4	1979	4x3	Ftf	ISR	Same as 2
USA	5	1981	3x3	Ftf	ISR	Finance, Business,
Health, lastyear						
USA	6	1981	3x3	Ftf	ISR	Finance/Business/Health,
next year						
USA	7	1981	4x3	Ftf	ISR	Satisfaction life etc
USA	8	1986	2x2x3	Ftf	ISR	Health/Income
USA	9	1986	3x2x2	Ftf	ISR	Savings/Transport/Safety
USA	10	1986	3x2x3	Ftf	ISR	Restless/Depressed/Relaxed
USA	11	1986	3x2x3	Ftf	ISR	Exited/Restless/Energy
USA	12	1986	4x2x2	Ftf	ISR	Health/Income
USA	13	1986	5x2x2	Ftf	ISR	Health/House/Income/Friends/ Life in general

Chapter 2

The adjustment of the MTMM design for estimation of the quality of questions of the European Social Survey: the split ballot MTMM approach⁵

Willem E. Saris

So far most MTMM experiments were based on the classical design suggested by Campbell and Fiske (1959) of three traits measured with three alternative methods. The problem of this design is that the respondents have to answer similar questions three times. This is a rather heavy response burden that may lead to satisficing and runs the risk of memory effects if the questions for the same traits are not separately long enough in time. In order to avoid these problems the suggestion is made by Saris, Satorra and Coenders (2004) to split the sample at random in several groups and ask each group only twice a question about the same trait. They suggested that using Multiple Group Maximum Likelihood estimation allows in that case the estimation of all parameters of the classical MTMM experiment.

With respect to the European Social Survey it was necessary to take care that all people in the main questionnaire would get the same questions. Therefore the 2-group design has been chosen for the ESS. In that case all respondents get form 1 of the question in the main questionnaire while one group gets form 2 in the supplementary questionnaire and the other groups gets form 3 of the same question in the supplementary questionnaire. This approach was chosen after evaluation whether the necessary estimates could be obtained even though we were aware of the fact that the 3 groups design was more efficient and would lead to less problems with respect to identifications. This chapter discusses the arguments for the choice of the new approach which has been called the Split ballot MTMM design.

2.1 Introduction

Over the last 40 years, many studies have been performed to evaluate the quality of survey questions. Most studies use random assignment of respondents to different question forms to see whether the form of the question makes a difference. These so called “split ballot experiments” have been used by Schuman and Presser (1981) and many others in the social sciences. Molenaar (1986) studied the quality of questions using nonexperimental research. In official statistics, test-retest models have been popular in evaluating questions (Forsman 1989). Heise (1969), Wiley and Wiley (1970), Alwin and Krosnick (1991) and Alwin (2007) used the quasi-simplex model based on panel data to evaluate the quality of questions. The testing of questions in cognitive laboratories has recently received a great deal of attention. As well as all these approaches, an alternative was applied by Frank Andrews (1984) which is called the Multitrait Multimethod or MTMM approach. After the death of Frank Andrews, his work was continued by European researchers (Scherpenzeel 1995, Scherpenzeel and Saris 1997, Coenders and Saris 2000, Corten and Saris, Aalberts and Saris 2002, Saris, Satorra and Coenders (2004), and finally led to a summary of this research in a book by Saris and Gallhofer (2007) which also introduces a computer program (SQP) that can predict the quality of questions before data are collected in the field (Oberski, Kuipers

⁵ This short summary is based on a paper published by W.Saris, A.Satorra and G.Coenders (2004) A new approach for evaluating quality of measurement instruments. *Sociological Methodology*, 3, 311–347.

and Saris 2004). In this paper, we concentrate on the MTMM approach. We will first explain what we mean by quality of a question, and then we will introduce the MTMM design and model. We will illustrate the approach and discuss its advantages and disadvantages.

2.2 Quality criteria for survey measures

The first quality criterion for survey items is *item non-response*. This is an obvious criterion, because missing values have a disrupting effect on the analysis, which can lead to results that are not representative of the population of interest.

A second criterion is *bias*, which is defined as a systematic difference between the real values of the variable of interest and the observed scores corrected for random measurement errors⁶. Real values can be obtained for objective variables and therefore the most preferable method is the one that provides responses corrected for random errors which are closest to the real values. A typical example comes from voting research. Participation in the elections is known after the elections. This result can be compared with the results obtained from survey research performed using various methods. It is a well-known fact that participation is overestimated when standard survey methods are used. A new method that does not overestimate the participation or produces a smaller bias is therefore preferable to the standard procedures.

In the case of subjective variables, in which the real values are not available, it is only possible to study the various distributions of responses for different methods. If differences between two methods are observed, at least one method is biased; however, it is also possible that both are biased.

These two criteria have received a lot of attention in split-ballot experiments. See Schuman and Presser (1981) for a summary. Molenaar (1986) studied the same criteria while focusing on non-experimental research (1986). In short, these criteria describe the observed differences of nonresponse and differences of response distributions.

Other quality criteria which have also been discussed at length are *reliability*, *validity*, and the *method effect*. Reliability is the complement of random errors and validity is the complement of systematic errors. Both criteria have been discussed extensively in psychology and other social sciences as criteria for the quality of measures. There are many different definitions of these criteria. Below we give the definitions which have been used in the MTMM literature for some considerable time, starting with a paper by Saris and Andrews (1991)

In order to do so we present a measurement model for two variables of interest, such as “satisfaction with the government” and “satisfaction with the economy.” The measurement model for the two variables is presented in Figure 1. In this model it is assumed that

- f_i is the trait factor i of interest measured by a direct question.
- y_{ij} is the observed variable (variable or trait i measured by method j).
- t_{ij} is the “true score” of the response variable y_{ij} .
- M_j is the method factor that represents a specific reaction of respondents to a method
- and therefore generates a systematic error.
- e_{ij} is the random measurement error term for y_{ij} .

$$\rho(f_1, f_2)$$

⁶ This simple definition serves the purpose of this text. However, a precise definition can be found in Groves (1989).

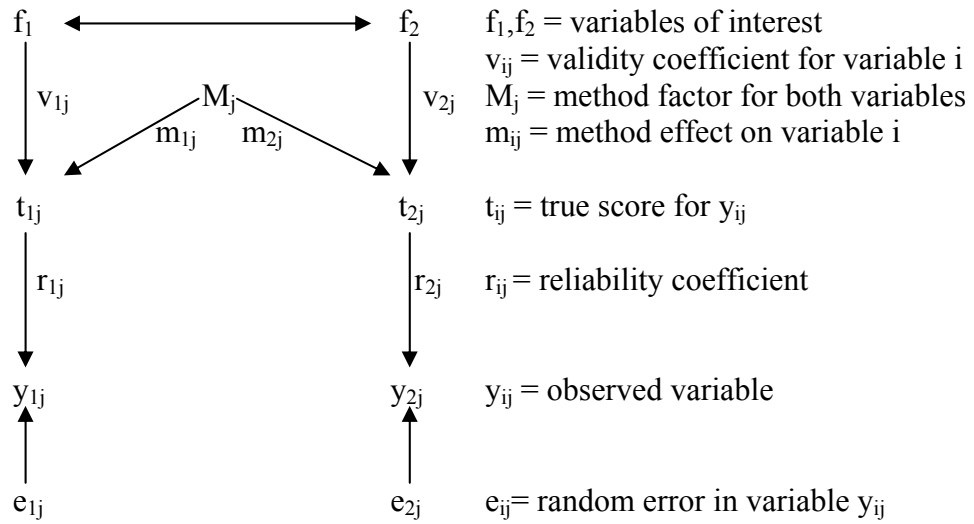


Figure 2.1: The measurement model for two traits measured using the same method.

The r_{ij} coefficients represent the standardized effects of the true scores on the observed scores. This effect is smaller if the random errors are larger. This coefficient is called the *reliability coefficient*. *Reliability* is defined as the strength of the relationship between the observed response (y_{ij}) and the true score (t_{ij}), that is r_{ij}^2 .

The v_{ij} coefficients represent the standardized effects of the variables of interest on the true scores for the variables that are in fact measured. This coefficient is therefore called the *validity coefficient*. *Validity* is defined as the strength of the relationship between the variable of interest (f_i) and the true score (t_{ij}), that is v_{ij}^2 .

The m_{ij} coefficients represent the standardized effects of the method factor on the true scores, called the *method effect*. An increase in the method effect results in a decrease in validity and vice versa. It can be shown that for this model $m_{ij}^2 = 1 - v_{ij}^2$, and therefore the method effect is equal to the invalidity due to the method used. The *systematic method effect* is the strength of the relationship between the method factor (M_j) and the true score (t_{ij}) denoted by m_{ij}^2 .

The *total quality of a measure* is defined as the strength of the relationship between the observed variable and the variable of interest, that is $(r_{ij}v_{ij})^2$.

The *effect of the method on the correlations* is equal to $r_{1j}m_{1j}m_{2j}r_{2j}$.

The reason for using these definitions as quality criteria becomes evident after examining the effect of the characteristics of the measurement model on the correlations between the observed variables.

It can be shown that the correlation between the observed variables $\rho(y_{1j}, y_{2j})$ is equal to the combined effect of the variables that we want to measure (f_1 and f_2) plus the spurious correlation due to the method factor as demonstrated in formula (1):

$$\rho(y_{1j}, y_{2j}) = r_{1j}v_{1j} \rho(f_1, f_2)v_{2j}r_{2j} + r_{1j}m_{1j}m_{2j}r_{2j} \quad (2.1)$$

Note that r_{ij} and v_{ij} , which are always less than 1, will decrease the correlation (see first term) while the effects of the method, if they are not zero, can generate an increase in the correlation (see second term).

If there are only two observed variables, the quality criteria and the correlation between the variables of interest cannot be estimated. A design for data collection is therefore needed that provides more information so that the parameters of the model can be identified.

2.3 The classical MTMM design and model

Campbell and Fiske (1959) suggested using multiple traits and multiple methods (MTMM). The classic MTMM approach recommends using at least three traits that are measured with three different methods, leading to nine different observed variables. An example of such a design is presented in Table 1.

Table 2.1. The classic MTMM design used in the ESS pilot study

The three traits were presented by the following three questions:

1. On the whole, how satisfied are you with the present state of the economy in Britain?
2. Now think about the national government. How satisfied are you with the way it is doing its job?
3. And on the whole, how satisfied are you with the way democracy works in Britain?

The three methods are specified by the following response scales:

(1) *Very satisfied*; (2) *Fairly satisfied*; (3) *Fairly dissatisfied*; (4) *Very dissatisfied*

<i>Very dissatisfied</i>	0	1	2	3	4	5	6	7	8	9	10	<i>Very satisfied</i>
--------------------------	---	---	---	---	---	---	---	---	---	---	----	-----------------------

(1) *Not at all satisfied*; (2) *Satisfied*; (3) *Rather satisfied*; (4) *Very satisfied*

Using this MTMM design, data for nine variables are obtained and a correlation matrix of 9×9 is obtained from those data. The model formulated to estimate the reliability, validity, and method effects is an extension of the model presented in Figure 1. Figure 2 illustrates the relationships between the true scores and the general factors of interest. Figure 2 shows that each trait (f_i) is measured in three ways. It is assumed that the traits are correlated but that the method factors (M_1, M_2, M_3) are not correlated because the reactions will be different for different methods. To reduce the complexity of the figure, no indication is given that for each true score there is an observed response variable that is affected by the true score and a random error, as was previously introduced in the model in Figure 1. However, these relationships, although not made explicit, are implied.

It is normally assumed that the correlations between the factors and the error terms are zero, but there is some debate regarding the actual specification of the correlations between the different factors. Some researchers allow for all possible correlations between the factors, while mentioning estimation problems⁷ (Kenny and Kashy 1992; Marsh and Bailey 1991; Eid 2000). Andrews (1984), Saris (1990) and Saris and Andrews (1991) suggest that the trait factors can be allowed to correlate, but should be uncorrelated with the method factors, while the method factors themselves are uncorrelated. When this latter specification is used, combined with the assumption of equal method effects for each method, almost no estimation problems occur in the analysis. This was demonstrated by Corten et al. (2002) in a study in which 79 MTMM experiments were reanalyzed.

⁷ This approach lends itself to non-convergence in the iterative estimation procedure or improper solutions such as negative variances.

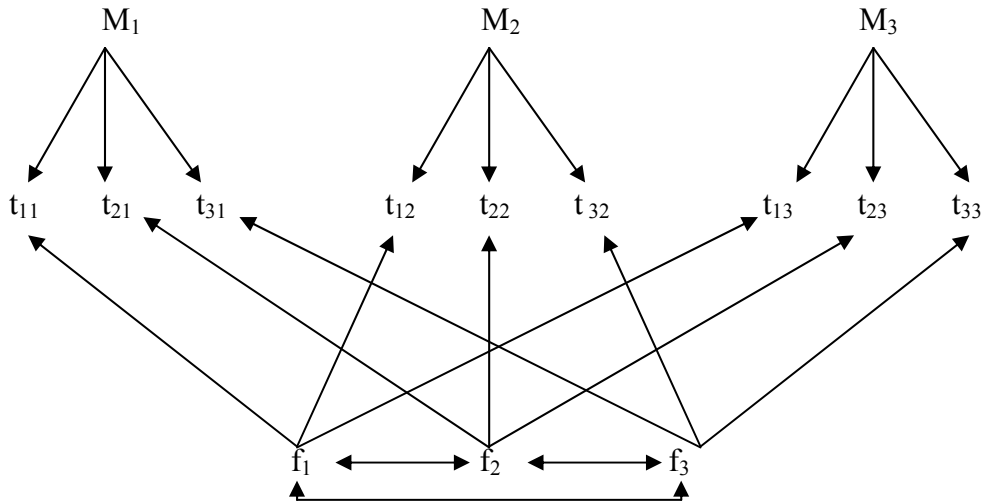


Figure 2.2: MTMM model illustrating the true scores and their factors of interest.

The MTMM design of 3 traits and 3 methods generates 45 correlations and variances. In turn, these 45 pieces of information provide sufficient information to estimate 9 reliability and 9 validity coefficients, 3 method effect coefficients and 3 correlations between the traits. There are a total of 24 parameters to be estimated. This leaves $45 - 24 = 21$ degrees of freedom, meaning that the necessary condition for identification is fulfilled. It also can be shown that the sufficient condition for identification is satisfied, and given that $df=21$, a test of the model is possible.

Many alternative models have been suggested for MTMM data. A review of some of the older models can be found in Wothke (1996). Among these is the *confirmatory factor analysis* model for MTMM data (Althausen et al. 1971; Alwin 1974; Werts and Linn 1970). An alternative parameterization of this model was proposed as the *true score* (TS) model by Saris and Andrews (1991), while the *correlated uniqueness* model has been suggested by Kenny (1976), Marsh (1989), and Marsh and Bailey (1991). Saris and Aalberts (2003) compared models presenting different explanations for the correlated uniqueness. Models with *multiplicative method effects* have been suggested by Campbell and O'Connell (1967), Browne (1984), and Cudeck (1988). Coenders and Saris (2000) showed that the multiplicative model can be formulated as a special case of the correlated uniqueness model of Marsh (1989). We suggest the use of the *true score (TS) MTMM model* specified by Saris and Andrews (1991) because Corten et al. (2002) and Saris and Aalberts (2003) have shown that this model has the best fit for large series of data sets for MTMM experiments. The classic MTMM model is locally equivalent with the TS model, meaning that the difference is only in its parameterization. See Appendix 1 for more details on why we prefer this model.

The Classical MTMM approach has its disadvantages. If each researcher performed MTMM experiments for all the variables of his/her model, it would be very inefficient and expensive, because he/she would have to ask six more questions to evaluate three original measures. In other words, the respondents would have to answer the questions about the same topic on three different occasions and in three different ways. This raises the questions of whether this type of research can be avoided; if this research is really necessary, and whether or not the work of the respondents can be reduced.

Most MTMM experiments to date have used the classic MTMM design or a panel design with two waves, in which each wave had only two observations for the same trait, while at the same time the order of the questions was random for the different respondents (Scherpenzeel and Saris 1997). The advantage of the latter method is that the response burden of each wave is reduced and the strength of opinion can be estimated (Scherpenzeel and Saris 2006). The disadvantages are that the total response burden is increased by one extra measure and that a frequently observed panel is needed to apply this design. Although this MTMM design has been used in many studies because of the presence of a frequently observed panel (Scherpenzeel 1995), we feel that this is not a solution that can generally be recommended. Given the limited possibilities of this particular design, other types of designs have therefore been produced, such as the split-ballot MTMM design (Saris, Satorra and Coeders 2004), which will be discussed in the next section.

2.4 The split-ballot MTMM design

In the commonly used split-ballot experiments, random samples from the same population receive different versions of the same questions. In other words, each respondent group gets one method. The split-ballot design makes it possible to compare the response distributions of the various questions and to assess their possible relative biases (Schuman and Presser 1981; Billiet et al. 1986).

In the split-ballot MTMM design, random samples of the same population are also used but with the difference that these groups receive two different forms of the same question. In total there is one less repetition than in the classical MTMM design and one more than in the commonly used split-ballot designs. We will show that this design combines the benefits of the split-ballot approach and the MTMM approach in that it enables researchers to evaluate measurement bias, reliability, and validity simultaneously, and that it does so while reducing the response burden. The suggestion to use split-ballot designs for structural equation models can be traced back to Arminger and Sobel (1991).

The two-group split-ballot MTMM design is structured as follows. The sample is split randomly into two groups. One group has to answer three survey items formulated using method 1, while the other group is given the same survey items presented in a second form, called “method 2.” in the MTMM literature. In the last part of the questionnaire all respondents are presented with the three items, which are now formulated in method 3 format. The design can be summarized as shown in Figure 2.3.

	Time 1	Time 2
Sample 1	Form 1	Form 3
Sample 2	Form 2	Form 3

Figure 2.3 The two-group split-ballot MTMM design.

In short, in the two-group design the researcher draws two comparable random samples from the same population and asks three questions about at least three traits in each sample: once with the same method and once with another form (method) of the same questions (traits) after sufficient time has elapsed. Van Meurs and Saris (1990) have demonstrated that the effects of memory are negligible after 20 minutes. This time gap is enough to obtain independent measures in most circumstances.

The design in Figure 3 matches the standard split-ballot design at time 1 and thus provides information on the differences in response distributions between the methods. Combined with the information obtained at time 2, this design provides extra information. The question of whether the reliability, validity and method effects can be estimated from this data still remains, since each respondent answers only two questions about the same trait and not three, as required for the classical MTMM design. The answer is not immediately evident, since the information necessary for the 9×9 correlation matrix comes from different groups and is by design incomplete (see Table 2). Table 2 shows the groups that provide data for estimating variances and correlations between questions using either the same or different forms (methods).

Table 2.2: Samples providing data for correlation estimation

	Method 1	Method 2	Method 3
Method 1	Sample 1		
Method 2	none	Sample 2	
Method 3	Sample 1	Sample 2	Sample 1+2

In contrast to the classical design, no correlations are obtained for form 1 and form 2 questions, as they are missing by design. Otherwise, all correlations in the 9×9 matrix can be obtained on the basis of two samples, but the data come from different samples.

Each respondent is given the same questions only twice, reducing the response burden considerably. However, the correlations between forms 1 and 2 cannot be estimated, leading to a loss of degrees of freedom when estimating the model on the now incomplete correlation matrix. This might make the estimation less efficient than the standard design in which all correlations are available, as in the three-group design. In large surveys the sample can be split into more subsamples and more than one set of questions hence evaluated. For more details of this approach, see Saris et al. (2004)

2.5 Estimating and testing models for split-ballot MTMM experiments

The split-ballot MTMM experiment differs from the standard approach in that different equivalent samples of the same population are studied instead of just one. Given that the random samples are drawn from the same population, it is natural to assume that the model is exactly the same for all respondents and the same as the model specified in Figure 2, which includes the restrictions on the parameters suggested by Saris and Andrews (1991). The only difference is that not all questions were asked in every group.

Since individuals were assigned to groups at random, and there is a large sample in each group, the most natural approach for estimation is the multiple -group SEM method (Jöreskog 1971). This approach is available in most SEM software packages. We refer to this approach as a multiple-groups structural equation model or MGSEM⁸.

⁸ Because each group will be confronted with partially different measures of the same traits, some software packages for multiple-group analysis will require some tricks to be applied. This is the case for LISREL, where the standard approach expects the same set of observable variables in each group. A simple trick to handle such a situation was described in the early work of Jöreskog (1971) and in the manual of the early versions of the LISREL program; such tricks are also described in Allison (1987). Multiple-group

As stated above, a common model is fitted across the samples, with equality constraints for all the parameters across groups. With the current software, and applying the theory for multiple-group analysis, estimation can be made by using the maximum likelihood (ML) method or any other standard estimation procedure in SEM. In the case of non-normal data, robust standard errors and test statistics are available in the standard software packages. For a review of multiple-group analysis in SEM models as applied to all the designs, see Satorra (1992, 2000).

The incomplete data set-up we are dealing with could also be considered as a missing data problem (Muthen et al. 1987). However, the approach for missing data assumes normality, while this design does not provide the theoretical basis for robust standard errors and corrected test statistics that are currently available in MGSEM software. Since the multiple-group option therefore offers the possibility of standard errors and test statistics which are protected from non-normality, we suggest that the multiple-group approach is preferable.

Given this situation, we suggest the MGSEM approach for estimating and testing the model using SB-MTMM data. In doing so, the covariance matrices are analyzed while the data quality criteria (reliability, validity coefficients and method effects) are obtained by standardizing the solution.

Although the statistical literature suggests that data quality indicators can be estimated using the SB-MTMM designs, we need to be careful when using the two group designs with incomplete data, because they may lead to empirical underidentification problems (Saris et al 2004). However under normal circumstances the model is identified and all parameters can be estimated. We will illustrate this approach below.

Many MTMM experiments have been carried out in recent decades (Scherpenzeel 1995). These experiments have provided information about the reliability and validity of 1087 questions. These questions were coded with respect to their characteristics and a meta-analysis was subsequently performed to determine the effect of the question characteristics on the quality criteria. The results of the meta-analysis have been reported in the book by Saris and Gallhofer (2007) which also introduces a program (SQP) for the prediction of the quality of questions based on this meta-analysis (Oberski et al 2004).

2.6 Conclusion and discussion

In this paper we hope we have shown that the Multitrait Multimethod approach to measurement problems in the social sciences can provide relevant information in terms of the reliability and validity of survey questions. In case of the use of the split ballot MTMM design, the approach can also provide information about the items missing values and bias, as well as other split ballot studies.

We argue that this approach is especially useful for subjective variables. It is often difficult to formulate alternative questions for objective variables, and to know whether memory effects can be excluded. The test-retest approach or the panel approach using the quasi simplex model is probably better for these variables.

We have also illustrated that relevant results can be obtained with the MTMM approach, suggesting that it is better to made use of item-specific scales than batteries of agree / disagree scales.

In Saris and Gallhofer (2007), we also presented the results of a meta-analysis of 87 MTMM experiments and a program (SQP) to predict the quality of survey questions.

analysis with the software EQS, for example, does not require the same number of variables in the different groups.

So far, this program can only predict the quality of questions in English, German and Dutch. Thanks to the experiments included in the ESS, it may be possible in the future to develop a new version of the SQP program that can predict the quality of questions in many other European languages.

The results of the MTMM experiments and the predictions of the program can be used to improve questions before the data are collected or for correction for measurement error after the data have been collected.

Chapter 3

New experiments in the ESS

Willem Saris

Irmtraud Gallhofer

Melanie Revilla

Diana Zavala

The Central Coordinating Team (CCT) of the European Social Survey has included from the very start next to the main questionnaire a supplementary questionnaire for methodological purposes in all countries. In this questionnaire alternative forms of some questions of the main questionnaire would be presented to the respondents in order to evaluate the quality of these questions using the SB-MTMM design with two subgroups. This means that all respondents get the chosen question forms in the main questionnaire while in the supplementary questionnaire two alternative forms are presented to randomly assign subgroups of the sample. As we have seen before such experiments will allow the estimation of the quality of all questions and allow for testing the comparability of the questions across countries. In the first part of this chapter we discuss which experiments in the different rounds have been introduced by purpose. In the second part we discuss the differences we have found between the questions that were not planned but occurred nevertheless in the process of the translation, layout and presentation of the questions to the respondents in the different countries

3.1 The Planned differences in the MTMM experiments

In this part we discuss the design of the SB-MTMM that has been planned in the first three rounds of the ESS.

3.1.1 Selection of experiments for round 1

It will be clear that the experiments cannot cover all variables used in the ESS. In the first round of the ESS the following crucial factors have been suggested for evaluation:

- a) open questions asking frequencies or amounts versus 7 point category scales
- b) dichotomous versus 5 points and 11 point scales
- c) 5 point agree/disagree items with statements versus item specific questions
- d) 11 point bipolar scales with show cards or without them
- e) 4 point bipolar scales versus 4 point unipolar scales and 11 point bipolar scales
- f) use of agree/disagree batteries compared with direct questions with construct specific responses

In this approach the choice of the topic is not so important but in the ESS we have to select for the experiment those topics which are in the main questionnaire already. The following choice was made for the different experiments:

- a) media use
- b) political efficacy
- c) social trust
- d) satisfaction with the economy, democracy and government
- e) trust in political institutions
- f) socio-political orientations

Other topics could have been chosen but we have chosen for sets of questions from the core questionnaire because they should get priority in the evaluation of their quality.

A compact summary of the design of the SB-MTMM experiments in the round 1 of the ESS can be found in Table 3.1. For the exact formulation of the questions we refer to the Appendix.

Table 3.1 Round 1: The SB-MTMM experiments					
Experim.	Var.	Meaning	main	SC-A	SC-B
Media 1	tvatot	- On an average weekday, how much time, in total, do you spend watching television?	8 categ. In hours	In hours and min	7 categ gener al
	rdtot	- On an average weekday, how much time, in total, do you spend listening to the radio?			
	nwsptot	- On an average weekday, how much time, in total, do you spend reading the newspapers?			
Pol. eff 2	polcmpl	- Politics seems so complicated that I can't really understand	5is	5ad	5ad
	polactiv	- I could take an active role in a group involved with political issues			
	polcds	- Easy to make my mind up about political issues			
Political orientatio n 3	ginveco	- The less the government intervenes in the economy, the better for the country	5ad batt	5ad	5is
	gingdif	- The government should take measures to reduce differences in income levels			
	needtru	- employees need strong trade unions to protect their working conditions			
Satisfacti on 4	stfecoc	- On the whole how satisfied are you with the present state of the economy in [country]?	11is	4is	6is fixed
	stfgov	- Now thinking about the [country] government, how satisfied are you with the way it is doing its job?			
	stfdem	- And on the whole, how satisfied are you with the way democracy works in [country]?			
Social trust 5	ppltrst	- Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?	11is	6is	2is
	pplfair	- Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?			
	pplhlp	- Would you say that most people look out for themselves or that they try to be helpful?			
Political trust 6	trstprl trstlgl trstplc	How much do you personally trust each of the institutions: - [Country]'s parliament - The legal system - The police	11is batter y	11is	11is score

Note: is=item specific scale, ad= agree/disagree scale, batt=battery, fix=fixed reference point

3.1.2 The selection of the MTMM experiments for Round 2

The selection of the experiments in the second round of the ESS are summarized below.

Experiment 1 How to ask numeric questions

In the MTMM experiments of round 1 we had seen that the frequency and amount of questions had very bad quality. Therefore, the format of such questions has been tested in the pilot of round 2 and a new version has been included in the main questionnaire. An experiment with alternative forms is done in order to see how these different versions work in the different countries. In Table 3.2 the experiment is summarized. The exact formulation of the questions can be found in the Appendix.

Experiment 2 different position of the items on the scale

In survey research very often batteries of statements are used, where within the statements an arbitrary position on the underlying dimension is specified. For example, it is said that something is “usually” done or “seldom” done. This choice is arbitrary but may have consequences for the results. One can also avoid such arbitrary choices and ask people to specify how frequently the activity happens on a scale from never to always. This experiment has been done with items about activities of doctors. Table 3.2 gives a summary of the experiment. The questionnaire is presented in Appendix.

Experiment 3: Item specific categories or batteries

In this experiment a comparison has been made between a standard battery and a set of separate questions with item specific response categories. The second form had 4 categories like the battery; the third form had 11 categories. The questions concern characteristics of a job. The summary of the experiment can be found in Table 3.2. The questions are presented in the Appendix.

Experiment 4: The use of different labels and positions of items

In this experiment we test the effect of the positions of the items on the underlying dimension as in experiment 2 but also the effect of scale with long labels at both sides of the scale. The number of categories was each time the same. The positions of the items were changed by changing the item from positive to negative while in the last form the positive and negative statements were placed at the end points of the scale. The topic was the role of men and women in society. The experiment is summarized in table 3.2. For details of the used questions we refer to the Appendix.

Experiment 5: The use of fixed reference points

In the ESS we usually use what has been called fixed reference points, i.e., labels that have a fixed position on the underlying opinion scale for example “extremely satisfied”. That is definitely the end point of the satisfaction scale. A non fixed reference point could be “very satisfied”. Some people will see it as the end point of the scale others don't. This difference of perception can cause differences in responses that have nothing to do with the substantive opinions. Therefore, fixed reference points have advantages. This experiment should show if this is indeed the case in all countries and also if 3 fixed reference points are better than 2. The topic for the experiments were the

satisfaction questions belonging to the Core of the ESS. In table 3.2 the experiment is summarized. The exact formulation can be found in Appendix.

Experiment 6: The effect of repetition for different items

These MTMM experiments are not possible without repeated observations. In our SB-MTMM design the number of repetitions has been reduced to 1 for all respondents but this can nevertheless have a positive (memory) or a negative (inaccurateness) effect on the quality of the data. This can be seen in an experiment, where exactly the same questions are repeated in the different parts of the data collection. This experiment is done with questions with respect to “trust in political institutions”. The summary of the experiment can be found in Table 3.2.; the exact formulation of the questions can be found in the Appendix.

Table 3.2. Round 2: The SB-MTMM experiments						
Experim.	Var	Meaning	ain	C-A	C-B	C-C
hwk 1	Hwktwd1	- On a typical weekday about how many hours, in total, do people in your household spend on housework for your home?	n hours + 6 pt scale	n hours	n hours + in %	
	Hwkpwd1	- And about how much of this time do you spend yourself?				
	Hwkpwdp	- And about how much of this time does your husband/wife/partner spend on housework?				
doc 2	dckptrt dctreqd dcdisc	- Doctors keep the whole truth from their patients - GPs treat their patients as their equals. - Before doctors decide on a treatment, they discuss it with their patient	IS	AD arely		AD sually
job 3	vrtypwrk jbscr hlthrwk	- There is a lot of variety in my work. - My job is secure - My health or safety is at risk because of my work	AD		IS	1IS
women 4	wncpwrk mnrspbm mnrgrtjob	- A woman should be prepared to cut down on her paid work for the sake of her family (main + SC-B) <i>Women should NOT be prepared... (SC-A)</i> - Men should take as much responsibility as women for the home and children. (main + SC-B) <i>Women should take more responsibility for the home and children. (SC-A)</i> - When jobs are scarce, men should have more right to a job than women. (main + SC-B) <i>When jobs scarce, women should have the same right to a job as men. (SC-A)</i>	AD attery	AD	IS	
satisf 5	Stfecov stfgov stfdem	- On the whole how satisfied are you with the present state of the economy in [country]? - Now thinking about the [country] government, how satisfied are you with the way it is doing its job? - And on the whole, how satisfied are you with the way democracy works in [country]?	1IS ixed	1IS ixed iddle label		1IS
trustin 6	trstprl trstlgl trstplt	How much do you personally trust each of the institutions: - [Country]’s parliament - The legal system - The politicians	1is		1is	1is

Note: is=item specific scale, ad= agree/disagree scale, batt=battery, fix=fixed reference point

3.1.3 The selection of the MTMM experiments for Round 3

There were two reasons for the proposals for MTMM experiments in this round. The first was that some experiments done in the pilot study of round 3 were not conclusive and the second was that some basic questions, used in the core, were not evaluated with respect to their quality. Given these two reasons the MTMM experiments for the Round 3 data collection has been formulated. We will discuss the proposed designs in sequence starting with the proposals based on the pilot experiments.

Proposals based on the pilot experiments

In experiment 1 of the Pilot we tried to see if the number of categories increases the quality of the question. This was indeed true for the 11 point scale compared with the 5 point scale. However the results for the 7 point scale were different from the expectations. The reversal of the numbers at the categories may be the cause.

Given that it is important to know if these results hold for all countries involved, we suggested repeating this experiment for all countries but with the corrections suggested above and limiting the experiment to the 3 positive items of the set of items.

The second experiment concerned the hypothesis that the effect of variation in the use of scales will reduce the method effects. Although the hypothesis sounded plausible the effects were not found in the pilot study because the method effects were not significant. A possible reason for the lack of method effects in this case is that the items in this scale represented positive and negative points of view and a positive item was always followed by a negative one and vice versa. This means that respondent has to think about the use of the scales anyway. If one answers the questions a bit attentively one has to switch from agree to disagree all the time. This seems to have happened here.

The conclusion was that this experiment cannot be done as it has been done. We suggested doing it in a different way. Our suggestion was to select only the three positive items from the items 45-50 and use them in the experiment. In doing so we can also see if balancing the scale reduces the method effect because in the main questionnaire a balanced scale will be used.

Proposals based on Core questions

Given the above specified experiments there was still space for more experiments in the supplementary questionnaire. We suggested to experiment with two topics of the core questionnaire that have not been evaluated yet. Those are the immigration questions and the consequences of more immigrants.

The first set is measured with a 4 point agree/disagree format in the main questionnaire. We suggested in group 1 to repeat the 5 point scale, in group 2 to use a 4 point scale and in group 3 a 7 point scale. In this way the variation of the scales experiment can be extended and we get more information about the effect of the number of categories.

The consequence of immigration is asked using an 11 point scale with anchored end points. This approach is arbitrary. One could also have used statements with an agree/disagree format. In groups we suggested to use a 5 point agree/disagree scale and 11 point agree/disagree scale in group 2. Finally we suggested using in group 3 a 7 point scale.

The summary of the proposal can be found in Table 3.3. For the exact formulation of the questions we refer to the Appendix.

Table 3.3 Round 3: The SB-MTMM experiments						
Exp	Var.	Meaning	Main = M1	gpA = M2	gpB = M3	gpC = M4
dngval 1	<i>dngval</i> <i>ppllfcr</i> <i>flclpla</i>	- I generally feel that what I do in my life is valuable and worthwhile - There are people in my life who really care about me - I feel close to the people in my local area	AD	AD	AD	AD
imbgeco 2	<i>imbgeco</i> <i>imueclt</i> <i>imwbent</i>	- It is generally bad for [country's] economy that people come to live here from other countries - [Country's] cultural life is generally undermined by people coming to live here from other countries - [Country] is made a worse place to live by people coming to live here from other countries	1IS	AD	1AD	AD
imsmetn 3	<i>imsmet</i> <i>imdftn</i> <i>impcntr</i>	- [Country] should allow more people of the same race or ethnic group as most [country's] people to come and live here. - [Country] should allow more people of a different race or ethnic group from most [country's] people to come and live here. - [Country] should allow more people from the poorer countries outside Europe to come and live here.	IS	AD	IS	AD
lrnnew 4	<i>lrnnew</i> <i>accdng</i> <i>plprftr</i>	- I love learning new things - Most days I feel a sense of accomplishment from what I do - I like planning and preparing for the future.	AD	AD	1IS	1AD

Note: is=item specific scale, ad= agree/disagree scale, batt=battery, fix=fixed reference point

So far we have presented the variation in the questions that have been made by purpose. There are, however, also differences which have been occurred in the questions and the data collection methods for other reasons.

Due to the fact that we have to use questions present in the questionnaire not just the above mentioned characteristics will vary across the experiments but also other characteristics for example:

- the position in the questionnaire,
- the distance to the next MTMM question
- the mode of data collection etc, (see restrictions) but also
- the length of the question text,
- the number of sentences, the number of labels etc.

Some of these differences are logical consequences of the formulation of the questions. Others are a consequence of the fact that the Central Coordinating Team (CCT) of the ESS does not have complete control over the way the questions are translated and presented to the respondents.

In preparation of the meta analysis of the MTMM experiments all questions have been coded on the characteristics that are included in the program SQP. As a consequence we have a complete overview of the differences between the questions in all countries. This will be the topic of the second part of this chapter.

3.2 The Unplanned differences in questions across the countries

Besides the variation in the question design planned by the researchers, many other differences have been detected due to the translations, the layout of the questions and the presentation of the questions to the respondents. The most important differences detected will be presented below. However, before we discuss this issue, we first pay attention to the procedure used to detect these differences. It is not so clear that one can detect these differences because the questionnaires have been translated in many different languages which complicate the comparison.

3.2.1. The procedure to detect the differences

In order to detect the characteristics of all the questions that were involved in the MTMM experiments in all the participating countries a new version of the program SQP has been used⁹. In the program more than 50 formal characteristics of survey questions, the answer categories, the show cards, the data collection method, the survey characteristics etc. are coded.

Because in many of the ESS countries different languages are spoken and used in the questionnaires, coders had to be found which were native speakers in all the languages. Fortunately it was possible to find sufficient native speakers for all languages.

In order to check the quality of the coding first some experiments were coded by two coders in order to see whether the agreement of the codes was sufficient to rely on a single coder. It turns out that coders often make errors, mostly by mistake. If the two coders spoke about the differences in their coding it was in general easy to come to a consensus about the correct code.

Given this experience we have decided that first two coders would make a consensus coding of the source questionnaire of each round. Consequently the coding of each coding of each question of each questionnaire in all languages and countries was compared with the consensus coding of the source questionnaire¹⁰. If a difference was detected the coordinator of the coding process spoke with the specific coder about the reasons for the difference. It could be that a mistake was made in the coding. However it was also possible that there was an unexpected difference in the coding in the question text in the specific country. In the former case the code was adjusted in the later case the code remained as it was so that now can be seen that a question on a specific characteristic was different from the characteristic in the source questionnaire. All the codes have been stored in the question data file included with the question text in the different languages. In the next section we will give some results with respect to the differences which have been found using this procedure.

3.2.2. The differences in characteristics between source and countries

Below we present the distribution of the questions in the source questionnaire and the questionnaires developed in the different countries. First we will give some results for characteristics of the questions that should not be different in the different countries, especially the concept of the questions, the domain of the question and the basic form of the questions. Each time we make a comparison between the distribution of the questions over the categories of the different characteristics found in the source

⁹ This program has been developed by Daniel Oberski and Thomas Gruner and is now a part of the new SQP program.

¹⁰ For this purpose again a special program "Compare" was made by Daniel Oberski.

questionnaire and the ones in the other countries after translation. In principle the proportions should be the same but by leaving out one question or making error small deviations can occur. The tables 3.4, 3.5 and 3.6 show indeed that the differences are minimal.

Table 3.4	Source		Other countries	
	Absolute	%	Absolute	%
Concept				
Evaluative belief	34	30.63	538	29.97
Feeling	42	37.84	700	39.00
Facts. background or behaviour	2	1.80	34	1.89
Evaluation	13	11.71	211	11.75
Norm	3	2.70	42	2.34
Policy	14	12.61	228	12.70
Action tendency	3	2.70	42	2.34
Total	111	100.00	1795	100.00

Table 3.5	Source		Other countries	
	Absolute	%	Absolute	%
Domain				
Health	6	5.41	102	5.68
Living conditions and background variables	12	10.81	197	10.97
Other beliefs	0	0.00	1	0.06
Work	20	18.02	340	18.93
Personal relations	16	14.41	231	12.86
Leisure activities	4	3.60	68	3.79
National politics: national government	7	6.31	119	6.63
National politics: national institutions	15	13.51	248	13.81
National politics: economic/financial matters	14	12.61	223	12.42
National politics: other	17	15.32	267	14.87
Total	111	100.00	1796	100.00

Table 3.6	Source		Other countries	
	Absolute	%	Absolute	%
Basic form				
No request present	48	43.24	801	44.60
Indirect request	40	36.04	552	30.73
Direct request	23	20.72	443	24.67
Total	111	100.00	1796	100.00

On the other hand differences can occur for different reasons. Table 3.7 shows that in some countries the procedure to use show cards for the questions was not followed all the time.

Table 3.7	Source		Other countries	
	Absolute	%	Absolute	%
Showcards				
Showcard not used	39	35.14	418	23.27
Showcard used	72	64.86	1378	76.73
Total	111	100.00	1796	100.00

At the time of this research such differences were not controlled by the Central Coordinating Team. This is now not possible anymore.

A similar phenomenon we see in table 3.8 presenting the distribution of questions with respect to unipolar and bipolar scales. Also in this case a difference is not necessary. In this case it is a bit more complicated because for example for trust one can formulate the question from “complete distrust to complete trust” or from “no trust at all to complete trust”. One can debate in this case whether the second scale is a proper translation of the first one. Some national coordinators, responsible for the translations, may have thought so but our coders did not think so. They coded the latter scale as unipolar.

Table 3.8	Source		Other countries	
Range of the used scale	Absolute	%	Absolute	%
Unipolar	39	35.14	846	45.07
Bipolar	72	64.86	1031	54.93
Total	111	100.00	1877	100.00

Finally we ask attention for the difference between the source questionnaire and the translations in other languages with respect to specifying “fixed reference points”. In the source questionnaire often scales are used with only the end point labelled with terms: “extremely bad to extremely good”. This is done because in this way the end points of the scales are clearly indicated and they got a fixed value on the numeric 11 points response scale. It can be seen in this table that this procedure was not always followed in the translations in the different countries. Often they use as labels for the end points like “very bad to very good”. However these labels were not coded as fixed reference points because people can think that these labels do not indicate the end points of the scales.

Table 3.9	Source		Other countries	
Fixed reference points	Absolute	%	Absolute	%
0	3	2.78	177	9.65
1	75	69.44	783	42.67
2	20	18.52	289	15.75
3	10	9.26	586	31.93
Total	108	100.00	1835	100.00

In a recent study (Zavala 2011) it was detected that in some countries, especially the Slavic countries, it is impossible to find a proper alternative for “extreme” and this has caused this difference.

It is, of course, not possible to give an overview of the distributions of all questions. That would require too much space. These examples illustrate what has been done in this coding phase and what the results are like.

3.3 Conclusions

In this chapter we have given an overview of the different experiments that have been done in the first three rounds of the ESS. It has been shown that several alternative formulations have been tested for different questions.

For all the questions which have been included in these experiments the quality of the questions have been estimated based on the MTMM experiments in which they were involved.

However it should also be clear that these experiments were done for specific topics (domains) so in principle we cannot simply generalize these results from these specific experiments suggesting to have found general results. In order to do so some experiments have been repeated for different topics. Besides that we have these experiments of the ESS combined with the earlier studies done using the same MTMM approach in order to get a more general result with respect to the quality of the questions. Over this complete data set a meta analysis has been done to make general statement about the effects of the different question characteristics on the quality of the questions.

In this context we have taken into account that not in all countries the instructions of the CCT have been followed. Because this happened, we have coded all questions on their characteristics and in the further analysis we take these differences into account.

Chapter 4

Estimation problems and solutions¹¹

Melanie Revilla

Willem Saris

Saris, Satorra and Coenders (2004) proposed a new approach to estimate the quality of survey questions, combining the advantages of two existing approaches: the multitrait-multimethod (MTMM) and the split-ballot (SB) designs. Implemented in practice, this new approach led to frequent problems of non-convergence and improper solutions. This paper uses Monte Carlo simulations to understand how the SB-MTMM approach can be improved to avoid the problems detected. The number of SB groups is a crucial element: the 3-group design is performing better.. However this leads to practical problems. Therefore it was studied how many respondents are needed in the third group to get acceptable estimates. Increasing the sample size of the groups in the 2 group SB-MTMM design is a possibility. For different reasons we have finally decided that the best solution for the estimation the parameters was a two step procedure: starting with Multiple group analysis assuming that for each experiment the parameters in all countries are the same; secondly testing for misspecifications in the model i.e. allowing for differences between the countries for parameters that are different. This approach works because we start with one model and approximately 40.000 cases.

4.1 Introduction

In Chapter 2 we have explained that for the ESS the SB-MTMM design was developed in order to evaluate the quality of questions across countries. The ESS used in each round a 2-group SB-MTMM design to collect data for several MTMM experiments in 20 - 30 countries. The survey is divided into a main questionnaire (same for all respondents: M_1), and a supplementary questionnaire, that differs for the two SB groups (M_2 in group 1, M_3 in group 2, cf. section 1). In round 3 we have a third group which got method 3. In that case the comparison was between 4 methods.

The 3-group design has also been implemented: for instance, in December 2008, the Longitudinal Internet Studies for the Social Sciences (LISS¹²) panel presented to its respondents a survey including some 3-group SB-MTMM experiments. The 3-group design however is more difficult to implement. Indeed, in the 2-group design the methods differ only at time 2 for the different SB groups, whereas in the 3-group design, they differ both at times 1 and 2. It is not possible with the 3-group design to have one main questionnaire similar for all respondents. Preparing the survey is therefore more demanding. Besides, researchers who want to analyse identical questions can only use two out of the three SB groups, so it reduces their sample size. Even if it concerns only the variables included in the MTMM experiments, many survey institutes prefer to use the 2-group designs. However, this leads to recurrent problems in the analyses.

¹¹ More information about this issue can be found in : Revilla M. and W.Saris (2011) The split-ballot MTMM approach: implementation and problems. Barcelona, RECSM working paper 19.

¹² Dutch Web panel based on probability sample. For more information, please see: <http://www.centerdata.nl/en/LISSpanel>

4.2 Main problems encountered in practice

Rindskopf already remarked in 1984 that in practice “structural equation models are often plagued by a variety of undesirable results” (p.109). He argued it was a consequence of empirical underidentification: “for most models, one cannot say that the model is identified but only that it may be identified if certain conditions are true. [...] The conditions for identification generally take the form of requiring that certain parameters not to be zero or that parameters not equal one.” (p. 110). If these conditions are not satisfied in a specific dataset, then undesirable results may arise.

Since 1984, much work has been done on structural equation models, but the issue of undesirable results is still present. For the SB-MTMM model, the “undesirable results” take mainly two forms: non-convergence (NC) and Heywood cases (HC). HC or “improper solutions” correspond to “negative variances or correlation estimates greater than one in absolute value” (Kolenikov and Bollen, 2008, p.1). Biased estimates may also be an issue but without knowing the true values, it is difficult to notice it.

NC is problematic since if the parameters cannot be estimated, no conclusion can be drawn. HC are also problematic. Negative variances may appear just because of sampling fluctuations if the true value of the parameter is close to zero (Van Driel, 1978). That is why it is often argued that HC can be simply solved by fixing to zero the negative but non-significant values. However, Rindskopf (1984) underlined that “the corrective action to take is not always obvious; for example, it is not always correct to remove a parameter from an analysis when it has negative error variance estimate, because the problem may be caused by another variable” (p. 110).

Despite this warning, fixing the negative non significant values to solve HC is a quite common procedure, implemented for example in Saris et al. (2004, p. 331).

Nevertheless, our analyses of real SB-MTMM data are in line with Rindskopf’s comment and suggest that fixing negative non significant estimates may have a large impact on other estimates of the model and may not really be a solution. This can be illustrated by the 2-group SB-MTMM experiment about satisfaction in the Netherlands collected in the first ESS round (2002-2003). The three traits deal with satisfaction with the “present state of the economy”, the “way the government is doing its job” and the “way democracy works in the country”. In the main questionnaire, respondents get an 11-point scale going from “extremely dissatisfied” to “extremely satisfied” (M_1). In the supplementary one, group 1 gets a 4-point scale going from “very dissatisfied” to “very satisfied” (M_2), whereas group 2 gets a 6-point scale going from “extremely dissatisfied” to “extremely satisfied” (M_3).

The covariance matrices are analysed using ML estimation for MG in LISREL¹³ (Jöreskog and Sörbom, 1991). The model used is the true score model (cf. Chapter 3). Since the respondents are randomly assigned to the SB groups, one does not expect significant differences across groups for the same questions, so the parameters in the second group are specified invariant (details in Saris and Gallhofer, 2007, chapter 12).

The estimation of this satisfaction experiment in the Netherlands leads to a HC: the method variances for M_2 and M_3 are negative, but according to a t-test not significant. We start by fixing M_2 variance. The variance of M_3 being still negative, we also fix it and get a proper solution (PS).

To determine if the model appropriately reproduces our data, we use the software JRule (Van der Veld et al., 2008) based on the testing procedure developed by Saris et al. (2009). Using information about types I and II errors, it provides a test for misspecifications at the parameter level. According to JRule, the method variances fixed are not misspecified and the model cannot be rejected. This seems to provide support to

¹³ An example of Lisrel input to analyze SB-MTMM experiments is available online: <http://bit.ly/gQI3sV>

the procedure of fixing negative variances. However, instead of M_2 and M_3 variances, we could also fix the variance of M_1 . This is an alternative way of getting a PS. Also this model cannot be rejected. In particular, no misspecifications are found for the method variance fixed (variance of M_1).

Even if we got proper solutions in both cases, the results seem determined by the choice of fixing one or the others method variances. The 11-point scale for instance (M_1) has the lowest quality of all three methods in the first situation (when fixing the method variances of the M_2 and M_3) but the highest in the second one (when fixing the method variance for M_1).

One could argue that the first model is the good one: fixing a non significant parameter seems more acceptable than fixing the only positive and significant variance. However, the second model cannot be rejected according to JRule¹⁴. We are more willing to think that getting so different estimates with two fitting models suggests that both sets of estimates are biased because of the decision of fixing some parameters. So getting HC may really lead to problematic situations where it is not clear what to do.

4.3 Frequency of these problems

NC and HC are all the more problematic as they are occurring very frequently. The first and fourth rounds of the ESS are used to illustrate this. In the first round, six SB-MTMM experiments (with three traits and three methods) dealing with media use, political efficacy, political orientation, satisfaction, social and political trust are analysed in 19 countries. In the fourth round, three SB-MTMM experiments dealing with media use, satisfaction and political trust are considered. 22 analyses are run based on the country and language of the interview. In total, $6 \cdot 19 + 3 \cdot 22 = 180$ SB-MTMM experiments are therefore studied¹⁵.

Table 4.1 reports the number of NC, HC and PS. One can notice that for the NC cases, one does not know if solving the non-convergence would lead or not to a proper solution.

Table 4.1: Results obtained when running 180 SB-MTMM models for ESS rounds 1 and 4

Experiments	NC	HC	PS	Total cases
Round 1				
Media use	15	4	0	19
Pol. efficacy	1	11	7	19
Pol. orientation	4	8	7	19
Satisfaction	3	9	7	19
Social trust	3	13	3	19
Political trust	2	10	7	19
Round 4				
Media use	16	6	0	22
Satisfaction	9	10	3	22
Political trust	1	13	8	22
Total across experiments (Total in %)	54 (30.0%)	84 (46.7%)	42 (23.3%)	180 (100%)

Note: NC = not convergent, HC = Heywood case, PS = proper solution

¹⁴ The two models are also very similar and cannot be rejected if we consider more global tests of the model as the Chi-square or fit indices as RMSEA.

¹⁵ For more details about the traits and methods used in each experiment, as well as for the list of countries (or countries/languages groups) analyzed in each round, please see: <http://bit.ly/hH07b7>

Table 4.1 shows that only in 23.3% of the datasets a proper solution is obtained, whereas 30.0% of the datasets lead to non convergence and 46.7% to Heywood cases. Differences between experiments may be observed: the media use experiment seems particularly problematic in both rounds, with no PS at all.

As seen in Chapter 3, Saris et al. (2004) mention that in some cases the 2-group design may not be empirically identified, in particular when there is no correlation between the traits. This is what seems to happen in the media use experiment. The correlations between the reported time spent watching television, listening to the radio and reading newspapers are almost zero. This may explain the problems encountered. For the other topics however, the results are worse than expected from the reading of Saris et al. (2004) and there is no clear explanation. In addition, for the same experiment, sometimes within one country from one language to another, the SB-MTMM experiment may in one case provide directly a PS but in the other not.

4.4 Possible reasons for the problems

Given this problematic situation Revilla and Saris (2011) did a Monte Carlo simulation study to determine under which conditions the NC and HC are occurring. Understanding when they are encountered may help finding how to solve them by preventing these conditions to happen. Based on the warnings made by Saris et al. (2004), three main explanations were considered:

- the role of the number of split-ballot groups: are there more problems in the 2-group SB design because of the incomplete design?
- the closeness of the true values to boundaries: are the HC occurring because the true values are close to zero?
- the similarities between different true values: are there more non convergence problems because of these similarities?

Revilla and Saris (2011) came to the following conclusions:

The problems occur with the 2-group design but not with the 3-group design. The number of SB groups used is the first main condition determining if the SB-MTMM approach is or is not performing well.

The more similar the true correlations between the traits in a 2-group design, the higher the probability of getting problems. Regression analyses suggested that the interaction between the absolute true values of the correlation between the traits and differences in correlations between the traits has a significant effect on the convergence and on the bias. So complex mechanisms are at work to determine when the 2-group design performs properly.

4.5 How can these problems are solved

Trying to identify under which conditions problems are occurring is interesting from a theoretical point of view, but it needs also to be related to practice. To get more insight in these problems Monte Carlo simulation were performed. Each experiment consisted of 500 simulations with 500 cases. So far, the analyses suggest it is preferable when designing MTMM experiments to choose traits that are sufficiently but not equally correlated. This may however be difficult to design. The true correlations between traits may be known from previous studies or an expected value may be deduced from the theory. But if, as shown in Revilla and Saris (2011), a set of correlations between traits of .1, .2, .8 leads to problems, whereas a set of correlations of .2, .3, .9 does not, a very precise knowledge is needed, which is most of the time not the case in practice. This section therefore focuses on potential solutions when facing problems, in particular HC. In order to determine how these problems can be solved

Monte Carlo simulations have been used to look for possible solutions. The simulations have been done with two sets of parameters: case 1 is a problematic set and case 2 is less problematic. The values are presented in table 4.2. Two situations are considered separately: one where the data has not been collected yet and one where the data has already been collected.

Table 4.2: List of values of the parameters

	Case1	Case2		Case1	Case2		Case1	Case2
Ga11	.74	.735	Te11	.35	.30	Ph21	.46	.60
Ga22	.82	.735	Te22	.23	.30	Ph31	.50	.10
Ga33	.74	.735	Te33	.35	.30	Ph32	.43	.30
Ga41	.70	.735	Te44	.45	.30	Ph11	1	1
Ga52	.74	.735	Te55	.39	.30	Ph22	1	1
Ga63	.74	.735	Te66	.39	.30	Ph33	1	1
Ga71	.86	.735	Te77	.17	.30	Ph44	.10	.16
Ga82	.86	.735	Te88	.17	.30	Ph55	.06	.16
Ga93	.83	.735	Te99	.22	.30	Ph66	.09	.16

4.6 If the data has not been collected yet

If the data has not been collected yet, the researchers have some freedom in order to solve the problems. Different potential solutions are tested below using simulations.

4.6.1 Increase the sample size?

Saris et al. (2004) show the sample sizes needed to get the same accuracy in the estimation are larger in the 2-group design. Increasing the sample size may therefore improve the performance of the 2-group design: the higher the sample size, the more accurate the estimates. Different sample sizes are therefore tested. The number of replications for each simulation is still 500. The analyses are done only for the 2-group design since in the 3-group design the results are already acceptable with 500 observations. Results for case 1 are given in the top part of table 4.3.

The table shows that, indeed, when the sample size increases, the number of NC decreases. Besides, the average estimate for M_I variance increases little by little and finally becomes positive. So the HC problem seems to be solved by increasing the sample enough. But “enough” means at least 5.000 observations are necessary in order to do get in average a positive variance for M_I and preferably 10.000 or more to really get an accurate solution. Theoretically, increasing the sample size is therefore, as expected, a solution. However, practically, the sample sizes needed in case 1 to reach accuracy are much too large for most of the surveys’ budgets. In the ESS, sample sizes are rarely higher than 2.000. Asking for five times this number is often unthinkable.

However, case 1 has been chosen for being particularly problematic. It is interesting therefore to look at another an example where the estimation was not good for the 2-group design with 500 observations, but not as bad as with case 1. The bottom part of Table 3 shows what is happening when increasing the sample size in case 2.

Table 4.3: Increasing sample size for case 1 and 2

2-group	Number of observations										
Case 1	500	800	1000	1500	2000	5000	7500	10000	15000	20000	true
Ga11	.9838	.9619	.9494	.8998	.8880	.8080	.7672	.7546	.7480	.7448	.74
(SD)	(.4230)	(.4331)	(.4032)	(.3734)	(.2996)	(.2171)	(.1151)	(.0792)	(.0690)	(.0584)	
Ga41	.5840	.5998	.6029	.6255	.6215	.6653	.6868	.6935	.6982	.7001	.70
(SD)	(.1680)	(.1638)	(.1601)	(.1487)	(.1354)	(.1098)	(.0852)	(.0701)	(.0611)	(.0548)	
Ga71	.7190	.7424	.7466	.7767	.7688	.8179	.8449	.8534	.8584	.8606	.86
(SD)	(.1987)	(.2005)	(.1955)	(.1815)	(.1667)	(.1327)	(.1024)	(.0845)	(.0731)	(.0647)	
Ph44	-.2007	-.1813	-.1581	-.0973	-.0664	.0224	.0726	.0859	.0919	.0951	.10
(SD)	(.6787)	(.7417)	(.6435)	(.6485)	(.4189)	(.3372)	(.1131)	(.0647)	(.0557)	(.0458)	
Ph55	.1263	.1150	.1116	.0983	.1026	.0773	.0654	.0620	.0595	.0586	.06
(SD)	(.0875)	(.0895)	(.0876)	(.0832)	(.0770)	(.0653)	(.0559)	(.0489)	(.0430)	(.0382)	
Ph66	.1771	.1617	.1595	.1411	.1466	.1130	.0967	.0919	.0887	.0875	.09
(SD)	(.1153)	(.1234)	(.1205)	(.1136)	(.1049)	(.0910)	(.0765)	(.0668)	(.0583)	(.0527)	
Number conv	266	281	302	324	340	410	450	462	492	496	500
Average bias	.1592	.1293	.1216	.0922	.0878	.0400	.0147	.0073	.0033	.0020	
Average MSE	.1771	.1872	.1573	.1442	.0937	.0833	.0387	.0347	.0334	.0325	
Case 2	500	800	1000	1500	2000	5000	7500				true
Ga11	.9540	.8428	.8060	.7812	.7640	.7396	.7377				.735
(SD)	(.5661)	(.3517)	(.2507)	(.1904)	(.1576)	(.0715)	(.0577)				
Ga41	.6413	.6891	.7056	.7154	.7247	.7352	.7352				.735
(SD)	(.1885)	(.1603)	(.1509)	(.1273)	(.1124)	(.0727)	(.0575)				
Ga71	.6426	.6917	.7077	.7191	.7284	.7359	.7367				.735
(SD)	(.1821)	(.1578)	(.1469)	(.1296)	(.1118)	(.0703)	(.0564)				
Ph44	.1213	.1293	.1455	.1504	.1533	.1592	.1597				.16
(SD)	(.1225)	(.2049)	(.0690)	(.0609)	(.0539)	(.0161)	(.0130)				
Ph55	.1717	.1659	.1626	.1605	.1604	.1587	.1596				.16
(SD)	(.0497)	(.0398)	(.0370)	(.0320)	(.0268)	(.0176)	(.0141)				
Ph66	.1706	.1640	.1620	.1595	.1599	.1584	.1589				.16
(SD)	(.0481)	(.0430)	(.0373)	(.0331)	(.0288)	(.0180)	(.0146)				
Number conv	348	381	419	443	465	496	499				500
Average bias	.0777	.0396	.0245	.0154	.0088	.0016	.0011				
Average MSE	.1168	.0781	.0598	.0528	.0491	.0432	.0424				

Again, as the sample size increases, the NC problem is reduced and the average estimates get more and more accurate. Moreover, the increase in sample size needed to improve the performance of the 2-group design is much smaller in case 2 than in case 1. Results are already quite accurate for 2.000 observations and for 5.000 they are really close to the true values. Results for case 1 were extreme. In other situations, increasing the sample size can be a solution since a reasonable sample size may solve the problems. The difficulty then is how to determine in advance the sample size needed for a specific experiment.

4.6.2 Use 3-groups with a small third group?

We saw that in average the 3-group design seems to solve most of the problems. On the contrary, when the 2-group design is applied to real data, a PS is obtained in only 23.3% of the cases considered (see table 1). Besides, more than 10.000 observations may be needed in some cases to get accurate estimates by increasing the sample size. This realised, one may reconsider the difficulty of implementation of the 3-group design and think more deeply about the possibility of using it.

What is bothering with the 3-group design is that not all respondents get one common method, such that researchers that want to use one measure of one variable for their research cannot use part of the respondents. To limit the number of respondents that cannot be used, we could think of a 3-group design with three groups of unequal sizes. In particular, having two main groups of more or less the same size, together with a third group with a minimum sample size could appear as a nice compromise, limiting the problems due to the implementation in practice, and still solving the NC problems and HC. Therefore, our next question is: what is the smallest possible sample size needed for the third group in order to solve the problems?

Since case 1 is the most problematic one, we start with it. The 500 observations are divided in different ways: first, the 2- and 3-group designs already considered before, with similar size for each group; then, different 3-group designs with unequal repartition of the observations: 49%, 49%, 2% in one case, 47.5%, 47.5%, 5% in another case, and 45%, 45% and 10% in the last case. The left part of Table 4.4 gives the results.

Table 4.4: Results for different repartitions of the 500 observations into groups

500 obs	Case 1						Case 2					
	2 group	3gps 2%	3 gps 5%	3 gps 10%	3 gps equal	true	2 gps	3 gps 2%	3 gps 5%	3 gps 10%	3 gps equal	true
Ga11	.9838	.7644	.7432	.7389	.7384	.74	.9540	.7694	.7479	.7382	.7322	.735
(SD)	(.4230)	(.1893)	(.0691)	(.0581)	(.0592)		(.5661)	(.1693)	(.1054)	(.0772)	(.0677)	
Ga41	.5840	.6922	.6976	.6998	.6994	.70	.6413	.7228	.7286	.7329	.7343	.735
(SD)	(.1680)	(.1041)	(.0787)	(.0709)	(.0600)		(.1885)	(.1346)	(.1011)	(.0852)	(.0666)	
Ga71	.7190	.8487	.8553	.8599	.8590	.86	.6426	.7203	.7260	.7319	.7325	.735
(SD)	(.1987)	(.1154)	(.0767)	(.0660)	(.0583)		(.1821)	(.1294)	(.0946)	(.0805)	(.0647)	
Ph44	-.2007	.0657	.0971	.1007	.1004	.10	.1213	.1558	.1595	.1612	.1614	.16
(SD)	(.6787)	(.2871)	(.0409)	(.0300)	(.0279)		(.1225)	(.0454)	(.0315)	(.0299)	(.0331)	
Ph55	.1263	.0616	.0595	.0581	.0591	.06	.1717	.1598	.1594	.1579	.1585	.16
(SD)	(.0875)	(.0598)	(.0411)	(.0337)	(.0282)		(.0497)	(.0440)	(.0402)	(.0380)	(.0334)	
Ph66	.1771	.0898	.0866	.0848	.0865	.09	.1706	.1572	.1564	.1549	.1560	.16
(SD)	(.1153)	(.0710)	(.0471)	(.0370)	(.0329)		(.0481)	(.0404)	(.0388)	(.0367)	(.0338)	
Number conv	266	500	500	500	500	500	348	499	500	500	500	500
Avg bias	.1592	.0133	.0029	.0015	.0013		.0777	.0114	.0055	.0028	.0022	
Avg MSE	.1771	.0543	.0341	.0330	.0327		.1168	.0520	.0470	.0450	.0443	

Results of table 4.4 are encouraging: by adding a third group with 10 observations (2%), the 500 replications become convergent and the variance of M_1 becomes in average positive. By having a third group of 25 cases (5%), the estimates are in average accurate, even in the problematic case 1. In order to co-validate this result, the same kind of simulations is also done for all other sets of values and

conditions which were qualified as poor or quite poor in Revilla and Saris (2011) and for some conditions where no problems were encountered. The results for case 2 can be found in the right part of Table 4.4, the other are not presented but the same pattern is found in all cases: already with 10 observations, the NC problem is solved and the bias is low in average.

In sum, in case the data has not been collected yet, we strongly recommend using a 3-group design. If it helps its implementation, unequal sample size for the three groups can be used with two main groups and one small group. However, it is important to notice that accurate estimates are obtained in average over 500 replications. Given the relative large standard deviations of the estimates of the parameters one can expect rather large uncertainties in the estimates which will also lead to large standard errors in the predicted values for the quality coefficients in the meta-analysis across countries.

4.7 If the data has already been collected

When the data has already been collected, adding a third group even of 10 cases is not possible. Since quite some data was collected using the 2-group design, in particular in the ESS, the next section looks for solutions to analyse properly this existing data.

4.7.1 Fix the negative variances to zero?

The classic way of dealing with HC consists in fixing to zero non significant negative estimates that should not be negative in theory. However, in the example of satisfaction in the Netherlands, we saw that different method variances fixed to zero led to models that could not be rejected but had very different estimates. It suggests that HC may be more problematic than one thinks, but also that fixing even non significant parameters may not be the proper thing to do. Nevertheless, we only looked at one example. Besides, we had no information about the true values. So it was not possible to know if one of the situations was biased whereas the other was correct or if both were biased. To investigate this point more systematically, we use simulations based on case 1 for the 2-group design, where the average estimate for the variance of method 1 is negative.

Table 4.5: Fixing method variances to zero

500 obs	Case 1			true
	2 group	2 group fix ph 44	2 groups fi ph55/ph66	
Ga11	.9838	.8627	.6642	.74
(SD)	(.4230)	(.0532)	(.0493)	
Ga41	.5840	.5995	.7788	.70
(SD)	(.1680)	(.0665)	(.0799)	
Ga71	.7190	.7347	.9552	.86
(SD)	(.1987)	(.0582)	(.0634)	
Ph44	-.2007	0	.1552	.10
(SD)	(.6787)		(.0205)	
Ph55	.1263	.1251	0	.06
(SD)	(.0875)	(.0302)		
Ph66	.1771	.1760	0	.09
(SD)	(.1153)	(.0287)		
Number conv	266	500	481	500
Average bias	.1592	.0999	.0758	
Average MSE	.1771	.0567	.0470	

Table 4.5 shows that fixing the first method variance (ϕ_{44}) to zero, all 500 replications become convergent. Besides, it leads to a PS: on average none of the parameters have prohibited values anymore (no negative variances also for the error terms). The same is true when fixing the second (ϕ_{55}) and third (ϕ_{66}) method variances, with few non convergent replications left. Nevertheless, the estimates are very different depending which method variance is fixed. Besides, both series of estimates are really biased. For example γ_{11} is 0.12 too high when the variance of M_1 is fixed to zero and 0.08 too low when the variances of M_2 and M_3 are fixed to zero. Moreover, the standard deviations are low such that the true value does not even appear to be in the confidence interval. We did not consider case 2 since they were no negative forbidden estimates.

Overall, the results suggest that fixing one method variance to zero, even when negative and non significant, is not a good solution, contrary to what is often argued and done in practice. What else can we do?

4.7.2 Add a very small third group with random data?

The next section investigates an idea derived from previous results. In section 0.2, it has been seen that using three groups, even with a third group of only 10 observations, improves in average the performance of the estimation. When the data has already been collected with a 2-group design, what would we get if we would simply invent data for this third group that we need but do not have? In light of previous results, we expect that if we use a really minimum sample size for this third group with random (invented) data, this will not harm much the estimates (small N) and at the same time will solve the problems due to the by design missing correlations.

To test this 500 datasets of 500 observations are generated in *Mplus* (Muthén and Muthén, 1998-2007) using the values of case 1. Then, five or ten fake observations for imaginary respondents that would get M_2 and M_3 are added to each dataset. In the first situation, these fake observations are completely randomly chosen and are the same for all 500 datasets. In the second situation, the values of the fake observations are inspired by values that are present in the dataset. Once the fake observations are added, the datasets are analysed. In order to see if the results are stable with different sets of true values, the procedure is repeated using the values of case 2. Table 4.6 gives the results for both cases.

Table 4.6: results using a small fake third group

500 obs	Case 1					Case 2				
	5 fake random	10 fake random	5 fake based data	10 fake based data	true	5 fake random	10 fake random	5 fake based data	10 fake based data	true
Ga11	.6903	.7245	.7848	.7639	.74	.6251	.7112	.7876	.7645	.735
(SD)	(.0694)	(.0610)	(.1674)	(.1131)		(.1210)	(.0736)	(.2977)	(.1654)	
Ga41	.7601	.7190	.6799	.6855	.70	.8995	.7629	.7271	.7257	.735
(SD)	(.0815)	(.0722)	(.1240)	(.0989)		(.1251)	(.0795)	(.1582)	(.1303)	
Ga71	.9291	.8860	.8316	.8411	.86	.8634	.7713	.7232	.7247	.735
(SD)	(.0776)	(.0678)	(.1372)	(.1053)		(.1077)	(.0763)	(.1498)	(.1267)	
Ph44	.1337	.1065	.0533	.0774	.10	.1862	.1730	.1525	.1579	.16
(SD)	(.0437)	(.0326)	(.1590)	(.1012)		(.0331)	(.0293)	(.0885)	(.0380)	
Ph55	.0251	.0636	.0688	.0673	.06	.1143	.1500	.1586	.1590	.16
(SD)	(.0359)	(.0286)	(.0734)	(.0561)		(.0492)	(.0392)	(.0465)	(.0434)	
Ph66	.0380	.0828	.0988	.0951	.09	.1251	.1413	.1563	.1563	.16
(SD)	(.0464)	(.0342)	(.0962)	(.0695)		(.0456)	(.0384)	(.0445)	(.0421)	
Number conv	361	500	471	500	500	499	500	496	496	500
Average bias	.0499	.0130	.0263	.0154		.0849	.0216	.0142	.0103	
Average MSE	.0315	.0315	.0470	.0391		.0408	.0411	.0629	.0511	

For both cases, with 10 fake observations almost all replications are convergent. With five cases, the convergence is a bit lower. Besides, even if the estimates are not perfectly accurate, for the 10 fake observations, they are in average acceptable and much better than when fixing negative variances to zero. Both cases with 10 fake observations are almost equivalent and very similar also to the case where data was generated with a 3-group design with unequal sample size groups 49%, 49% and 2% (cf. Table 4.4).

Obviously, one would however have to be careful with the interpretation of the results if following such a procedure, mainly because a good performance in average does not mean that in one specific experiment the procedure will give accurate estimates. Also in this case we see that the standard deviations are rather large which will also lead to large uncertainties in the prediction of the quality coefficient.

4.7.3 The chosen solution

In section 5.1 we have seen that the problems are solved in case of very large samples. In the ESS each experiment has been done in at least 25 countries with samples of around 1500 cases. If each country is analyzed separately there are only 1500 cases and one would have problems. But if we assume for the moment that the model is the same in all countries this means that for each experiment we have at least 37.500 cases taking all countries together. Such a sample would be enough to avoid non convergence for most of the experiments. Using Multiple group analysis of any SEM program one can estimate such a model. We have used in this case the program LISREL 8.5.

It should be clear that we do not believe that this assumption with respect to the equality of the model for all countries is correct. So the next step would be to test for misspecifications in the model for the different countries because we expect that some parameters, indicating the reliability and validity will be different in different countries. For the detection of the misspecifications in the first model we have used the program JRule (Van der Veld et al 2008) based on the work of Saris et al (2009). Using this program the model is corrected, introducing more free parameters in the different countries till the differences between the estimated values of the parameters were so small from one run of the program to the other that one could conclude that it made no sense to continue with the adjustments of the models. We used as a criterion that the differences in estimated values should be in general smaller than .02. We thought that such a difference is not of substantial importance and gives sufficient precision with respect to the estimation in the meta analysis discussed later.

The results of such a sequential process of model corrections can depend quite heavily on the first steps made in the process. Therefore we have decided that each dataset for each experiment have to be analyzed in the above way independently by two researchers. Because the two researchers can come to different results the last step is that they compare the differences and decide together which corrections have to be introduced in the model in order to get a jointly accepted result.

It turned out that this approach worked for all experiments available in the first 3 rounds of the ESS except for the media data. In the latter case this approach did not work because the correlations between the traits, use of TV, Radio and newspaper is so close to zero that the model is not empirically identified even with close to 40.000 cases. However, for all other topics this procedure worked satisfactorily in the sense that the analyses converged to a jointly accepted solution which is difficult to improve and which shows rather small standard errors for the quality estimates.

In order to evaluate the quality of this estimation procedure we have summarized in table 4.7 the correlations and the mean of the absolute differences between the

resulting parameter estimates for different persons and groups for each of the experiments in round 4.

Table 4.7 The correlations between the parameter estimates for the different persons and groups.

experiment	Phase 1			Phase 1 adjusted			Comparison of groups		
	Who	corr	mean diff	Who	corr	mean diff	Who	corr	mean diff
Imbgeco	1 - 2	.9951	.0325	1 - 2	.9999	.0009	g1- g2	.9845	.0479
Imsmetn	1 - 2	.9898	.0238	1 - 2	.9999	.0033	g1-g2	.9775	.0522
Dngval	3 - 2	.9929	.0329	3 - 2	.9999	.0044	g1-g2	.9944	.0206
Lrnnew	3 - 2	.9979	.0281	3 - 2	.9999	.0016	g1-g2	.9672	.0876

While we had some doubt concerning the quality of the estimations of individual analysts, two independent analyses were done by different persons (1,2,3 and 4). However, in table 4.7 we see that the result of the different individual analysis was not so bad. The correlations between the obtained parameter estimates were between .9898 and .9979 and the mean differences between the parameter values were between .0238 and .0329. This is, of course, a very good result. In order to avoid idiosyncratic estimates we asked the two analysts to look at each other's final results and try to find a common solution for relative large differences. In doing so the correlations between the remaining results were .9999 for all topics and the mean differences reduced to maximally .0044. Finally, to be completely sure about our results we did the whole procedure once more but this time other analysts did the analysis and created a group's result. For example, for the experiment Imbgeco now analysts 3 and 4 did the analysis and adjusted their results. The results of this new group got were compared with the results of the first pair of analysts. To our surprise the correlations between the groups were lower than between individual analysts and the mean differences were larger. This result suggests that there is still some arbitrariness in the analyses. The larger differences can only be explained if the two groups go different routes in the improvement of the model and create a basic model which is somewhat different so that for many parameters differences occur. One can see nevertheless that in general the similarity of the quality estimates is very similar over all countries. We have decided at some point to use the results of the first pair of individual analysts as the final results and derived from these two sets a point estimator of all quality indicators for all countries. The description of these results will be described in the next chapter.

Here we have to say one thing more. It has been possible to use this approach because a program was developed that orders the data of 25 countries and 2 or 3 groups, runs the analyses, picks up from the huge output the essential information for the researcher, and stores this information in such a way that the program can pick up the information from different analyses for comparison. This comparison program was used by the two researchers in order to determine whether their solutions were so similar that they did not have to continue with the adjustments. The programs which make these analyses possible are for free available for other users¹⁶.

¹⁶ The programs have been developed by Daniel Oberski and Thomas Gruner see the appendix

4.8 Conclusion

Given the serious problems we detected in the analysis of the SB-MTMM data of each country for all experiments, a very elaborate study was done to detect the reasons of these problems and the solutions to these problems. In this process many possibilities were tested. Finally, we decided to estimate the reliability and validity coefficients of all questions in all countries and experiments by a two-step procedure: first, a model was estimated assuming that the quality coefficients were the same in all countries. So the initial estimates were based on at least 37.500 cases. After that, two researchers independently checked which quality coefficients had to be estimated independently of the others because they were indicated to be different, the two researchers created together a joint solution for all questions in all experiments.

Appendix

Due to fact that certain choices in this process are arbitrary and may lead to differing quality estimates, each analysis was done independently by two different analysts, after which the analysts compared their results and introduced incremental model changes until no further improvements could be found in each of their models.

This yielded a very complex analysis with very many steps and comparisons of different versions of the experiments of an analysis for the four different analysts. To make this procedure possible, a computer application program “MAC” (see below) was developed.

The program allows the analyzers to input the LISREL model syntax, run it, and obtain outputs and a comparison of the quality estimates with previous versions or other analyzer’s versions. Each run of an analysis was stored using the version control system *git* (2009). The analyzers could also view the exact differences (“diff”) between their model syntax and that of another version or analyzer, as well as obtain a side-by-side comparison of the quality estimates. This allowed them to pinpoint the exact model changes that may have led to any differences in estimates. An online repository of this history, combining the repositories of all analyzers, is available. Below the programs for these tasks have been mentioned.

Developers	Project name	URL	Purpose	Technologies
D. Oberski	SQL coding program	http://sqp.nl/	Allow coders to code all different SQL characteristics into a database, Allow for comparison between different choices; Export codes to dataset; Automatic codes for some characteristics using Natural Language Processing	Python 2.7, Django, Treetagger, Hunspell
D. Oberski, T. Gruner	MTMM Analysis Comparison and Archiving (MAC)	https://github.com/recsm/automtmm/	Keep a full git repository of each analyzer’s analysis history automatically; Allow analyzers to edit and run LISREL analyses from within MAC; Parse LISREL outputs and display and compare estimates in the program.	Git, Adobe Air, Python, LISREL
T. Gruner, D. Oberski	MTMMArchive	https://github.com/recsm/MTMMArchive	Online git repository of all MTMM analyses of all analyzers with their full analysis history.	Git, github

Chapter 5

The variation in quality of the questions across countries and methods used

Diana Zavala

Melanie Revilla

Laur Lilleoja

Willem Saris

The previous chapter has shown which solution has been found for the estimation problems of the SB-MTMM experiments. Having found this solution the data could be analyzed with the suggested procedure without too many problems. The results of these analyses are reliability, validity and quality estimates of all questions involved in MTMM experiments. These quality indicators have been added to the database of ESS questions. These quality estimates can be used by scholars in the analysis of ESS data to correct for measurement error. How this can be done in a simple way is illustrated in chapter 8. Detailed information can be found in Saris and Gallhofer (2007, chapter 15). In this chapter we will give some results with respect to the quality of the questions and so illustrate why it is necessary to correct for measurement errors.

In this chapter we concentrate on the results obtained in the first three rounds of the ESS. In the data base of questions more questions from previous MTMM experiments are available.

First we look at the distribution of the quality of the questions across the ESS questions. In total 2460 questions have been evaluated in round 1 – 3 of the ESS. The distribution of the quality is presented in Figure 5.1.

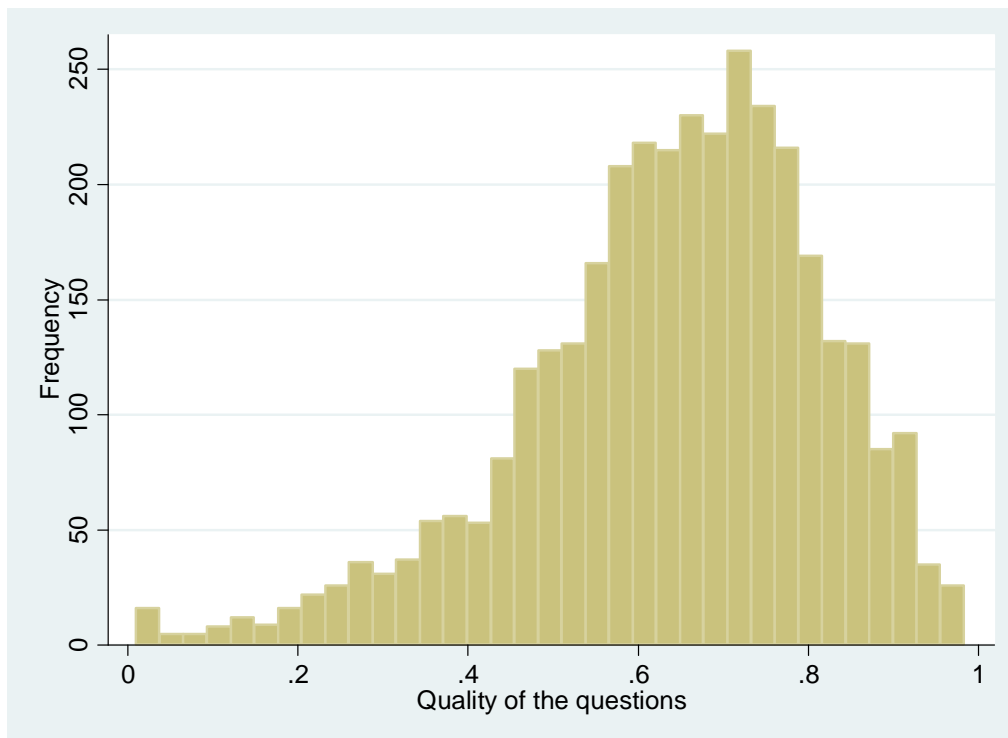


Figure 5.1 The distribution of the quality of the questions

The mean quality is not so bad (.64) and higher than the average result reported by Alwin (2007) with respect to studies in the US using a different model. In average 64% of the observed variables is explained by the latent variables of interest. In the US this was 50%. We also see there are questions where the quality is very low (<.40). One may wonder whether these questions are good enough to be used.

5.1 Differences in quality across countries

Important is whether the quality of the questions is not too different across countries because if that is the case one can't compare relationships between variables across countries. In table 5.1 we present the average quality of the questions for all countries which participated.

Table 5.1 The mean quality of the questions in the countries

Country	Mean total quality	Standard deviation	Minimum	Maximum
Austria	0.645	0.158	0.10	0.98
Belgium	0.603	0.172	0.02	0.93
Czech Republic	0.632	0.162	0.03	0.87
Denmark	0.621	0.188	0.02	0.92
Estonia	0.604	0.183	0.02	0.89
Finland	0.624	0.178	0.03	0.96
France	0.578	0.188	0.02	0.96
Germany	0.637	0.156	0.03	0.95
Greece	0.691	0.163	0.12	0.96
Ireland	0.594	0.180	0.03	0.95
Netherlands	0.672	0.183	0.01	0.98
Norway	0.647	0.143	0.24	0.94
Poland	0.631	0.180	0.03	0.96
Portugal	0.662	0.169	0.26	0.97
Slovenia	0.616	0.168	0.03	0.92
Slovakia	0.585	0.190	0.04	0.92
Spain	0.622	0.172	0.11	0.96
Sweden	0.670	0.123	0.39	0.88
Switzerland	0.651	0.181	0.03	0.97
United Kingdom	0.630	0.160	0.03	0.92
Ukraine	0.606	0.167	0.18	0.96
Total	0.639	0.175	0.01	0.98

The mean quality for all countries is 0.64 and the standard deviation is 0.175. The country with the highest mean quality is Greece (0.69) and the country with the lowest mean value is France (0.58). The difference is not that large. On the other hand we see that the variation in quality within all countries is rather large. This suggests that probably large differences can be seen if we look at different topics within each country. But that could mean that the averages are relatively comparable but the quality of specific questions can be very different across countries.

So we looked next at the differences in quality for some questions across the countries and concentrate on the questions in the main questionnaire. The question with the least variation in the quality across countries is an important question about immigration which is asked in the last round as part of the core.

B37 STILL CARD 14 How about people from the poorer countries outside Europe? Use the same card.

- Allow many to come and live here 1
- Allow some 2
- Allow a few 3
- Allow none 4
- (Don't know) 8

The results for the question with the least variation in quality are presented in table 5.2. This table shows that the variation is indeed very minimal. The lowest value is in the Ukraine (.69) and the highest in Switzerland (.84).

Country	Mean quality
Austria	0.823
Belgium	0.717
Denmark	0.826
Estonia	0.717
Finland	0.748
France	0.787
Germany	0.796
Ireland	0.741
Netherlands	0.732
Norway	0.781
Poland	0.773
Portugal	0.814
Slovakia	0.717
Slovenia	0.717
Spain	0.797
Switzerland	0.847
Ukraine	0.686
United Kingdom	0.748
Overall mean	0.765
Standard deviation	0.046

For a correlation of .6 between two variables in both countries with these qualities it would already mean that in the Ukraine the observed correlation¹⁷ would be $.6 \times .69^2 = .414$ and the correlation in Switzerland would be $.6 \times .84^2 = .504$. This possible difference in observed correlation would not be a substantial difference. This difference would completely be a consequence of the difference in data quality as found in the ESS. Note that in both cases the correlation between the observed variables would be considerably lower than the true correlation between these variables (.6).

The largest variation has been found for the question of the core about government intervention. The question is formulated as follows:

¹⁷ This result is based on the equation 2.1 where $q^2 = (r.v)^2$ assuming that the method effect is minimal.

CARD 16 Using this card, please say to what extent you agree or disagree with each of the following statements. **READ OUT EACH STATEMENT AND CODE IN GRID**

B43GinvEcoThe less that government intervenes in the economy, the better it is for [country]

1. strongly agree
2. agree
3. agree neither disagree
4. disagree
5. strongly disagree

The qualities of the questions across the different countries of this question are presented in table 5.3.

Table 5.3. Quality across countries of item with largest variation	
Country	Mean quality
Austria	0.638
Belgium	0.360
Czech Republic	0.590
Denmark	0.359
Germany	0.383
Finland	0.337
France	0.365
Greece	0.687
Ireland	0.352
Netherlands	0.400
Norway	0.362
Poland	0.601
Portugal	0.943
Slovenia	0.364
Spain	0.746
Sweden	0.434
Switzerland	0.380
United Kingdom	0.408
Overall mean	0.484
Standard deviation	0.175

In this case the differences are indeed much larger. The lowest value is .337 in Finland and the highest .943 in Portugal. Fortunately not all differences are so large because this would lead to very large differences in observed correlations even though the correlation variables would be the same. To illustrate this with a correlation of .6 between the latent variables, this would mean that in Finland the correlation would be $.6 \times .337^2 = .20$ and in Portugal $.6 \times .943^2 = .56$. This is just a consequence of differences in the size of the measurement errors.

Because the size of the measurement errors has such a big effect on correlations and other measures for relationships these quality estimates have been estimated. They can be used to correct for measurement errors as we will show in chapter 8.

5.2 Differences in quality for Domains and Concepts

Saris and Gallhofer (2007) have shown that there are significant differences in the validity and the reliability –and as consequence in the quality— for items from different domains, concepts and other associated characteristic.

Domain	Mean total quality	Standard deviation	Minimum	Maximum
Health	0.473	0.151	0.06	0.89
Living conditions and background variables	0.586	0.114	0.21	0.81
Work	0.661	0.138	0.31	0.97
Personal relations	0.587	0.144	0.09	0.87
Leisure activities	0.533	0.118	0.28	0.78
National politics: national government	0.705	0.120	0.24	0.96
National politics: national institutions	0.760	0.081	0.07	0.93
National politics: economic/financial matters	0.595	0.181	0.14	0.96
National politics: other	0.685	0.136	0.38	0.96
Total	0.640	0.154	0.06	0.97

Table 5.4 below shows differences depending on the domain resulting of questions involved in the MTMM experiments¹⁸. Items asking about ‘health’ have the lowest quality 0.473. Questions about politics vary depending on the specific topic, items on national institutions and national government reported the overall highest quality, 0.760 and 0.705 respectively while in items about ‘economic or financial matters’ the quality was much lower 0.595.

Concept	Mean total quality	Standard deviation	Minimum	Maximum
Norm	0.723	0.110	0.43	0.96
Policy	0.703	0.151	0.39	0.96
Action tendency	0.691	0.090	0.38	0.91
Feeling	0.676	0.126	0.07	0.96
Evaluative belief	0.606	0.171	0.06	0.97
Facts, background or behaviour	0.527	0.100	0.34	0.78
Evaluation	0.527	0.129	0.18	0.89
Total	0.640	0.154	0.06	0.97

¹⁸ It should be said that the differences in quality between the different categories of a explanatory variable can also come from other characteristics which are related with this variable. Therefore a multivariate analysis would give a better indication of the effect of the variable domain. This picture will be given in the next chapter.

Table 5.5 shows that there are differences in the quality depending on the concept measured. Norms and policies have the highest quality while items about facts, background or behaviour and evaluations reported the lowest quality.

5.3 Effect of the question formulation on the quality

It can also be expected that question formulation has an effect on the quality of the questions. In an earlier paper it was reported by Saris et al (2010) that Agree/disagree batteries had a very negative effect on the quality. Without repeating the full report here we can illustrate this here with two examples. In round 3 experiments were done to study these effects. These experiments have been summarized in table 5.6.

In the second experiment, concerning consequences of immigration the first measures were item specific scales while the repetition in the 3 subgroups were Agree/disagree batteries with different numbers of categories.

Table 5.6 Round 3: The SB-MTMM experiments						
Exp	Var.	Meaning	Main = M1	gpA = M2	gpB = M3	gpC = M4
2	<i>imbgeco</i>	- It is generally bad for [country's] economy that people come to live here from other countries	11IS	5AD	11AD	7AD
	<i>imulect</i>	- [Country's] cultural life is generally undermined by people coming to live here from other countries				
	<i>imwbcnt</i>	- [Country] is made a worse place to live by people coming to live here from other countries				
3	<i>imsmet</i>	- [Country] should allow more people of the same race or ethnic group as most [country's] people to come and live here.	4IS	5AD	4IS	7AD
	<i>imdfctn</i>	- [Country] should allow more people of a different race or ethnic group from most [country's] people to come and live here.				
	<i>impcntr</i>	- [Country] should allow more people from the poorer countries outside Europe to come and live here.				

Note: is=item specific scale, ad= agree/disagree scale, batt=battery, fix=fixed reference point

In the immigration experiment 4 methods have been used. Three of them are collected in the supplementary questionnaire in randomly assigned subgroups of the samples in each country. This means that the data with these three methods have been collected at the same point in time and under the same conditions in randomly assigned groups. The results of this experiment are presented in table 5.7.

In table 5.7 we have presented the results with respect to the average quality across the participating countries of experiment 3 comparing an Item specific scale with two Agree/disagree scale all three measured in randomized subgroups of the total sample in the supplementary questionnaire.

Table 5.7 The quality of the questions concerning immigration

	Question 1	Question 2	Question 3
IS 4	.905	.914	.908
AD 5	.568	.629	.607
AD 7	.525	.597	.562

The table indicates the enormous difference in quality between the different scale types where the Item specific scales (IS) turn out to have much higher quality than the two agree/disagree scales (AD). This result is in agreement with an earlier publication of Saris et al. (2009).

Another result that has been found before (Revilla 2011) is that the 5 point agree/disagree scale has a higher quality than a 7 points and an 11 point scale as can be seen in Table 5.8 based in experiment with respect to the consequences of immigration (imbgeco).

Table 5.8 The quality of the questions concerning consequences of immigration

	Question 1	Question 2	Question 3
AD 5	.576	.649	.649
AD 7	.352	.462	.490
AD 11	.267	.413	.452

These examples show very clearly the effect the choice of the method can have on the quality of the questions.

5.4 Conclusions

In this chapter we have seen that the differences in quality across countries for some questions can be large. As a consequence one cannot compare relationships between variables across countries without correction for the quality of the questions. It is for this reason that the ESS has decided to include MTMM experiments in the standard operations of the ESS. In chapter 7 we will show how the information about the quality of questions can be obtained and in chapter 8 we will show how relationships can be estimated correcting for measurement error.

We have also seen that the choices made in the design of the questionnaires can have considerable effect on the quality of the questions. Therefore we will show in chapter 7 how the quality of questions can be predicted before the data are collected and how the new program SQP2.0 can provide suggestions for improvement of the questions.

Chapter 6

The prediction of the quality of the questions

Daniel Oberski

Thomas Gruner

Willem Saris

As has been shown in the previous chapters we created a database of questions containing characteristics of the questions, the response options, introductions, and showcards, as well as characteristics of the questionnaire and the data collection method. From the MTMM experiments we obtained quality measures for the questions: the reliability, validity and quality coefficients, which we added to the characteristics database. In total 3011 questions are available in the database. Using all this information about the questions, a prediction model was estimated using random forests of regression trees (Breiman 2001).

One advantage of this approach is that is able to generate good predictions based on a large number of correlated features, and allows for possible interactions, insofar as they are estimable. The procedure also provides prediction intervals and standard deviations for the predictions. In this chapter the development of the prediction model will be explained. This model is implemented in a new version of the Survey Quality Predictor (SQP 2) application program, which will be presented in the next chapter.

The database of MTMM experiments consists of 87 old experiments used to develop SQP1 and 15 new experiments done in more than 25 countries. Only those experiments corresponding to questions coded were analyzed. In total we have both the question characteristics and the quality estimates for 3483 questions (1051 unique method-trait combinations) in different languages/countries. The MTMM analyses done separately for each country were unstable and it was decided to stabilize them by introducing cross-country equality restrictions in certain parameters. Afterwards these restrictions were tested against the observed data by examination of the modification indices and expected parameter changes (see chapter 5). As was mentioned there, the analyses were done by two researchers with the aid of specially developed software.

After the analyses for all experiments were done and stored in the online git repository, reliability and validity estimates were extracted from the repository and written to a plain text file using a script written in Python. In this way we obtained two quality estimates for each question: one for each of the two analyzers who had separately analyzed the data. We then combined these two estimates in the following way.

First the estimates (reliabilities and validities) were logit-transformed. To combine the estimates from the different analyzers, we then estimated a random effects model using the logit transformed estimates with item-country combination as a random factor¹⁹. Within-analyzer variance was found to be negligible and removed from the model. Overall, 97% percent of the variance in reliability estimates was estimated to be due to the item-country combination. There was thus some variance across analyzers, though it was relatively small.

From the random effects analysis we obtained point estimates of the logit-transformed estimates and their standard errors, taking into account between-analyzer variance. These were used to construct point estimates and 95% confidence intervals for the original reliability and validity estimates. The size of these confidence intervals (difference between upper and lower bounds) ranged between 0.0009 and 0.1363 for the

¹⁹ Since one language per country was analyzed, these might equally well be labeled “item-language” combinations.

reliability coefficients and between 0.0009 and 0.2793 for the validity coefficients. The overall average reliability coefficient estimate was 0.841 and the overall average validity coefficient was 0.923.

Point estimates and intervals of these estimates were stored in the database of questions coded with SQP. For information about the way we dealt with missing data we refer to Appendix B. Histograms of the reliability and validity estimates and their logit transformations are shown in figures 6.1 and 6.2.

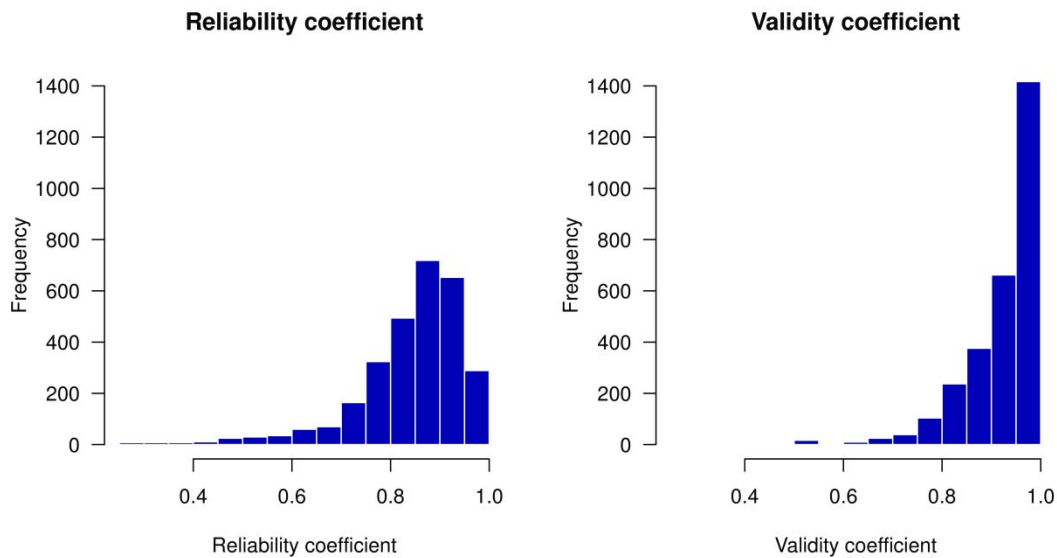


Figure 6.1 Reliability and validity coefficient estimates obtained from the MTMM experiments, without transformation.

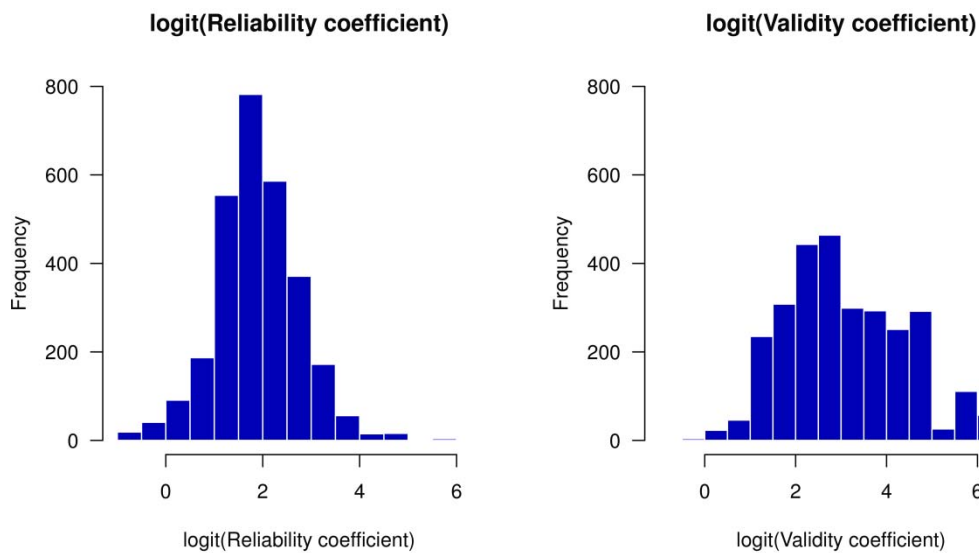


Figure 6.2 Reliability and validity coefficient estimates obtained from the MTMM experiments after logit transformation.

After estimating the reliability and validity of the coded questions in the previous step, we obtained a database in which question characteristics were joined with the reliability and validity estimates. After deletion of questions that were either not coded or not analyzed, a data set with 3483 questions was obtained.

Figures 6.3, 6.3 and 6.4 show the codes obtained for the characteristics “domain”, “concept”, and “number of categories” (for categorical questions). These tables are intended to give the reader an impression of the range of topics covered, and the type of questions analyzed. Unsurprisingly given the topics covered in the European Social Survey, most questions measure subjective variables: attitudes, opinions, etc., measured with categorical scales. Though there are some factual questions and frequency scales, these clearly constitute the minority of questions.

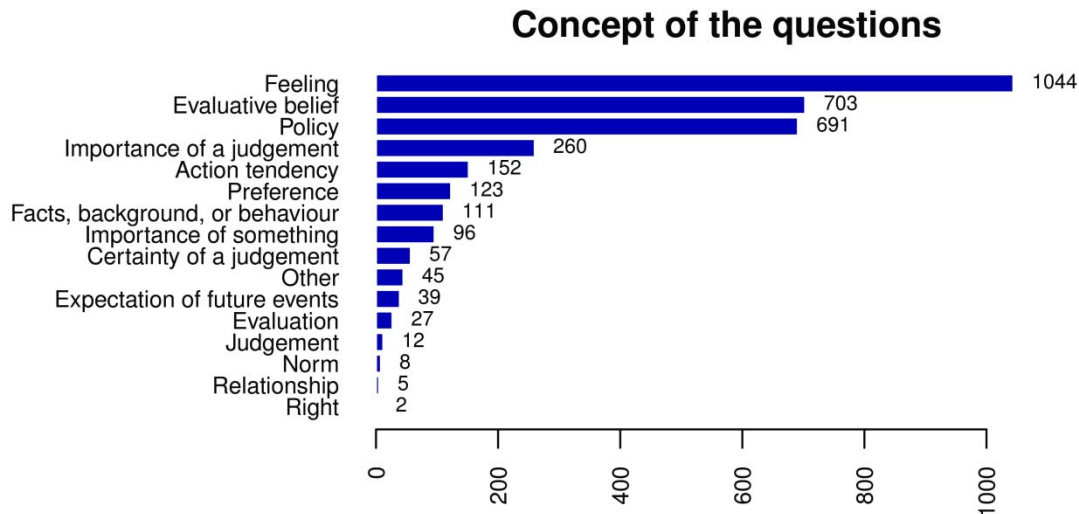


Figure 6.3 The frequency distribution of the different concepts in the sample

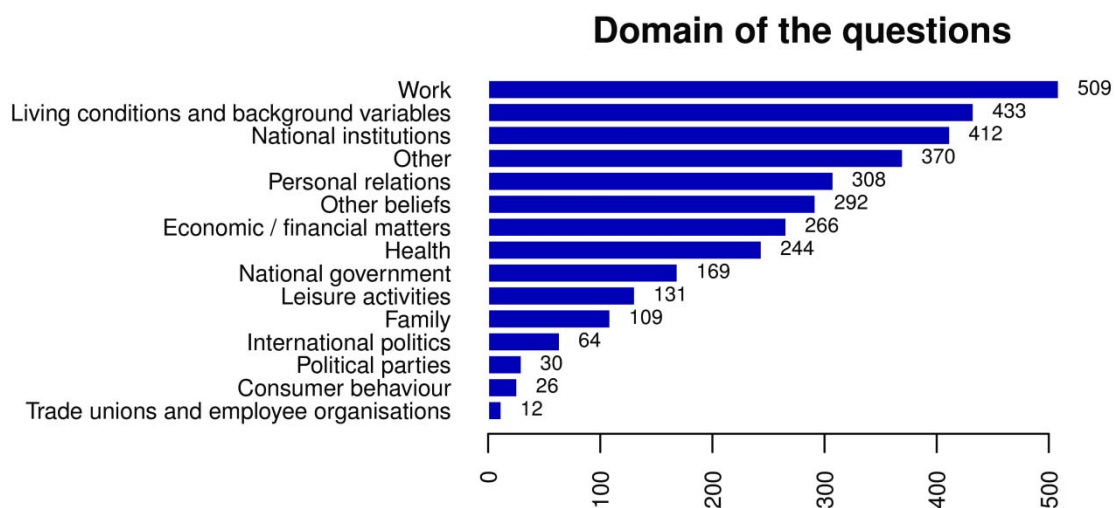


Figure 6.4 The frequency distribution of the different domain topics in the sample

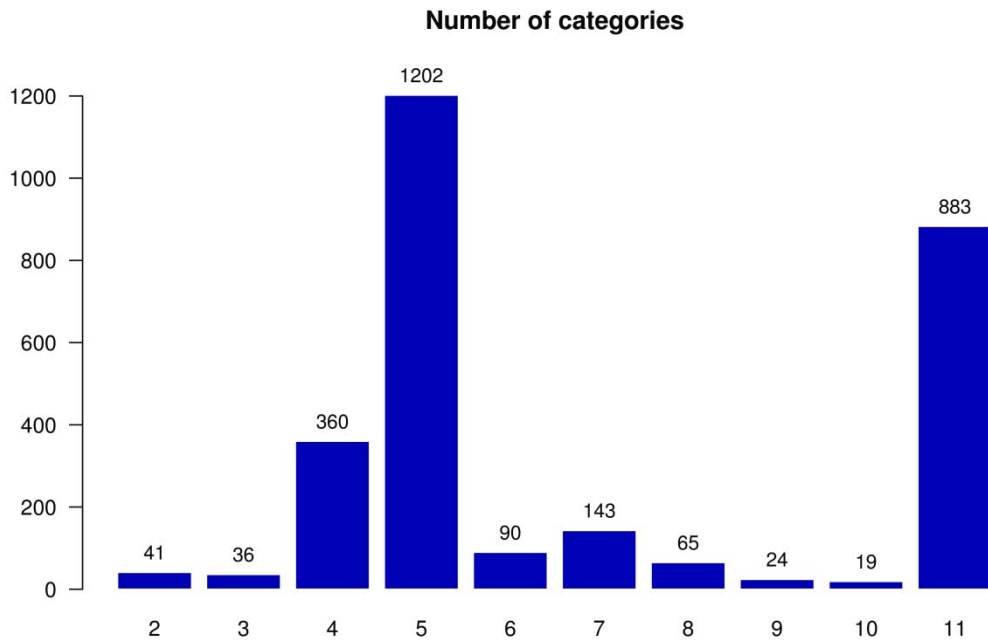


Figure 6.5 The frequency distribution of the different category scales in the sample

6.1 The Meta-analysis

We fitted separate prediction models for the logit-transformed estimates of the reliability and validity coefficients (r and v). Predictors were obtained using Breiman’s (2001) random forests of regression trees, as implemented in the *R 2.13.1* package *randomForest* (R Development Core Team 2011; Liaw & Wiener 2002). A random forest is an ensemble predictor, that is, a collection of many individual predictors whose individual predictions are combined to form the final prediction. In random forests, the individual predictors are regression trees grown with the CART algorithm. In our case we used 1500 trees for each of the two models for $\text{logit}(r)$ and $\text{logit}(v)$.

The ensemble is formed by taking, for each tree, a bootstrap sample with replacement of questions, so that some questions are included in the sample or “in bag”, and others are excluded or “out-of-bag”. On average over the entire forest, a question was out-of-bag about 184 ± 11 times - that is, it is *not* present in about 12% of the trees in the forest. The trees are not only random in the sense of the observed question distribution, but also in the sense of the variables (“features”) selected for inclusion in the tree growth algorithm: for each analysis, 20 out of the 62 meta-variables are selected at random (without replacement).

Each of the regression trees is grown on one of the bootstrapped datasets in the following manner. The dataset is split into two groups (“nodes”) based on that split on a question characteristic which yields the smallest possible mean squared prediction error for the $\text{logit}(r)$ or $\text{logit}(v)$. For each new group the same procedure is repeated until the resulting group would have 5 or fewer observations or no improvement in mean square prediction error can be found. This algorithm is known as the CART algorithm.

In practice CART trees may suffer from overfitting problems. Their predictive power can be limited, and this has led to pruning techniques, whereby the lower nodes of the tree are removed from the predictor so as to prevent overfitting. In the random forest algorithm, a different approach is taken. Instead of growing just one regression tree, many trees – in our case 1500 – are grown without pruning, but based on a double randomization of both observations and variables used in the prediction. This deals with

the overfitting problem by subsuming randomness due to overfitting in the between-tree variance, and automatically using those features that are commonly selected in all bootstrap samples to determine the average, and final, prediction

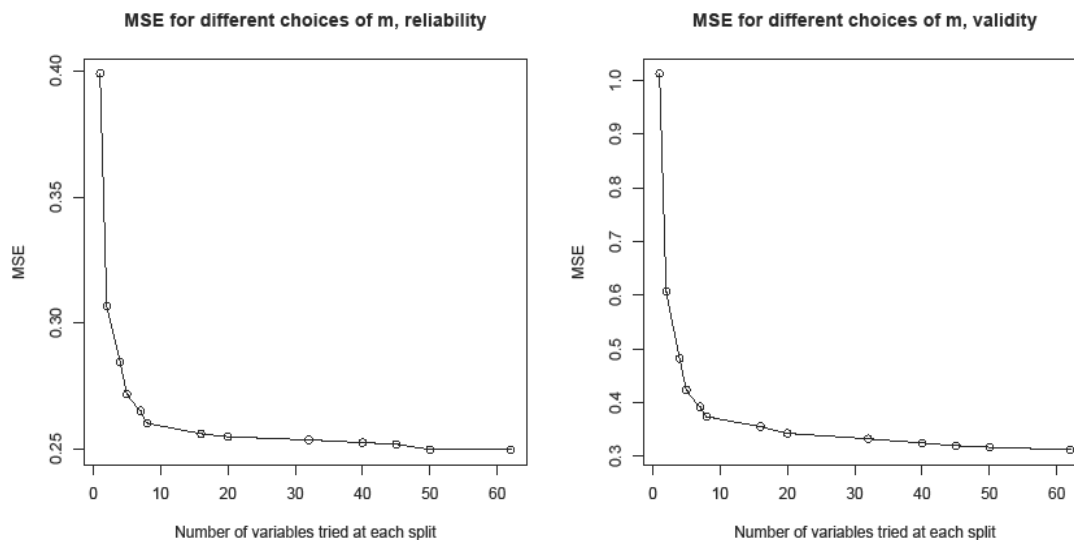


Figure 6.6 Mean squared prediction error of the random regression forest predictor for different choices of the number of features (m).

A key parameter in the random forest algorithm is the number, m , of variables selected at random for each tree. On the one hand, growing trees with more features gives more predictive power to each tree, which will reduce the mean squared prediction error. On the other hand, increasing the number of features will increase the correlations between tree predictions, reducing the mean squared prediction error, or requiring more trees to obtain predictions of the same accuracy. Figure 6.6 plots the estimated out-of-bag mean squared prediction error for reliability and validity coefficients for different choices of the number of randomly selected features m .

It can be seen that the mean squared error is not reduced much further after 20 features, which is the default chosen by the *randomForest software*. We therefore chose to retain this default choice i.e. $m=20$ in our approach.

6.2 Quality prediction

The final prediction obtained from the random forest ensemble is the mean of the predictions of the individual trees. An example of a single regression tree is given in figure 6.7. The tree in figure 6.7 gives a prediction of the logit of the reliability coefficient for a question with given characteristics. For example, suppose the question “Do you think the government does a good job?” is asked, with answers ranging from “the worst job” to “quite a good job”. The top node splits off depending on the domain, in this case national government (domain=101). Afterwards we follow the split on concept to the left-hand node, because the question asks an “evaluative belief” (concept = 1). The average number of syllables per word is then the next relevant variable, which in this case is 1.22; less than 1.5. Finally, there is only one fixed reference point, so that the prediction of the logit value ends up being 2.2. This prediction was based on 34 observations. This logit value can be transformed in a value of a reliability coefficient by taking the inverse of this value or $\text{invlogit}(2.2) = 0.90$. So according to this tree the reliability would be predicted to be 0.90. The final prediction from the random forest is then the average of 1500 such predictions.

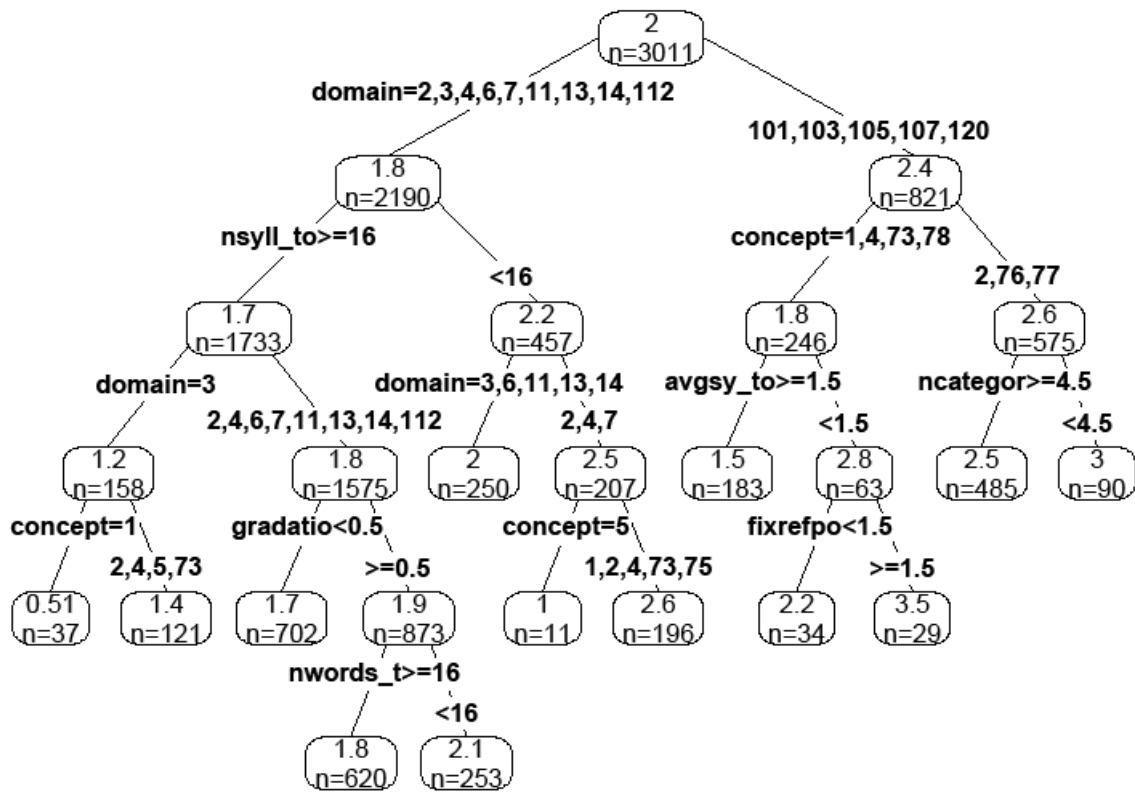


Figure 6.7 Example CART tree for prediction of the logit-transformed reliability coefficient.

The tree given in figure 6.7 is only given as an example output of the CART procedure, and does not necessarily correspond to any tree used in the final prediction. In fact 1500 of these trees are created and the overall prediction is then taken as an average over all 1500 trees in the “forest”. Given the amount of information available in all these trees a distribution of the predictions is obtained and this information can be used to determine the means and specify prediction intervals and standard deviations.

6.3 Suggestions for improvement of questions

Looking once more at the tree in Figure 6.7, we can see that a much higher prediction of $\text{invlogit}(3.5) = 0.97$ would have been given if there had been more fixed reference points, for example if the final category had not been “quite” but “the very best”. It can be seen by looking at the terminal nodes that a large range of different predictions can be obtained depending on the characteristics of the question.

Given the available ensemble predictor one can, for each predicting variable, vary the code and see what the effect would be on the predicted quality. In this “what-if” analysis one can get a mean prediction for each possible code of the variable, keeping all other codes the same. Some of these predicted values may be lower than the predicted value of the real question but others may be higher. In this way one can get an impression of what improvement in the prediction is possible by changing this characteristic of the question or the study while keeping all other characteristics the same. However we speak purposely of “impression” because in general one characteristic of a question can not be changed without also changing other characteristics of the question. For example, increasing the number of categories will change the number of words and syllables, and possibly also the instruction or even the labelling of the scale, etc. So one has to be careful with these suggestions. A more adequate procedure is to reformulate the question and check the prediction of the new

question. In addition, it should be kept in mind that the current model is only a prediction model and not a causal model. Therefore there is no guarantee that actually changing this characteristic will have the predicted effect on the quality.

So far we spoke of only one prediction variable. Looking for the possible improvements can already be very tedious for one variable if this variable has many categories which all have to be checked separately. This can be a rather lengthy process. Therefore it makes sense to consider which variables are the most important ones for the predictions.

6.4 Variable importance

There has been a discussion in the literature about the way to determine which prediction variables are the most important. For details of this discussion we refer to Appendix D. We have chosen the conditional approach. Figure 6.8 gives the conditional variable importance measures for predictions of the validity coefficient, and figure 6.9 does the same for the reliability coefficient. These graphs might be taken as being of interest for the future exploration of relative importance various factors may have in affecting the validity and reliability coefficients.

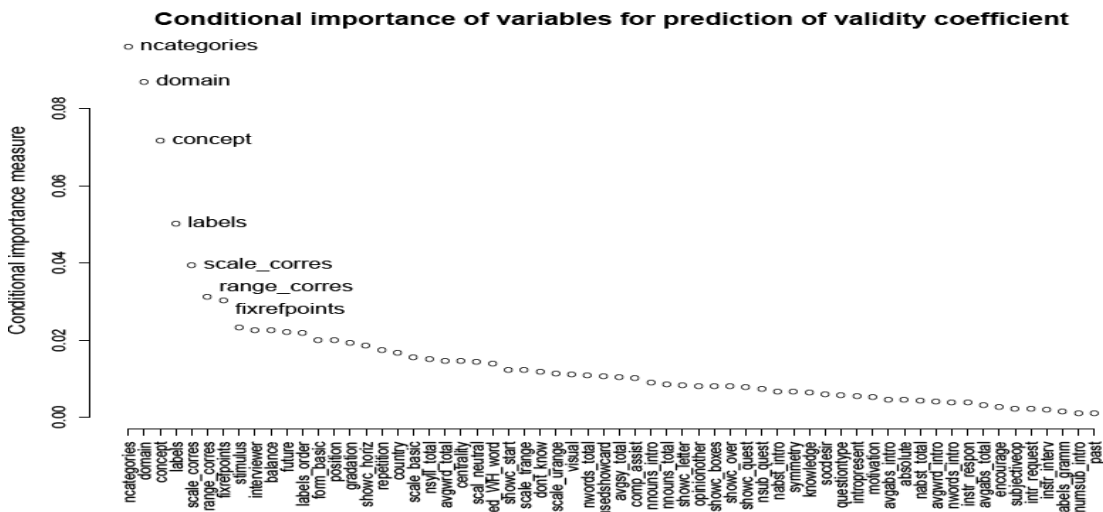


Figure 6.8 Conditional (“unbiased”) importance measures of the prediction variables for the prediction of the validity coefficient, ordered by importance.

It is interesting to note that some characteristics are important for both random (reliability) and systematic errors (validity), while others seem to act more on one or the other. For example, it is clear that both quality measures vary greatly by the topic (domain and concept). However, also three survey design characteristics are important in both predictions: “labels” (fully, partial, or none), “scale_corres” (a recode of unipolar/bipolar scales, see above), and “range_corres” (whether the numbers on the labels correspond to the direction of the meaning). The number of fixed reference points

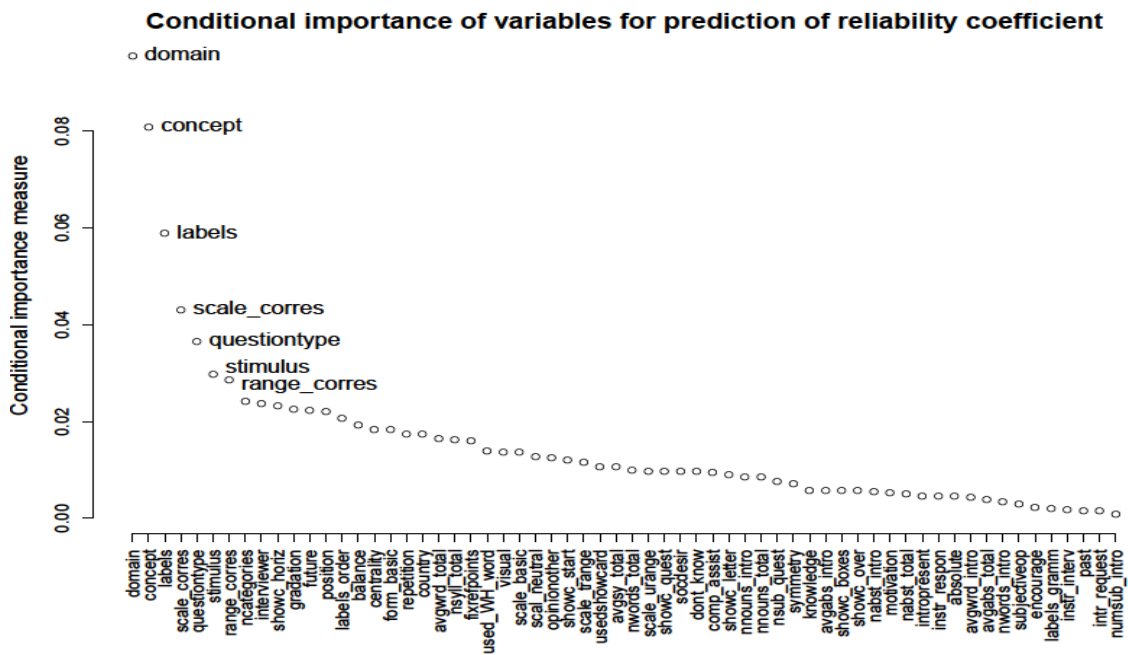


Figure 6.9 Conditional (“unbiased”) importance measures of the prediction variables for the prediction of the validity coefficient, ordered by importance.

and number of categories appears to be more important for predicting the validity coefficient than for predicting the reliability coefficient. Conversely, stimulus (used in batteries) and question type are more important in the prediction of the reliability coefficient.

We have used the importance of the different variables to reduce the computation time for the evaluation of possible improvements of the questions. In doing so we used the following rules:

1. The variables which are directly related to the trait measured and can't be changed will be ignored. These variables are: country, domain, concept, future, past and present, social desirability and centrality.
2. We have selected the most important 20 variables from the two figures 6.8 and 6.9, starting with the common variables and adding the most important single predictors.
3. In the calculations one gets firstly the result for the first 20 variables. After that one can also ask for all the other ones

The first decision reduces the number of variables for which computations have to be done from 53 to 45. One may wonder whether position should be included as well but we did not do so because the position can indeed be changed.

The second decision was made because the quality is determined by the combination of reliability and validity. Therefore important predictors for both should be taken into account

The third decision was made in order to provide the user more quickly with the results for the variables which in general have the most effect. However, because the results can be different for different questions we also allow for further information about the possible effect of the other 25 variables.

The three decisions together led to the following list of 20 predictors which are probed first in the calculations:

- Ncategories
- Labels
- Scale corres
- Range corres
- Fixrefpoints
- Stimulus
- Interviewer
- Balance
- Labels order
- Form basic
- Position
- Gradation
- Showc_horizon
- Scale basic
- Nsyll_total
- Avgwrld_total
- Scal_neutral
- Used WH word
- Visual
- Showc start

For the meaning of these variables we refer to Appendix A

6.5 Evaluation of the quality of the prediction models

For each tree, the mean square error of the predictions from the tree is calculated using only questions that are out-of-bag for that tree. After growing the entire forest, the prediction error for the overall forest is calculated by combining the out-of-bag prediction error estimates. Thus, the mean square prediction error estimate is automatically based on cross-validation samples.

From these mean squared error estimates one can calculate an R^2 measure of the predictive power of the forest as a whole. The R^2 was 0.84 and 0.65 for the validity (v) and reliability coefficient (r) logits, respectively. The squared correlations between predicted and observed coefficients on the original scales were 0.69 for the reliability coefficients and 0.72 for the validity coefficients. Figures 6.10a and 6.10b show predictions of reliability and validity coefficients (on the original scale) versus residuals.

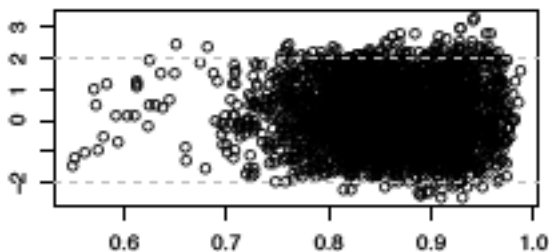


Figure 6.10a Fitted vs. residuals, reliability coefficient.

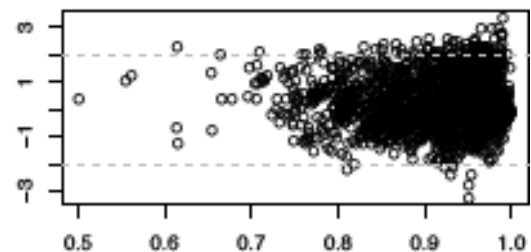


Figure 6.10b Fitted versus residuals, validity coefficient.

It can be seen that there is a scarcity of questions with very low reliabilities. There are also three outliers in the prediction of the validity coefficients. Besides these features no general pattern is visible.

Another prediction of interest is the product $q^2 = r^2 \times v^2$, also known as the “quality” of the question. The scatter plot of predicted versus observed values for the quality is shown in figure 6.11.

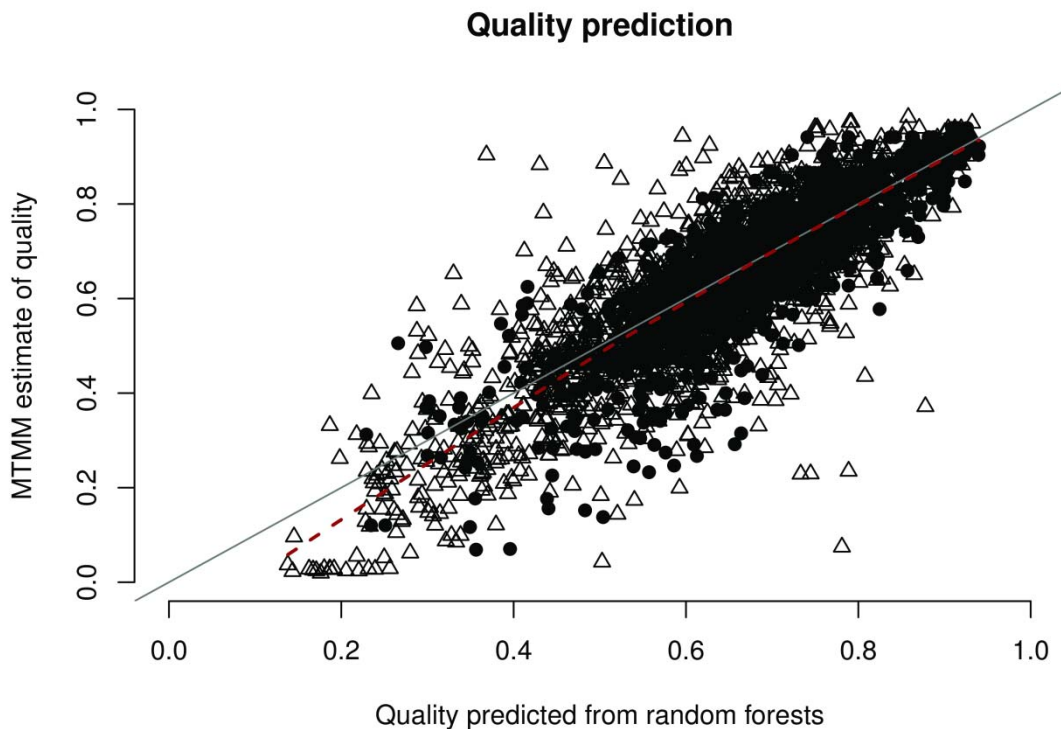


Figure 6.11 Predicted quality (q^2) versus observed quality. The triangles indicate questions from the ESS, the dots questions from the old experiments. A loess smoother (dotted line) is plotted alongside the 45 degree line of unbiasedness.

The prediction does the reasonably good job that can be expected based on the high R^2 measures for validity and reliability coefficients. However, for the few questions with low quality parameters, the predictions systematically too high, as shown by the deviation of the dotted line from the gray 45 degree line.

Overall, we believe that the predictor does a reasonably good job of providing information about the expected quality of a question, with the caveat that the prediction worked less well for the few questions with a very low quality (more than 60% measurement error). When employing the predictor to obtain quality predictions of questions that were not in this study, it should also be remembered that the questions in the dataset cover only a certain range of application.

6.6 Conclusions

The procedure for prediction of the quality of questions is considerably different from the previous procedure used to make the predictions for SQP1.0. In that case the model was a regression equation based on the absolute values of the reliability and the validity. The reason for the change is that the new procedure gives better predictions and avoids the problem of unacceptable predictions, larger than 1 or smaller than 0. Another advantage of this new procedure is that we do not only provide a point estimate but also a prediction interval. Finally, the new program is based on a much larger

database of questions than the older version and incorporates questions from a diverse range of topics, languages, and countries.

In the next chapter we will present the SQP 2.0 program. This program allows the user to code the question characteristics in a user-friendly interface, and provides predictions of the reliability and validity estimates based on the random forest predictors. The program also allows for a direct comparison of the results of the predictions with the results of the MTMM experiments that are available in the database. The results of the present procedure are indeed much better than using SQP 1.0. The explained variances for reliability and validity were in the past respectively .47 and .61; with the new prediction procedure the explained variance increased to respectively .60 and .85. This is a considerable improvement. It should be said that the predictions will never be perfect because some questions may be so different that the database does not contain sufficient similar questions. This holds at this moment especially for questions about facts, frequencies, and events.

Appendix A: Obtaining the SQP codes

The coding program was developed by Oberski (2010). Automatic codes were used for: no. words, sentences, syllables (via Hunspell morphological analyzer (Németh 2005), <http://hunspell.sourceforge.net/>), nouns (via Treetagger <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>) (Schmid 1994).

Non-automatic codes were obtained by: Teams, training, consensus coding, quality control on codes

A list of characteristics coded in the SQP program is given below.

Codes collected in a meta-data file by the program. This file was then cleaned.

It was decided that only questions in countries that participated in all rounds of the ESS should be coded, and only in their main language (the language spoken by most people in the country).

Additional meta-variables were added for information about the data collection modes and position in the questionnaire by hand.

Meta-variables for non-ESS experiments were also available. These were all recoded into the newer coding system used by the ESS program. The ESS cleaned dataset and the older experiments recoded into the new system were then joined together to form a new meta-dataset of questions and their characteristics.

In the SQP coding program, a splitting rule implied a very detailed categorization of the domain and concept of the question could be obtained. Upon inspection of the codes, it was found that most of these new categories were empty. Therefore the domain was collapsed to only the main domain choices, with a split only for “national politics”. Similarly, for concept the categorization was restricted to the main concept choices except that a split on “other simple concepts” was added.

A newly coded variable was the so-called “scale correspondence”. This variable is formed from the characteristics “unipolar/bipolar underlying scale” and “unipolar/bipolar response scale”. The codes for the “scale correspondence” characteristic were determined as shown in the table below.

Table A1 Coding of the "scale correspondence" characteristic.

	Range of the conceptual scale	
Range of the response scale	Unipolar	Bipolar
Unipolar	1	2
Bipolar or n/a	-	3

Below is the list of all variables used in the meta analysis

Characteristic	Type
Domain [domain]	Categorical
Domain: national politics [natpoldomain]	Categorical
Domain: European politics [dom_european]	Categorical
Domain: international politics [intpoldomain]	Categorical

Domain: family [dom_family]	Categorical
Domain: personal relations [dom_personal]	Categorical
Domain: work [dom_work]	Categorical
Domain: consumer behaviour [dom_consumer]	Categorical
Domain: leisure activities [dom_leisure]	Categorical
Domain: health [dom_health]	Categorical
Domain: living conditions and background variables [dom_backgrou]	Categorical
Domain: other beliefs [dom_other]	Categorical
Concept [concept]	Categorical
Social Desirability [socdesir]	Categorical
Concept: other simple concepts [conc_simple]	Categorical
Concept: complex concept [conc_complex]	Categorical
Centrality [centrality]	Categorical
Reference period [ref_period]	Categorical
Formulation of the request for an answer: basic choice [form_basic]	Categorical
WH word used in the request [used_WH_word]	Categorical
Use of stimulus or statement in the request [stimulus]	Categorical
'WH' word [WH_word]	Categorical
Request for an answer type [questiontype]	Categorical
Use of gradation [gradation]	Categorical
Balance of the request [balance]	Categorical
Presence of encouragement to answer [encourage]	Categorical
Emphasis on subjective opinion in request [subjectiveop]	Categorical
Information about the opinion of other people [opinionother]	Categorical
Absolute or comparative judgment [absolute]	Categorical
Response scale: basic choice [scale_basic]	Categorical
Number of categories [ncategories]	Numeric
Don't know option [dont_know]	Categorical
Number of frequencies [nfrequencies]	Numeric
Maximum possible value [scale_max]	Numeric
Labels of categories [labels]	Categorical
Theoretical range of the scale bipolar/unipolar [scale_trange]	Categorical
Labels with long or short text [labels_gramm]	Categorical
Number of fixed reference points [fixrefpoints]	Numeric
Range of the used scale bipolar/unipolar [scale_urange]	Categorical
Interviewer instruction [instr_interv]	Categorical
Respondent instruction [instr_respon]	Categorical
Extra motivation, info or definition available? [motivation]	Categorical
Introduction available? [intropresent]	Categorical
Knowledge provided [knowledge]	Categorical
Number of sentences in introduction [nsents_intro]	Numeric
Number of sentences in the request [nsents_quest]	Numeric
Number of words in introduction [nwords_intro]	Numeric
Number of subordinated clauses in introduction [numsub_intro]	Numeric
Request present in the introduction [intr_request]	Categorical
Number of words in request [nwords_quest]	Numeric
Total number of nouns in request for an answer [nnouns_quest]	Numeric
Total number of abstract nouns in request for an answer [nabst_quest]	Numeric
Total number of syllables in request [nsyll_quest]	Numeric
Number of subordinate clauses in request [nsub_quest]	Numeric
Number of syllables in answer scale [nsyll_ans]	Numeric
Total number of nouns in answer scale [nnouns_ans]	Numeric
Total number of abstract nouns in answer scale [nabst_ans]	Numeric
Show card used [usedshowcard]	Categorical
Horizontal or vertical scale [showc_horiz]	Categorical
Overlap of text and categories? [showc_over]	Categorical
Numbers or letters before the answer categories [showc_letter]	Categorical

Scale with numbers or numbers in boxes [showc_boxes]	Categorical
Start of the response sentence on the showcard [showc_start]	Categorical
Question on the showcard [showc_quest]	Categorical
Picture on the card provided? [showc_pict]	Categorical
Neutral category [scal_neutral]	Categorical
Symmetry of response scale [symmetry]	Categorical
Order of the labels [labels_order]	Categorical
Correspondence between labels and numbers of the scale [scale_corres]	Categorical

Appendix B Imputation

The meta-analysis dataset contained reliability and validity estimates for questions, as well as question design characteristic codes provided by the coders and the automatic coding program. Not all characteristics were coded for all questions, however. Particularly, a series of question design characteristics of the showcards were added to the codes after the “old” experiments had already been coded. For these “old” experiments there was therefore no information on showcards. In addition, there were some instances of questions that had not been completely coded for one reason or another. Therefore the meta-analysis dataset has missing data.

We wished to deal with the missing data, without increasing the apparent precision of the final prediction artificially. For this reason we chose to multiply impute the missing data using the chained equation approach of (van Buuren and Groothuis-Oudshoorn (2011), as implemented in their R package *mice*. Multiple imputation of missing data was conditional on all other design characteristics, and 3 randomly imputed datasets were obtained.

We then performed the random forest analysis separately for each multiply imputed dataset, obtaining 3 separate sets of 500 trees. The 500 trees were then combined into one single prediction ensemble of 1500 trees. For this reason the prediction intervals obtained from the random forest ensemble also take into account the uncertainty in the imputations (Rubin 1987). For a more detailed description of this approach, see Nonyane & Foulkes (2007).

Appendix C The Software developed for SQP 2.0.

Developers	Project name	URL	Purpose	Technologies
D. Oberski, T. Gruner, GUI design by M. Cassidy	SQP 2.0	http://devel.sqp.nl/	New user interface to the coding program; Provide predictions with prediction intervals of question quality - reliability, validity, common method variance – based on characteristics choices; Display point estimates obtained from MTMM analyses; Provide what-if-scenarios showing the effect of a change in characteristic on the prediction.	Python 2.7, Django, AJAX libraries, Pyro, r2py2
D. Oberski	SQP prediction engine	(available upon request)	Provide raw predictions from raw question characteristic codings using randomForest objects.	R 2.13.1

Appendix D the estimation of importance of prediction variables

The random forest procedure also provides so-called “variable importance” measures. These are marginal deteriorations in mean square prediction error when the information in a particular variable is removed. The “importance” of a variable is calculated by randomly permuting the observed values of that variable and then recalculating the out-of-bag mean square error of predictions. If the reduction in mean square error is large, the importance is said to be high (Breiman 2001). This measure is sometimes called the “permutation importance”.

One issue with this approach is that it does not take into account the correlation between different predictive variables. Trivially, for example, the total number of syllables in the question and the average number of syllables per word will be highly correlated. Breiman’s variable importance measure will give both a similar importance measure. If these are both high, this should not be interpreted to mean that *both are indispensable* characteristics of the question for prediction of the quality. It may still very well be that using only one of them would give equally good predictions. In short, the variable importance measure is marginal, not conditional. Presumably, though, either one or the other or both are important in the predictive sense, and the marginal variable importance measures are still useful for this purpose.

A second issue to note with the variable importance measures obtained from regression trees is that, since there are many more possible splits for variables with many categories, the more categories a variable has, the more often it will be split upon, i.e. the more “important” it will be. This is not necessarily a problem, as it conveys simply that variables with more categories contain more information. Other authors have criticized these measures on these two grounds, however.

On the grounds that Breiman’s variable importance measures are marginal, Hothorn et al (2006) criticized these measures and spoke of “bias”. That is, Breiman’s measures are biased as measures of the expected deterioration in the model predictive power if the variable were left out of the analysis entirely. To counter this “bias”, they proposed forests of conditional regression trees (cforests). We fitted forests of conditional regression trees to our dataset as well, using the R package *party* (Strobl et al. 2008). It should be noted that the predictions for quality coefficients obtained from these conditional random forests correlated 0.98 with the predictions obtained using the original algorithm. As can be expected, however, the variable importance measures were very different. The marginal permutation importance measures for the models used to predict reliability and validity coefficients are shown in figures 7 and 8.

Figure 9 gives the conditional variable importance measures for predictions of the validity coefficient, and figure 10 for the reliability coefficient. These graphs might be taken as being of interest for the future exploration of relative importance various factors may have in affecting the validity and reliability coefficients. They do not necessarily provide information on the functioning of the prediction implemented in the random forest predictor described above.

It is interesting to note that some characteristics are important for both random and systematic errors, while others seem to act more on one or the other. For example, it is clear that both quality measures vary greatly by the topic (domain and concept). However, also three survey design characteristics are important in both predictions: “labels” (fully, partial, or none), “scale_corres” (a recode of unipolar/bipolar scales, see above), and “range_corres” (whether the numbers on the labels correspond to the direction of the meaning). The number of fixed reference points and number of categories appears to be more important for predicting the validity coefficient than for

predicting the reliability coefficient. Conversely, stimulus (agree-disagree-type scales) and question type are more important in the prediction of the reliability coefficient.

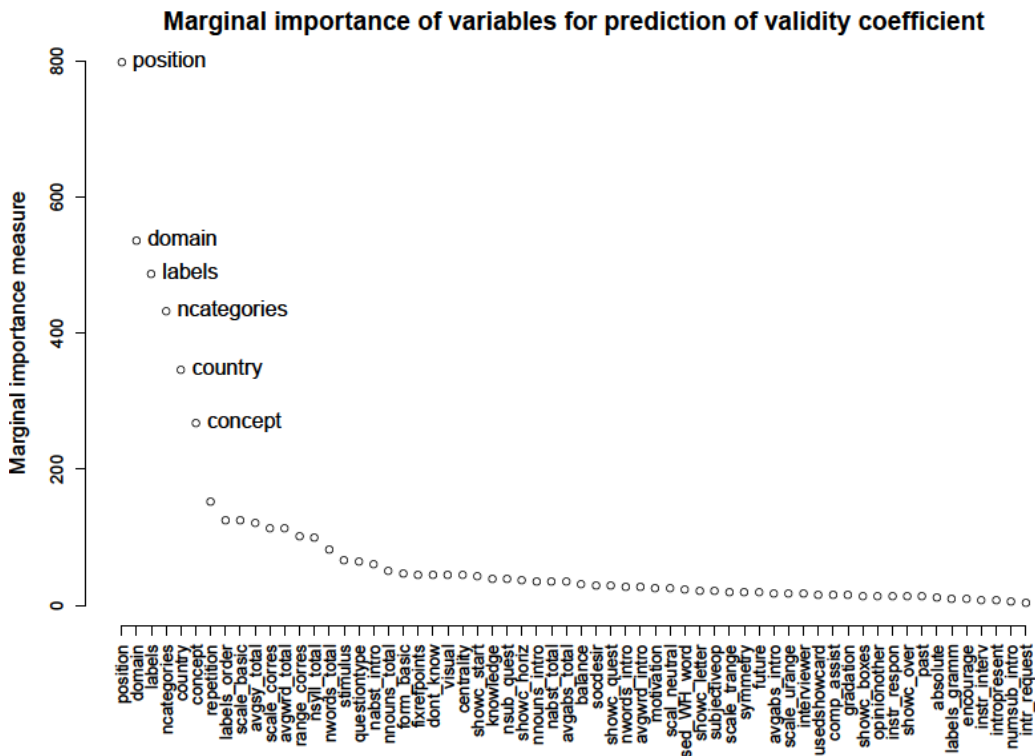


Figure 6D.1 Marginal importance measures for the validity coefficient prediction.

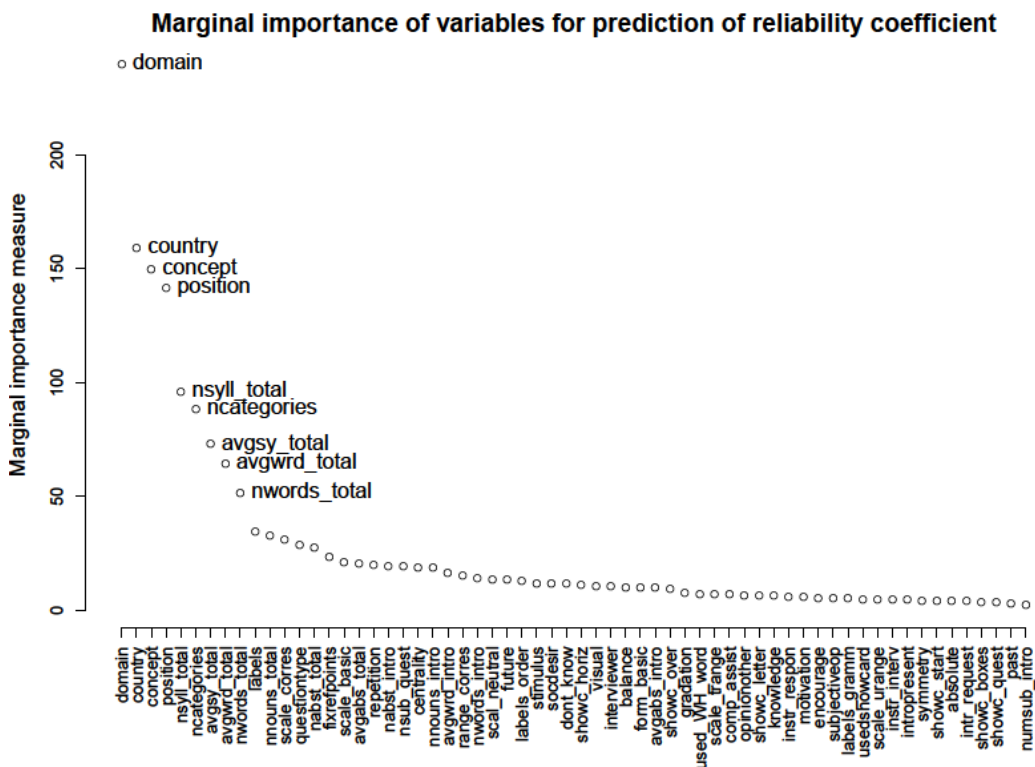


Figure 6D.2 Marginal importance measures for the reliability coefficient prediction.

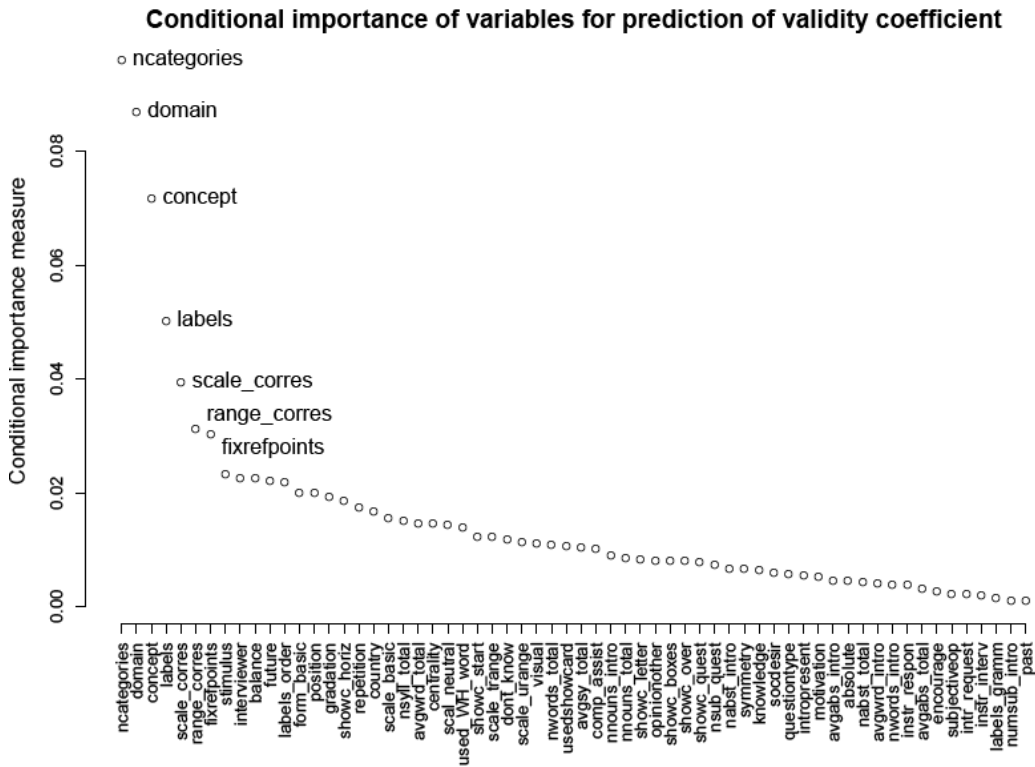


Figure 6D.3 Conditional (“unbiased”) importance measures of the prediction variables for the prediction of the validity coefficient, ordered by importance.

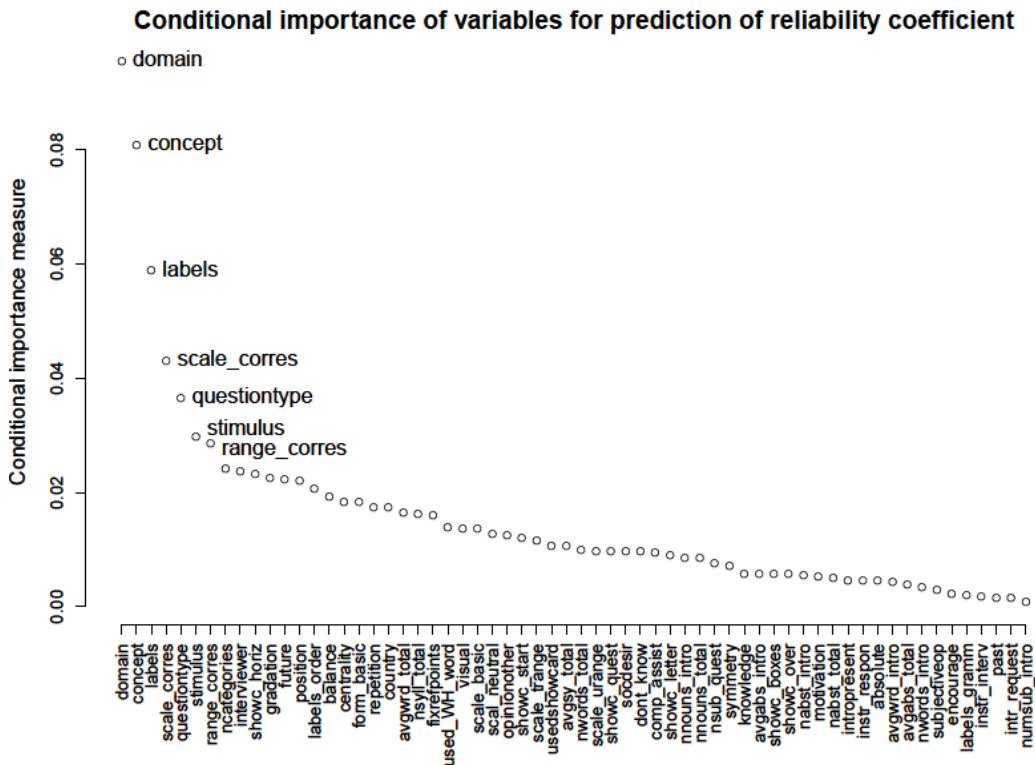


Figure 6D.4. Conditional (“unbiased”) importance measures of the prediction variables for the prediction of the validity coefficient, ordered by importance.

Chapter 7

The program SQP 2.0 for prediction of quality of questions and its applications

Daniel Oberski

Thomas Gruner

Willem Saris

The results of the last chapter have been used to develop a new version of the program SQP for the prediction of the quality of questions. SQP 2.0 has several advantages above version 1. The most important advantage is that this program is based on many more questions from many more countries. As a consequence the new program can make predictions of the quality of questions in many different languages. Another advantage of the new program is that the program not only provides point estimates but also confidence intervals for the predictions. Furthermore the program provides in a simple way suggestions for improvements of the questions. Finally a technical advantage is that the estimation is based on the logit of the quality coefficients so that the predictions can never exceed the value 1 what was sometimes the case in the old program if the questions were specified with a combination of optimal characteristics. Last but not least the predictions are considerably better than those of SQP1.0

There are in principle three different ways in which the new version of the SQP program can be used. The first option is directed to questions which were involved in MTMM experiments. In chapter 5 we have shown that the question data base contains at this moment all questions which have been involved in the MTMM experiments of the rounds 1-3 of all countries which participated in the ESS plus the questions which have been studied in the past (Saris and Gallhofer 2007). For all these questions the quality is available in the data base. Soon the set of questions will be extended with the questions of round 4 and 5. For these questions the quality estimates will be available but the coding of the questions will follow later. The program SQP can be used to obtain these quality estimates.

The second option is directed to questions of the ESS which have not been involved in MTMM experiments. In the future all other questions not involved in an MTMM experiment but asked in the ESS will be added to the data base. It will be clear that for these questions no quality estimates are available. Therefore, in order to obtain these estimates the user of the program has to code the characteristics of the question and the program provides the estimates of the quality of the questions.

The third option is directed to questions which are formulated for new studies. It will be clear that in that case the user first has to introduce the questions in the system before the coding can be started and the program can provide the prediction of the quality and suggestions for improvement of the question.

In the next pages we will discuss these different option in the sequence indicated above. However before we discuss the different option we will first introduce some basic steps to start up the program. If one goes to the internet and selects .SQP.nl one gets to see the home page of SQP. If one clicks on start the program opens the first page of the program. On that page the program asks you to register as a user. So if this is your first use of the program, click on "register now" and answer the questions that follow. If you remember your user name and password you can next time go directly

through this step by entering your user name and password and click on Login. If this is done you end up on the screen presented in Figure 1.

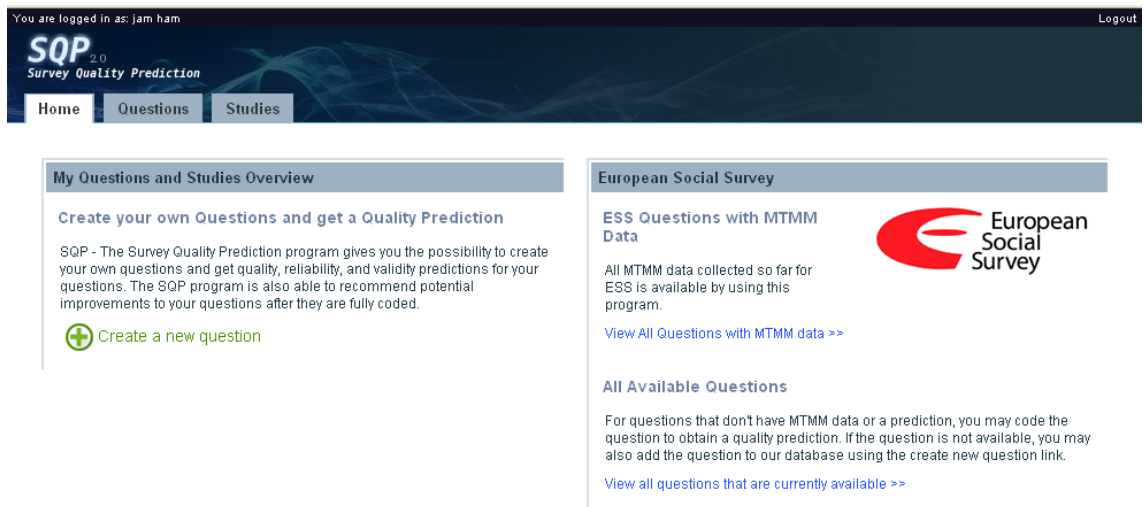


Figure 7.1 The Home page of the SQP program where one can make the basic choice what one wants to do

As can be seen the home page allows making the choices which we have suggested above. In the next sections we will illustrate what can and should be done if one makes each of the specified choices. We start with the choice of the MTMM questions.

7.1 The quality of questions involved in the MTMM experiments

If we select “View all questions with MTMM data”, by clicking on this text, we end up in the next screen presented in Figure 7.2. On this screen the user can make a selection for specific questions. The study, the language and the country of the questions can be specified by making a selection of the buttons at the top left. One can even ask for a specific question by typing the number or name of the question

You are logged in as: jam ham Logout

SQP
Survey Quality Prediction

Home Questions Studies

Home > MTMM Questions

Filter Questions

Show Questions From:

All Studies

All Languages

All Countries

Containing Text:

Selection Criteria:

Only with Predictions

Only with MTMM

My Questions

+ Add New Question

Key

My Questions and Codings

Authorized Predictions

Other User Predictions

MTMM Data Available

Question	Study	Language	Country	Quality
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	German	Austria	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Dutch	Belgium	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Czech	Czech Republic	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Danish	Denmark	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	French	France	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	German	Germany	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Greek	Greece	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	English	Ireland	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Hebrew	Israel	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Dutch	Netherlands	M
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Norwegian	Norway	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Polish	Poland	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Portuguese	Portugal	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Slovene	Slovenia	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Spanish	Spain	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Swedish	Sweden	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	English	Great Britian	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	Finnish	Finland	M U
A8 / PPLTRST / MOST PEOPLE CAN BE TRUSTED OR YOU CANT BE TOO CAR	ESS Round 1	German	Switzerland	M
A9 / PPLFAIR / MOST PEOPLE TRY TO TAKE ADVANTAGE OF YOU, OR TRY T	ESS Round 1	German	Austria	M

Showing questions 1 to 20 of 2703 total

Figure 7.2 The first page 20 MTMM questions of different countries

Imagine that we want to look at questions of Round 3 of the ESS, asked in Ireland. Then can do so by the specifications presented in Figure 7.2.

You are logged in as: jimtraud

SQP
Survey Quality Prediction

Home Questions Studies

Home > MTMM Questions (ESS Round 3, Ireland, English)

Filter Questions

Show Questions From:

ESS Round 3

English

Ireland

Containing Text:

Selection Criteria:

Only with Predictions

Only with MTMM

My Questions

+ Add New Question

Key

My Questions and Codings

Authorized Predictions

Other User Predictions

MTMM Data Available

Question	Study	Language	Country
B35 / IMSMETI / ALLOW MANY/FEW IMMIGRANTS OF SAME RACE/ETHNIC GROUP AS MAJORITY	ESS Round 3	English	Ireland
B37 / IMPCITR / ALLOW MANY/FEW IMMIGRANTS FROM POORER COUNTRIES OUTSIDE EUROPE	ESS Round 3	English	Ireland
B38 / IMBGECO / IMMIGRATION BAD OR GOOD FOR COUNTRY'S ECONOMY	ESS Round 3	English	Ireland
B39 / IMUECLT / COUNTRY'S CULTURAL LIFE UNDERMINED OR ENRICHED BY IMMIGRANTS	ESS Round 3	English	Ireland
B40 / IMWBCHT / IMMIGRANTS MAKE COUNTRY WORSE OR BETTER PLACE TO LIVE	ESS Round 3	English	Ireland
E26 / LRHIEW / LOVE LEARNING NEW THINGS	ESS Round 3	English	Ireland
E27 / ACCDING / FEEL ACCOMPLISHMENT FROM WHAT I DO	ESS Round 3	English	Ireland
E28 / PLPRFTR / LIKE PLANNING AND PREPARING FOR FUTURE	ESS Round 3	English	Ireland
E40 / DIGVAL / FEEL WHAT I DO IN LIFE IS VALUABLE AND WORTHWHILE	ESS Round 3	English	Ireland
E45 / FLCLPLA / FEEL CLOSE TO THE PEOPLE IN LOCAL AREA	ESS Round 3	English	Ireland
HS1 / testb1 / [Country] should allow more people of the same race or ethnic group as most [country's] people to come and live here	ESS Round 3	English	Ireland
HS2 / testb2 / [Country] should allow more people of a different race or ethnic group from most [country's] people to come and live here	ESS Round 3	English	Ireland
HS3 / testb3 / [Country] should allow more people from the poorer countries outside Europe to come and live here	ESS Round 3	English	Ireland
HS4 / testb4 / It is generally bad for [country's] economy that people come to live here from other countries	ESS Round 3	English	Ireland
HS5 / testb5 / [Country's] cultural life is generally undermined by people coming to live here from other countries	ESS Round 3	English	Ireland
HS6 / testb6 / [Country] is made a worse place to live by people coming to live here from other countries	ESS Round 3	English	Ireland
HS7 / testb7 / I love learning new things.	ESS Round 3	English	Ireland
HS8 / testb8 / Most days I feel a sense of accomplishment from what I do.	ESS Round 3	English	Ireland
HS9 / testb9 / I like planning and preparing for the future.	ESS Round 3	English	Ireland
HS10 / testb10 / I generally feel that what I do in my life is valuable and worthwhile.	ESS Round 3	English	Ireland

Showing questions 1 to 20 of 46 t

Figure 7.3 The first 20 questions asked in Ireland and involved in MTMM experiments.

By this selection the number of questions is considerably reduced to 46. Let us say that we want to see the results for a specific one, B38. I can type the number in the text box but I can also directly click on the question in the screen. If I do so I get the screen of Figure 7.4.

You are logged in as: imtraud

SQP
Survey Quality Prediction

Home Questions Studies

Home > MTMM Questions (ESS Round 3, Ireland, English) > B38 / IMBGECO / IMMIGRATION BAD OR GOOD FOR COUNTRY'S ECONOMY

Filter Questions

Show Questions From:

ESS Round 3

English

Ireland

Containing Text:

Selection Criteria:

Only with Predictions

Only with MTMM

My Questions

[Add New Question](#)

Key

- My Questions and Codings
- Authorized Predictions
- Other User Predictions
- MTMM Data Available

Question

B38 / IMBGECO / IMMIGRATION BAD OR GOOD FOR COUNTRY'S ECONOMY

ESS Round 3 Ireland - English

Request for Answer Text:
Would you say it is generally bad or good for Ireland's economy that people come to live here from other countries? Please use this card.

Answer options:

- 00 Bad for the economy
- 01
- 02
- 03
- 04
- 05
- 06
- 07
- 08
- 09
- 10 Good for the economy

Information	Quality	Options
MTMM Estimate	0.557	View MTMM Results >>
Authorized Prediction	0.596	View Prediction Detail >>
My Quality Prediction		Code question to create my own quality prediction >>

Showing questions 1 to 20 of 461

Figure 7.4 The question B38 of Round 3 in English of Ireland

The popup screen presents how the questions was formulated and it indicates what information is available for this question. First of all , we see that quality of the question estimated in the MTMM experiment is given which is .557. This means that a bit more than 56% of the variance in the observed variable comes from the variable that it should measure. It also means that close to 44% of the variance is error.

Sometimes there are also other estimates of the quality of the question available, predictions based on the coding of the question and the prediction program discussed in the last chapter. MTMM questions are often coded and therefore a prediction of the quality by the program can also be obtained. This is also true in this case. The so called “authorized” prediction is .596. This prediction is called “authorized” because it is based on the coding of this question which has been checked on correctness by our research team at RECSM.

We see that in this case the predicted values are not very different from the value obtained by the MTMM experiment. In order to get more information about the quality of the question especially splitting the quality up in reliability and validity. This can be done for the MTMM results by clicking on “View MTMM Results” but one can also click on View prediction details. In that case one gets the details of the MTMM and the SQP predictions results. Choosing the latter one gets the screen of Figure 7.5.

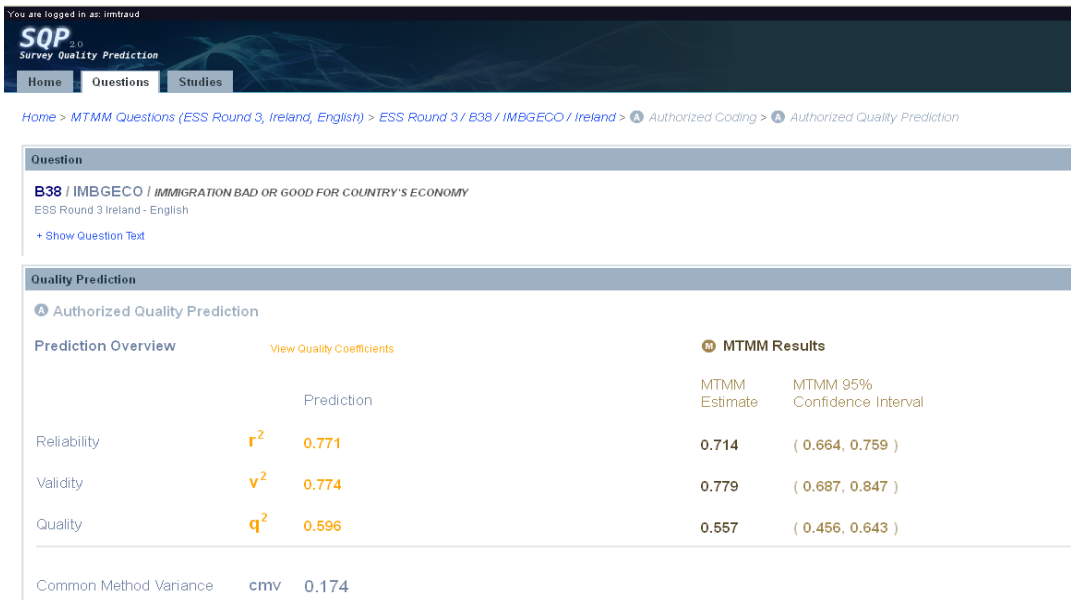


Figure 7.5. Detailed information about the quality indicators of question B38 from Ireland.

We see that in this case the estimates by MTMM and the predictions by SQP2.0 are rather similar. In general this can be expected given the high correlation between these two estimates reported in the last chapter but there are exceptions because occasionally questions can be deviant for the most common questions or because the analysis has led to a rather deviant result.

On the screen is also presented the Common Method Variance (CMV). That is an estimate of the correlation that the method would produce between variables which measure the same variables and have the same quality. In this case one can say that due to the method used the correlations would be .174 too high. How this information with respect to the data quality can be used in data analysis to correct for measurement errors will be discussed below in section 7.4.

In order to get a different picture of the quality, one can also ask for the quality coefficients by clicking on “View quality coefficients”. By doing so we get the screen of Figure 7.6

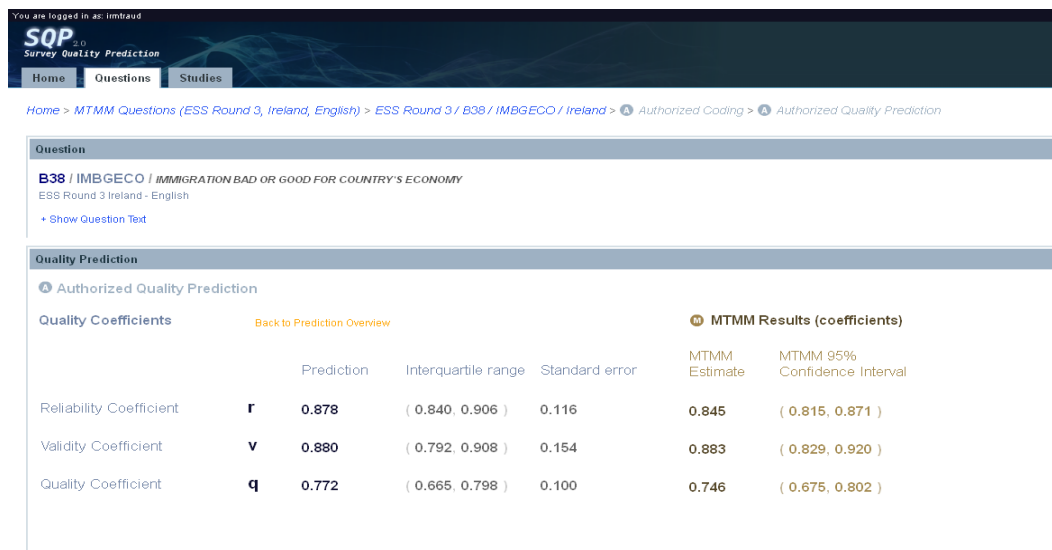


Figure 7.6 The comparison of the quality coefficients

The quality coefficients are comparable with factor loadings. In the MTMM experiments these coefficients have been estimated. They are the square root of the quality estimates. In this screen also the uncertainty is presented for both sets of estimates. It is clear that they overlap for a large part. This is what you expect if the two estimates give approximately the same result. It should, however, be clear that these two estimates are based on very different data. One is based on the MTMM data and the other on the coding of the question and the prediction procedure described in the previous chapter.

If one would like to see the codings, one can click on “View prediction codes”. Doing so we get screen presented in Figure 7.7.

The screenshot shows the SQP interface with the following content:

Question Panel (Left):

Question

B38 | IMBGECO | IMMIGRATION BAD OR GOOD FOR COUNTRY'S ECONOMY
ESS Round 3 Ireland - English

Request for Answer Text:
Would you say it is generally bad or good for Ireland's economy that people come to live here from other countries? Please use this card.

Answer options:

- 00 Bad for the economy
- 01
- 02
- 03
- 04
- 05
- 06
- 07
- 08
- 09
- 10 Good for the economy

Characteristic Panel (Right):

Characteristic	Choice
Domain	National politics
Domain: national politics	Economic / financial matters
Concept	All other simple concepts
Concept: other simple concepts	Evaluation
Social Desirability	A bit
Centrality	Rather central
Reference period	Present
Formulation of the request for an answer: basic choice	Indirect requests
WH word used in the request	Request without WH word
Request for an answer type	Interrogative
Use of gradation	Gradation used
Balance of the request	Balanced or not applicable
Presence of encouragement to answer	No particular encouragement present
Emphasis on subjective opinion in request	Emphasis on opinion present
Information about the opinion of other people	No information about opinions of others
Use of stimulus or statement in the request	No stimulus or statement
Absolute or comparative judgment	An absolute judgement
Response scale: basic choice	Categories
Number of categories	11
Labels of categories	Partially labelled
Labels with long or short text	Short text

Figure 7.7 The codes selected for the different characteristics of the question B38

At the right hand side we see the codes of all the characteristics of the question. At the left side is the text of the question indicated. These are the authorized codes of the characteristics approved by the team of RECSM.

7.2 The quality of non MTMM questions in the data base

Moving to the second option of the program SQP, we have to go back to the home page and click on “View all questions that are currently available”. Questions which have not been involved in MTMM experiments but are present in the data base can only be evaluated by predictions using the SQP prediction program. There are two possibilities: the questions have already be coded or not. Looking at Figure 7.8 we see both examples in round 2 of the ESS from Ireland, especially question G22 and G23. The latter has already been coded and approved by RECSM, indicated by the A behind the text of the question while the former G22 has not been coded by nobody so far

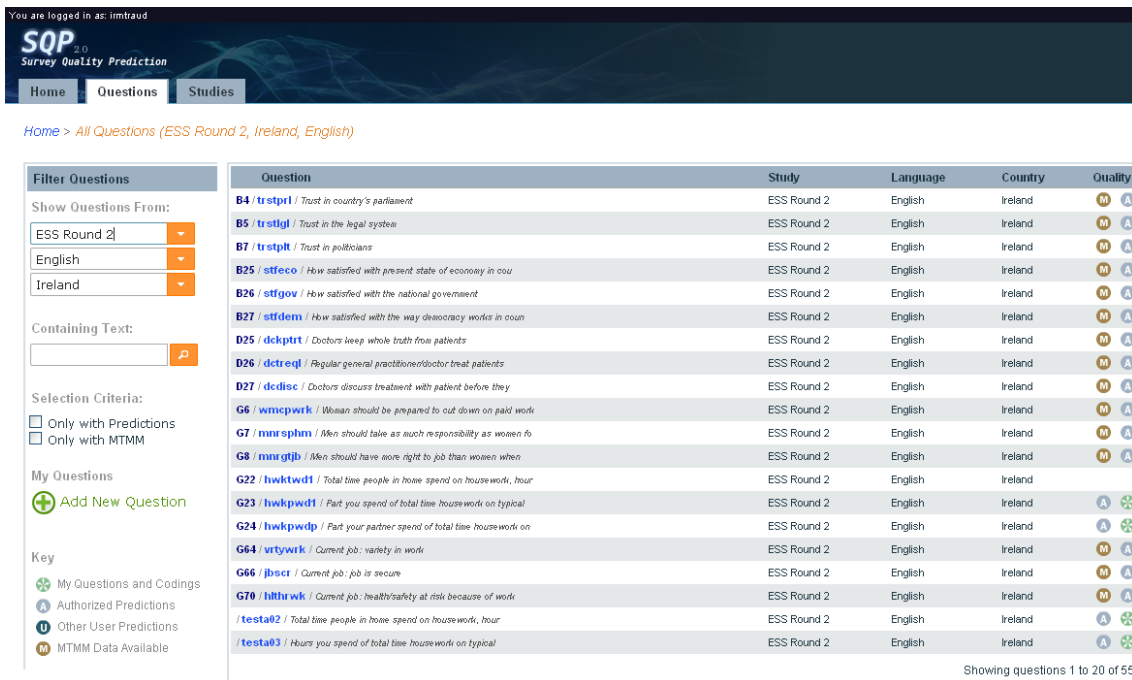


Figure 7.7 The overview of the question of Round 2 from Ireland

If we select question G23 first we get again the pop up screen for this question as before presenting the quality prediction by SQP based on the approved coding. We can also ask again the details of the quality estimates and the specification of the codes. So far it goes the same as before.

If we select G22 the process is different because G22 has not been coded so far. So if we select this question, we get the screen presented in Figure 7.8.

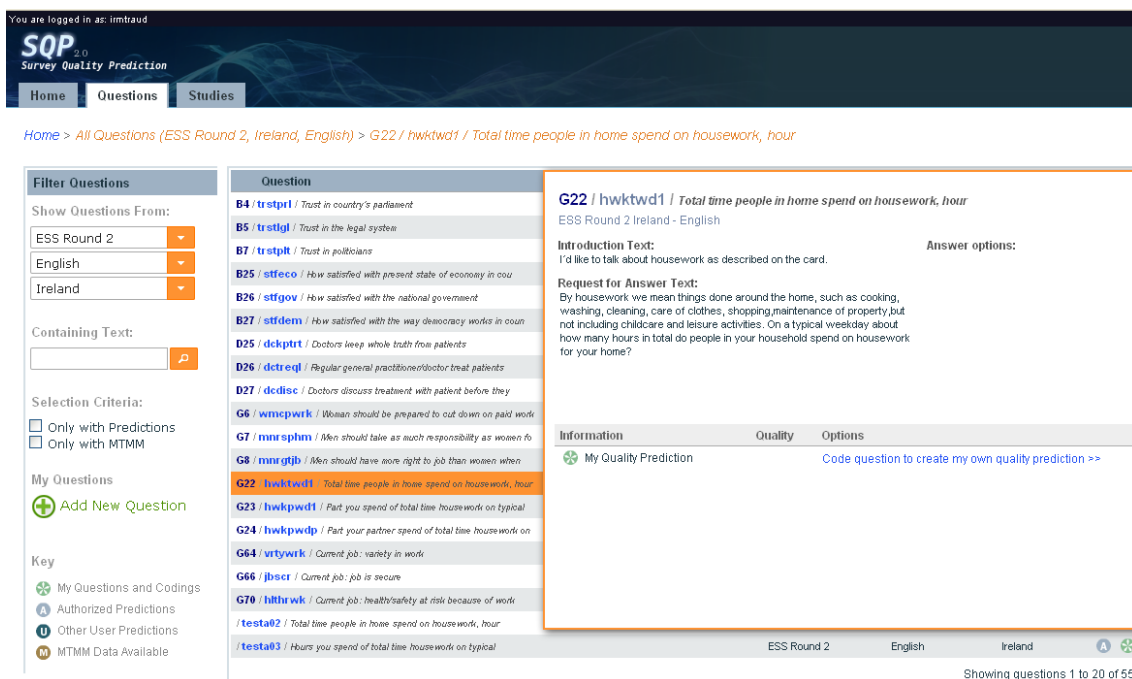


Figure 7.8 The screen if in for Ireland in round2 the question G22 has been chosen

In order to get a prediction of the quality of this question the first thing to do is to do the coding. If you click on "Code question to create my own prediction" the next screen is presented in Figure 7.9

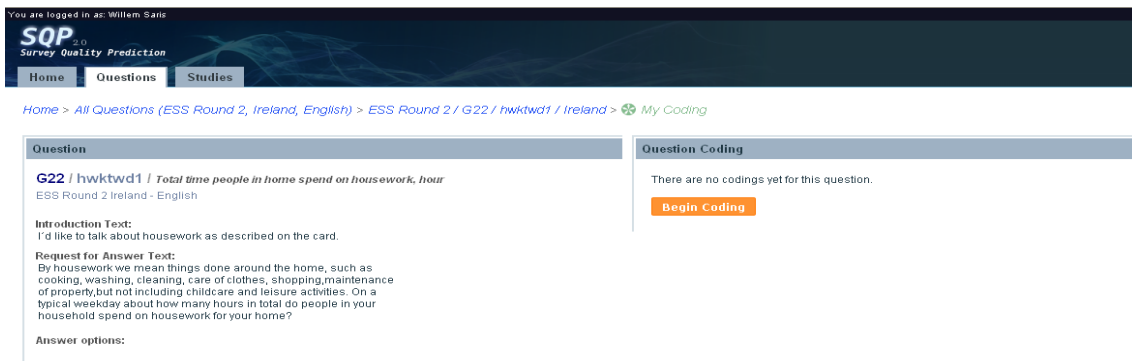


Figure 7.9 The screen with the question and the option to begin coding

Selecting “Begin coding” brings one to the screen presented in Figure 7.10. On the left side, in the lower part, the question and the answer categories are presented. On the top left side is the first characteristic is mentioned that should be coded. This is the domain of the question. The possible categories have been indicated. In yellow some information about this characteristic is indicated. If you select a category the choice is presented at the right side of the screen and the next characteristics to be coded appears at the left side. This characteristic is coded in the same way and this process goes on till all characteristics are coded.

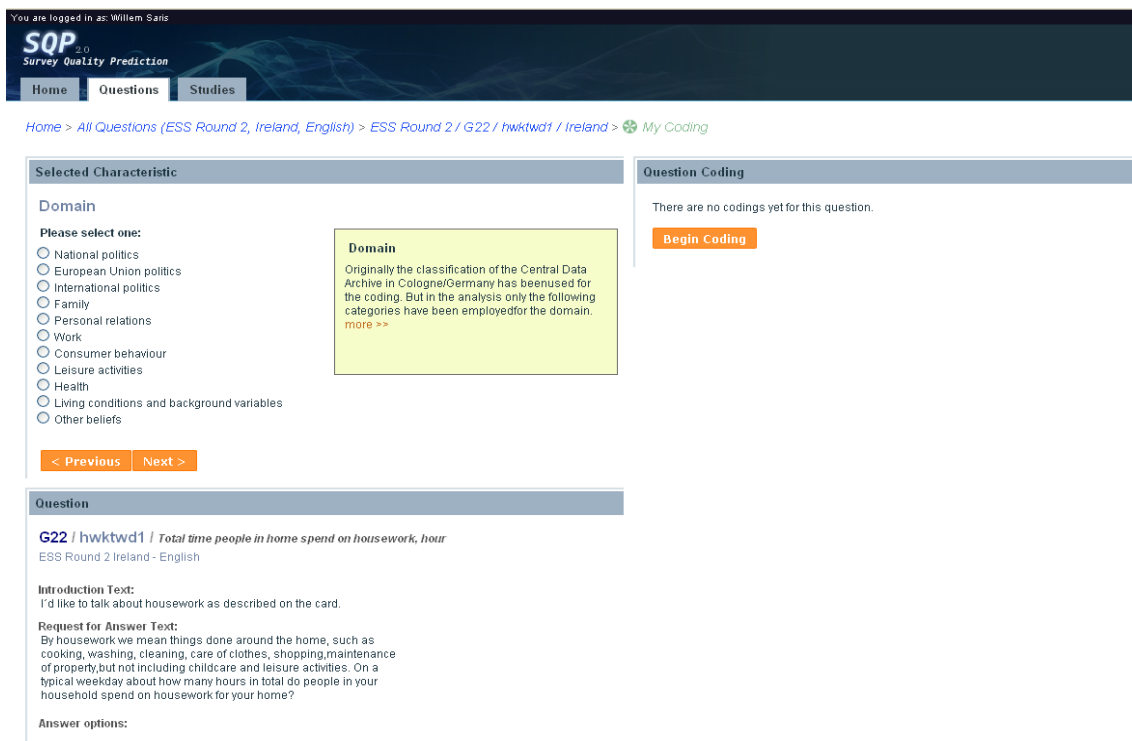



Figure 7.10 The first screen of the coding procedure

Some times the program makes a suggestion for a possible answer. For example it suggests how many sentences and words there are in the questions. In that case you can accept the suggestion by clicking on next or you can correct the number and click on “next” to go to the next

characteristic. When the coding is done for all characteristics the screen of Figure 7.11 appears.

Home > All Questions (ESS Round 2, Ireland, English) > ESS Round 2 / G22 / hwktwd1 / Ireland > My Coding

Question Coding Complete

 **Coding Complete!**

This question has been completely coded.

Question Quality Prediction

Get a prediction of the quality of this question based on the choices made for each characteristic.

[Get Quality Prediction >](#)

Return to the question list.

Go back to your search results for questions.

[Back to Question List](#)

Characteristic	Choice
Domain	Family
Domain: family	Household matters
Concept	Facts, background, or behaviour
Social Desirability	A lot
Centrality	Rather central
Reference period	Present
Formulation of the request for an answer: basic choice	Direct request
WH word used in the request	WH word used
WH word	How (quantity)
Request for an answer type	Interrogative
Use of gradation	Gradation used
Balance of the request	Balanced or not applicable
Presence of encouragement to answer	No particular encouragement present
Emphasis on subjective opinion in request	No emphasis on opinion present
Information about the opinion of other people	No information about opinions of others
Use of stimulus or statement in the request	No stimulus or statement
Absolute or comparative judgment	An absolute judgement
Response scale: basic choice	Frequencies or amounts
Number of frequencies	24
Don't know option	DK option only registered
Interviewer instruction	Present
Respondent instruction	Absent
Extra motivation, info or definition available?	Present
Knowledge provided	Definitions only
Introduction available?	Available

Figure 7.11 The screen after the coding has been completed

Now you can ask for a prediction of the quality of the question by clicking on the text "Get Quality Prediction". One can also continue coding or go back to the question list. If predictions are requested screen 7.12 will appear in this case.

You are logged in as: irmtraud

SPQ
Survey Quality Prediction

Home Questions Studies


Home > All Questions (ESS Round 2, Ireland, English) > ESS Round 2 / G22 / hwktwd1 / Ireland > My Coding > My Quality Prediction

Question

G22 / hwktwd1 | Total time people in home spend on housework, hour
ESS Round 2 Ireland - English

[+ Show Question Text](#)

Quality Prediction

 **My Quality Prediction**

Prediction Overview [View Quality Coefficients](#)

	Prediction
Reliability	r^2 0.731
Validity	v^2 0.926
Quality	q^2 0.677

Common Method Variance cmv 0.054

Figure 7.12 The quality prediction for question A1 in Great Britain

In this case only the prediction of SPQ is presented because no quality estimate was obtained for this question. If one would like the predictions of the quality coefficients which are the square root of the quality predictions it self one has to click on the button at the right saying "view quality coefficients". In that case one gets also the prediction intervals.

7.3 The prediction of the quality of new questions

For the third option of the program we have to go back to the home page of SQP and select the option “Create a new question”. If one chooses to introduce a new question one gets first a screen asking information about this study and question. In this case we specify that we do an study called immigration in English and the name of the variable will be called equality and the name of the question is the same. This information is also presented in the screen below (Figure 7.13)

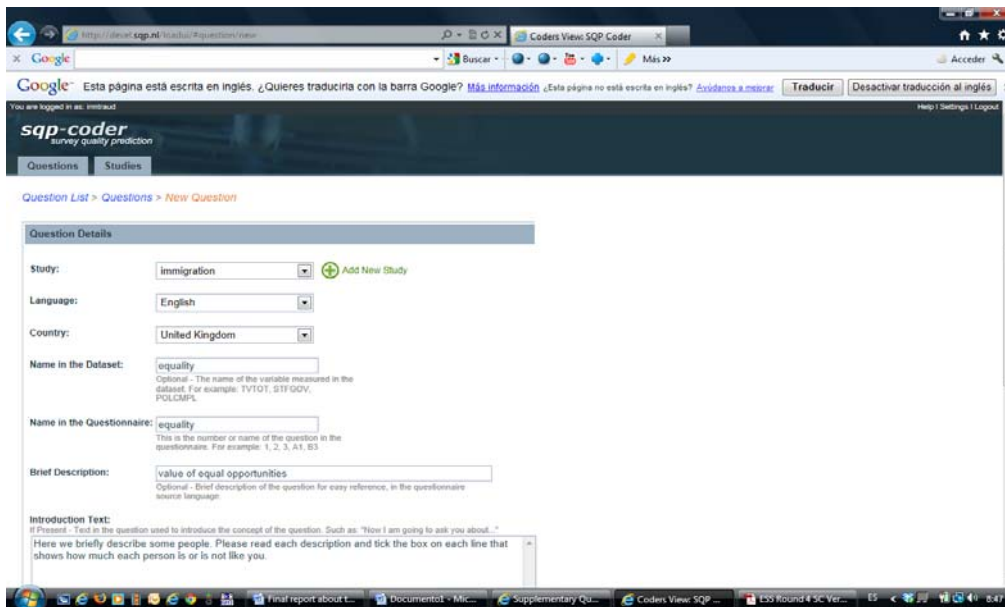


Figure 7.13 The screen registrating the basic information about the question

The next step is that we have to introduce the question it self. In this case we have chosen to introduce the question about the value equal opportunities of the Schwartz Human Value scale. This question has an introduction, a question with stimuli and 6 response categories. Figure 7.14. presents this specification

Introduction Text:
If Present - Text in the question used to introduce the concept of the question. Such as: "Now I am going to ask you about..."

Here we briefly describe some people. Please read each description and tick the box on each line that shows how much each person is or is not like you.

Request for Answer Text:
Text in the question that requests an answer such as: "Please select the option...", "How much time..."

How much like you is this person?

He thinks it is important that every person in the world should be treated equally.
He believes everyone should have equal opportunities in life.

Answer options:
Answer options or numbers in the answer scale. One option per line.

1 Very much like me
2 Like me
3 Somewhat like me
4 A little like me
5 Not like me
6 Not like me at all

Save Question

Figure 7.14 The form to specify the question

After that we have saved the question we get to see the screen presented in Figure 7.15.

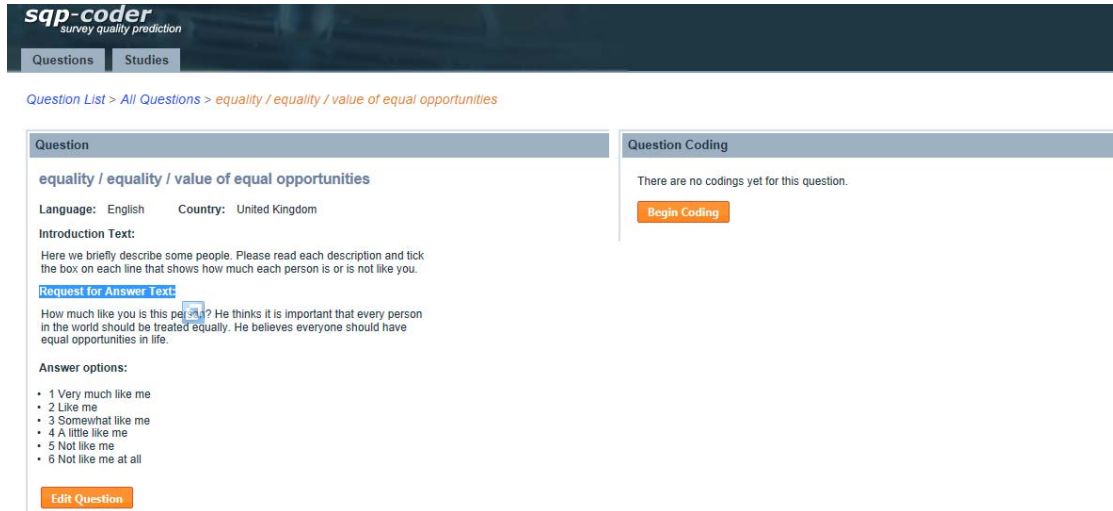


Figure 7.15 The question and the button to start the coding

The next thing that has to be done is the coding of the question. This starts by clicking on “begin coding”. If we have done the coding as show before and asked for the prediction of the quality we get the result presented in Figure 7.16

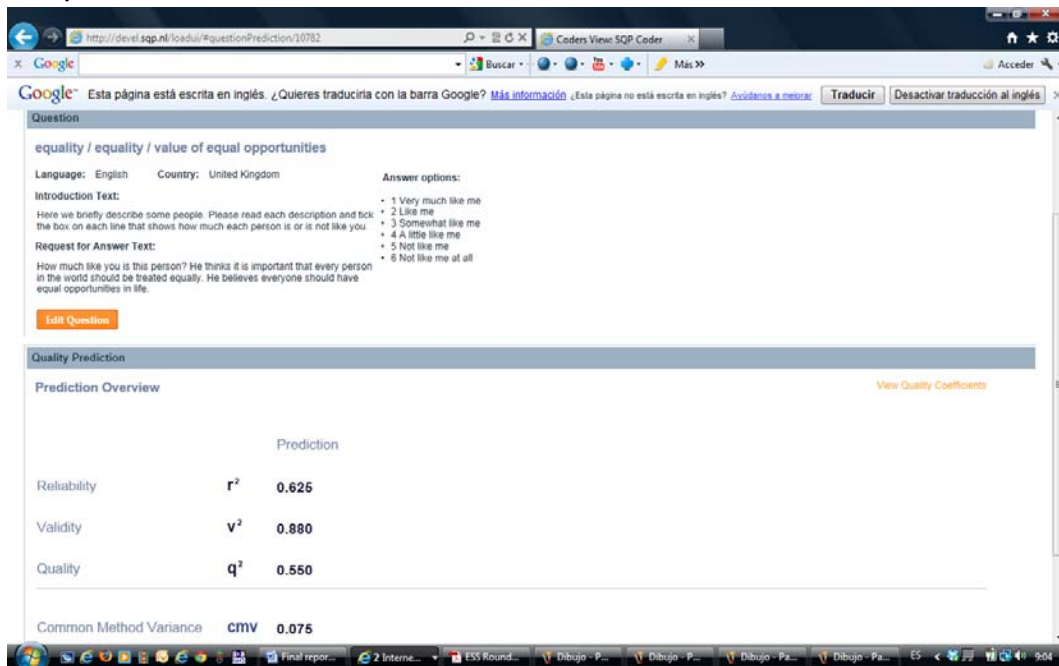


Figure 7.16 The prediction of SQP2.0 of the quality of this question

It will be clear that this question is not very good. So we can ask for suggestions for improvements. In this case the result after evaluation of all characteristics gives the result presented in Figure 7.17

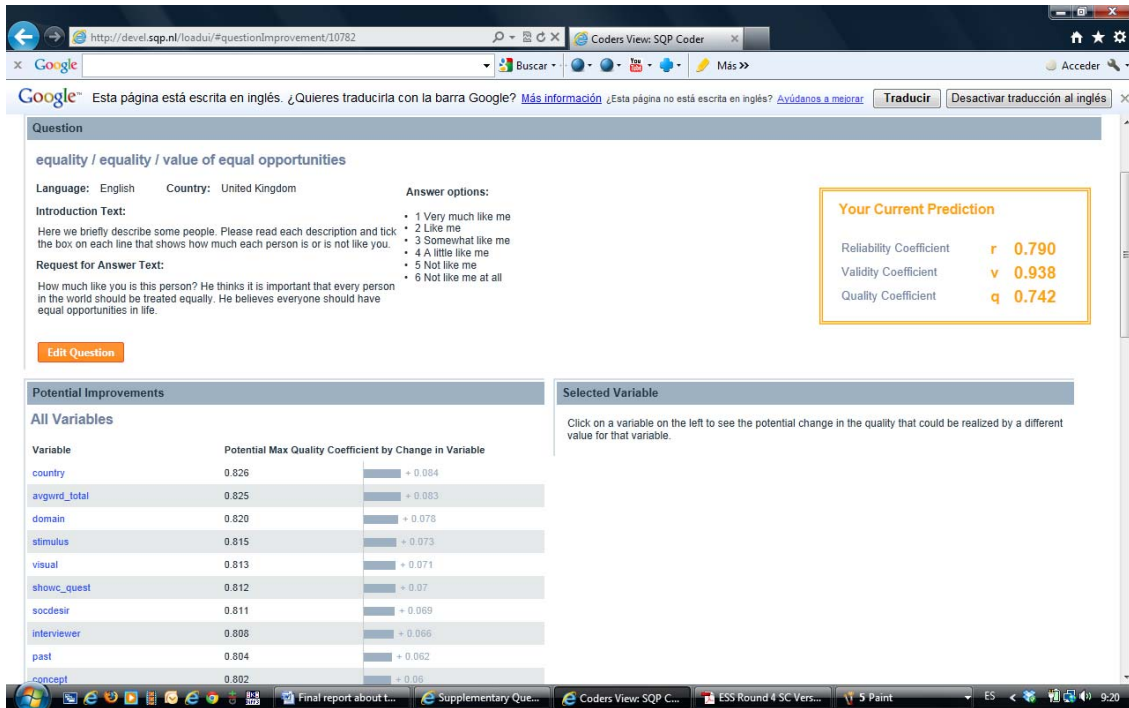


Figure 7.17 Several suggestions for improvement of the question

This analysis shows that several improvements can be made. We see that choosing an other country would help. This is of course an impossible option. Possible alternatives are presented by the characteristics avgwrд_total, stimulus, visual etc. One should realize that this table gives the improvement for one question characteristic at the time keeping all other the same as they are. This means that by combination of several of these characteristics one may be able to improve the question considerably. The program gives suggestions for this but one have to test the new version again one can not just add the different improvements together. In the next section we will discuss this issue in more detail.

7.4 Applications of the program

There are three relevant applications of the program SQP. The first is the improvement of questions before the data have been collected. The second is the use of the quality estimates for correction for measurement errors in the analysis between variables. The third application is the evaluation of the quality of composite scores for complex concepts. Of these three possible applications two will be discussed here. The latter possibility will not be discussed here. The evaluation of composite scores has been extensively discussed in Saris and Gallhofer (2007). So we start with the improvement of questions before the data collection.

7.4.1 Improvement of the quality of questions

In the last section we discussed the measurement of the value “equality” , an item of the Schwartz Human Values scale as introduced in the ESS. We have seen in

Figure 7.17 that SQP suggests many considerable improvements, especially with respect to: the length of the text, the use of stimuli, the data collection, the concept etc.

To start with the last issue. The question asked about the similarity of the respondent to the person described in the stimulus where in the stimulus a mixture of two concepts are presented: a value statement and a norm. This is what Saris and Gallhofer (2007) have called a complex concept because the item asks a similarity about other concepts. Besides that two different concepts have been combined in the stimulus. This could lead to a lot of confusion at the side of the respondent. We have also seen that use of batteries of statements have a negative effect therefore, following the suggestions of Saris and Gallhofer (2007), we would suggest to measure the value with a item specific question like:

How important or unimportant is it for you that all people are treated equally?

1. *Completely unimportant*
2. *Important*
3. *Neither unimportant neither important*
4. *Important*
5. *Extremely important*

This question is much shorter, it is a bipolar item specific scale and no statement is used. Let us see how good the quality of this question is according to the program SQP. In order to check this we introduce the question again in the program, code the question and ask for the quality prediction. The result is presented in Figure 7.18.

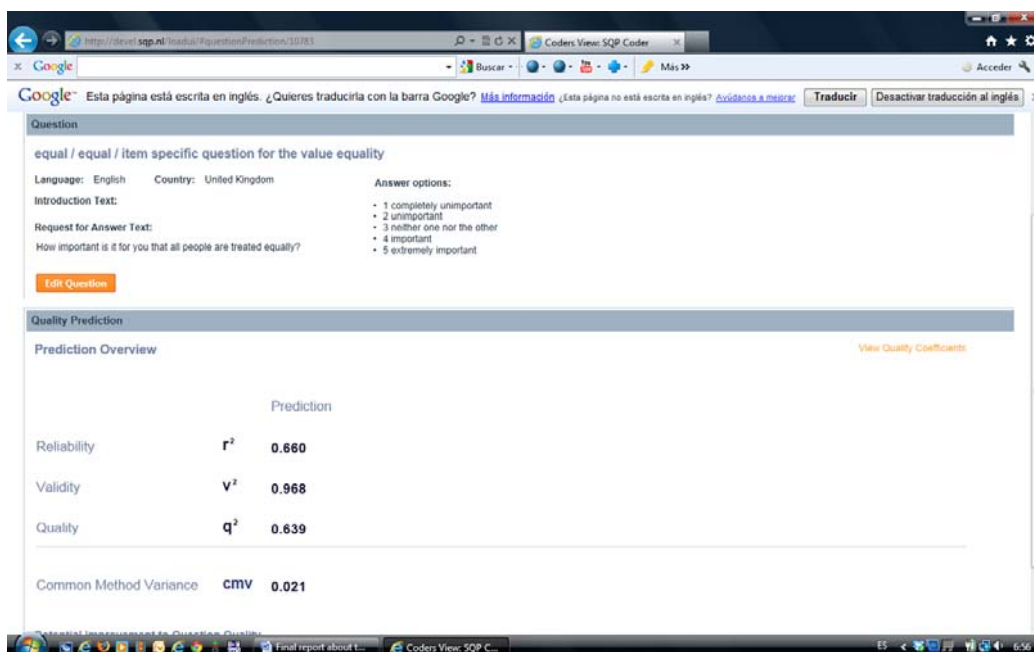


Figure 7.18 the quality of the reformulated question with respect to equality

The question of Schwartz had a quality of .55, the new question has a quality of .64. In explained variance this would mean that the explained variance of the observed variable by the variable of interest, the value equality, has been increase with nearly 10%. One can also look at further possible improvements but the explained variance will never be perfect which means that there remain always measurement errors.

Therefore, correction for measurement error is also important as we will see in the next section.

Note that the improvement in quality was mainly obtained by the increase in the validity. This means that using this formulation the systematic effect of the method i.e. the complement of the validity, has been reduced. This can also be seen in the reduction in the Common method variance which is now rather small.

A more detailed picture of the quality can be obtained by clicking on the text at the right “view quality coefficients”. If we do so we get the screen presented in Figure 7.19.

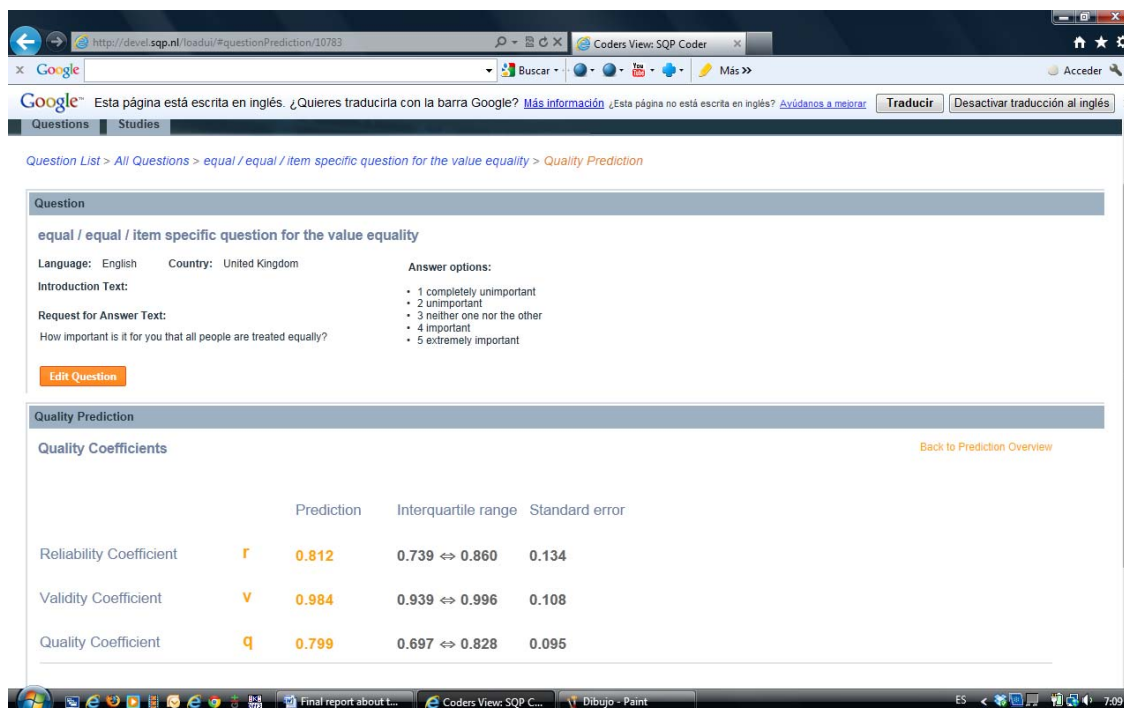


Figure 7.19 the quality coefficients, interquartile range and standard error

The quality coefficients are the square root of the quality indicators themselves. These are the coefficients which are estimated in the MTMM experiments. In this screen we see the uncertainty which exists in these estimates presented in the interquartile range and the standard error. It will be clear that a considerable range of uncertainty remains.

Nevertheless, the attraction of this approach is that we get these estimates before data have been collected. The MTMM experiments are time consuming and expensive. These quality estimates are obtained with minimal efforts and allow researchers to improve their data collection before they spend a lot of money on their data collection. It is not possible to take into account more than 50 question characteristics while formulating a question. SQP makes it possible to evaluate the questions made on these characteristics and suggest improvements. This is the major advantage of this procedure.

7.4.2 Correction for measurement error in the analysis

As we said before, measurement errors will remain, no matter how good we do our best to improve the questions. That means that the estimates of the relationships between the variables will be affected by these errors. Therefore it is necessary to correct for these errors. In this section we want to show by a simple example how this can be done.

The example we want to use is a model to explain the opinions about immigration. Variables to explain this opinion have been collected in the second round of the ESS. Some of the questions have already been discussed. We suggest for this example the model presented in Figure 7.20.

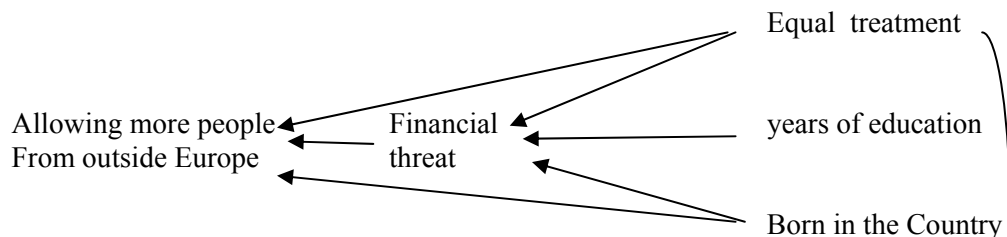


Figure 7.20 A simple model for explaining opinions about immigration

In round 3 of the ESS data for these variables have been collected in Ireland. The questions used are the following:

Immigration (Imm)

B37 STILL CARD 14 How about people from the poorer countries outside Europe? Use the same card.

- Allow many to come and live here* 1
- Allow some* 2
- Allow a few* 3
- Allow none* 4
- (Don't know)* 8

The quality of this question (.74) was estimated in a MTMM experiment.

Financial consequences (FIN);

B38~ CARD 15 Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? Please use this card.

- | | | | |
|----------------------------|--|----------------|---------------|
| <i>Bad</i> | | <i>Good</i> | |
| <i>for the</i> | | <i>for the</i> | |
| <i>economy</i> | | <i>economy</i> | <i>(Don't</i> |
| 00 01 02 03 04 05 06 07 08 | | 09 10 | <i>know)</i> |
| | | | 88 |

The quality of this question (.557) was estimated by MTMM (Figure 7.4)

Equal treatment (Equal):

Here we briefly describe some people. Please read each description and tick the box on each line that shows how much each person is or is not like you.

How much like you is this person? He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.

- 1 Very much like me*
- 2 Like me*
- 3 Somewhat like me*
- 4 A little like me*
- 5 Not like me*
- 6 Not like me at all*

The quality of this question (.55) was predicted by SQP

Years of education (Edu):

F7 How many years of full-time education have you completed?

WRITE IN:

(DON'T know) 88

The quality of this question (.78) was predicted by SQP.

Nationality (National):

C20 Were you born in [country]?

Yes 1 GO TO C23

No 2 ASK C21

(Don't know) 8 GO TO C23

The quality of this question (.78) was predicted by SQP.

The correlations between these variables obtained in Ireland in round 3 of the ESS were as indicated in Table 7.1.

1.00				
-.421	1.00			
.163	-.109	1.00		
-.156	.172	-.063	1.00	
-.085	.145	-.021	.103	1.00
Imm	fin	egal	edu	nat

Table 7.1 The correlations between these variables obtained in Ireland (n=1700)

On the basis of these data the effects presented in Figure 7.20 can be estimated with and without correction for measurement error. Normally the analysis is done without correction for measurement error. In that case the estimation is done on the basis of the correlation matrix of Figure 7.1 without any adjustment.

If one wants to correct for measurement error one has to make on change in this matrix which is that the 1's on the diagonal should be substituted by the quality estimates. So the adjusted matrix for this example is presented in Table 7.2.

.740				
-.421	.557			
.163	-.109	.550		
-.156	.172	-.063	.780	
-.085	.145	-.021	.103	.780
Imm	fin	egal	edu	nat

Table 7.2 The correlations with on the diagonal the quality estimates for the variables

It will be clear that now the matrix is not a correlation matrix anymore, however by transforming this matrix in a correlation matrix (using a program), one will get the correlation between these variables corrected for measurement error²⁰. The result is presented in Table 7.3.

Imm	1.00				
fin	-0.66	1.00			
equal	0.26	-0.20	1.00		
edu	-0.21	0.26	-0.10	1.00	
nat	-0.11	0.22	-0.03	0.13	1.00

Table 7.3 The correlations corrected for measurement error

It will be clear that all correlations have been increased by this correction for measurement error. As a consequence, we should also expect that the estimates of the effects will be different, in general higher. The effects have been estimated with the ML estimator of LISREL. The inputs for these analyses have been presented in Appendix 7.1. The results without and with correction for measurement error have been presented in Table 7.4.

The most striking result is that the effect of the opinion about the “financial consequences” on the opinion “to allow more immigrants” has been changed from -.39 to -.64. This is a bit less than a doubling of the effect. For other effects the changes are not so big in absolute value but they are for several parameters approximately the same relative to the coefficient in the analysis without correction for errors. However we also see that the coefficient don’t get always larger. Occasionally this is not the case.

	Without correction On immigration	with correction for errors on immigration
By		
Financial consequences	-.39	-.63
Equal treatment	.11	.13
Education	-.08	-.03
Nationality	-.02	-.04
Total explained (R ²)	.20	.45
	On financial consequences	on financial consequences
By		
Equal treatment	-.10	-.17
Education	.15	.22
Nationality	.13	.19
Total explained (R ²)	.06	.13

Table 7.4 The estimates of the effects of the variables on the Opinions about Immigration and Financial consequences without and with correction for measurement error.

²⁰ This approach is a bit too simple because we ignore possible extra correlations due to method effects and the fact that we take the quality estimates as given values. For more details we refer to Saris and Gallhofer (2007), Lance et al (2010) and Oberski (2011).

We gave this example to show that taking into account the quality of the questions (i.e. correction for measurement error) can have a considerable effect on the results of analysis of the relationships between variables. Therefore we are of the opinion that the information about the quality of questions is essential for the analysis of survey data and even more so in comparative research.

Looking at the Appendix 7.1 one can see that the procedure to take the quality into account is very simple. One only has to substitute the 1's on the diagonal of a correlation matrix by the quality coefficients and transform (with the program) the covariance matrix in a correlation matrix and analyze the data as before and one gets the results corrected for measurement errors.

7.5 Conclusions

In this chapter we have shown that the program SQP2.0 can be used to obtain (1) the quality estimates that were obtained by MTMM experiments (2) the quality predictions by SQP of questions that are in our data base but were not part of a MTMM experiment and (3) the quality predictions by SQP of new questions that a researcher would like to evaluate. We have also shown that the program provides in a simple way suggestions for improvement of questions.

In the last part of this chapter we have illustrated by a simple example how the quality estimates can be used to correct for measurement error in regression or more general structural equation models. This last topic is in fact the reason why the ESS and we pay so much attention to quality of questions or measurement errors. The example has shown that one can get very different results for the parameters of interest if one corrects for measurement error. This is even more important for comparative research because in comparative research the correlations across countries may be different not because of differences in relationships between the variables of interest but just because of differences in measurement errors or quality of the measures. So we think that comparative research is only possible with the correction for errors as we have indicated in the last section. It is for this reason that we do all the efforts discussed in this report.

Appendix 7.1 The LISREL inputs to estimate the parameters of the model in Figure 7.20

The LISREL input for the analysis without correction for measurement errors

```
immigration Ireland
da ni=5 no=1700 ma=km
km
1.00
-.421 1.00
.163 -.109 1.00
-.156 .172 -.063 1.00
-.085 .145 -.021 .103 1.00
labels
Imm fin eqal edu nat
model ny=2 nx=3 be=fu.fi ga=fu,fi ps=di,fr
fr be 1 2 ga 2 1 ga 2 2 ga 2 3
fr ga 1 1 ga 1 2 ga 1 3
out
```

The LISREL input for the analysis with correction for measurement errors

```
immigration Ireland
da ni=5 no=1700 ma=km
cm
.740
-.421 .557
.163 -.109 .550
-.156 .172 -.063 .780
-.085 .145 -.021 .103 .780
labels
Imm fin eqal edu nat
model ny=2 nx=3 be=fu.fi ga=fu,fi ps=di,fr
fr be 1 2 ga 2 1 ga 2 2 ga 2 3
fr ga 1 1 ga 1 2 ga 1 3
out
```

These two inputs show that the only part that has been changed is the diagonal of the correlation matrix where we have introduced the quality coefficients obtained in MTMM experiments or by prediction using SQP 2.0.

Chapter 8

Conclusions and future developments

Willem Saris

Any measurement will contain errors. These errors will effect the estimates of means and relationships between variables. These problems are even larger in comparative research because the differences in measurement errors can cause differences across countries which have nothing to do with substantial differences. Therefore the Central Coordinating team of the ESS decided to introduce a supplementary questionnaire next to the main questionnaire in all data collections in order to determine the size of the measurement errors in all countries. These estimates can be used to correct for measurement errors and in this way make the data across countries comparable.

In this report we have presented the results of these experiments in the first three rounds. We will also show how these results can be used in practice in a simple way.

Let us start with a summary of the large amount of results we have obtained in the context of the ESS infrastructure.

Database of questions. Because in each round MTMM experiments have been done a data base has been created with alternative forms of questions which are supposed to measure the same variable. The combination with the obtained quality estimates allows the user of the data base to select for specific variables the optimal form. For details see chapter 3.

A new design for MTMM studies. The classical MTMM experiment requires that all respondents answer three questions measuring the same variable. This may lead to memory effects or satisficing. Therefore we looked for an alternative which has been found in the Split ballot MTMM design. In this design each respondent has to answer on twice a similar question for the same variable. For the ESS especially, we developed the 2 group design where all respondents get the same form of the question in the main questionnaire. In the supplementary questionnaire the sample is split randomly in two groups which get each a different form of the question to measure the same variable. It was shown that using this design all quality criteria could be estimated although some estimation problems were expected. For details see chapter 2.

A new procedure for the analysis of the data. In the analysis of the data of the Split ballot MTMM experiments it turned out that the expected problems occurred more frequently than expected. Therefore a study was made of the problems and of possible solutions. It turned out that the solution was to start the analysis with a Multiple group analysis assuming the same model across all countries and relaxing this assumption on the basis of detection of misspecifications in the model. This analysis was done by two researchers independent of each other and after that a comparison was made and optimal estimates were produced for both analyses. For details we refer to chapter 4.

A new program for analysis of the data. The analysis of the MTMM experiments had become a rather complex process. Therefore several new programs were developed to facilitate this analysis. These programs allow the analyzers to input the LISREL model syntax, run it, and obtain outputs and to make a comparison of the quality estimates with previous versions or other analyzer's versions. Each run of an analysis was stored using the version control system *git* (2009). The analyzers could also view the exact differences ("diff") between their model syntax and that of another version or analyzer, as well as obtain a side-by-side comparison of the quality estimates. This allowed them to pinpoint the exact model changes that may have led to any differences in estimates. An online repository of this history, combining the repositories of all analyzers, is available. For details see chapter 4.

A database of 3483 coded questions. The data base of ESS contains questions in many different languages. For the estimation of the effects of the questions characteristics on the quality of the questions the characteristics had to be coded. Because of the different languages this is a difficult issue. However we found in Barcelona enough native speakers in all languages available in the ESS. So all questions involved in the MTMM experiments have been coded by native speakers using a new program for coding of questions. The obtained results were compared with the codes obtained for the source questionnaire which was coded by two coders of our team. If differences between the codes in the different languages and the source questionnaire were detected these difference were discussed and solved to get a consensus concerning the coding of the foreign languages. For details see chapter 5.

A new procedure for quality prediction. Given that in the data base for 3483 questions the question characteristics were coded and the quality estimates, reliability, validity, and quality, were available a prediction procedure had to be chosen. For this purpose a new prediction approach, the Random Forest program of Breiman (2003), has been chosen based on the logits of the quality estimates. The advantages of this choice above linear regression used earlier is that no impossible predictions are possible (>1), that one get construct 95% prediction intervals. It turned out that the predictions were much better than with the old program SQP. For details , see chapter 6

A new program SQP2.0 Based on the work mentioned above a new program SQP 2.0 has been created for the predictions and improvement of questions. With this program users can obtain the estimated quality of the questions that were involved in MTMM experiments. They can also get a prediction of the program SQP 2.0 and suggestions of improvements. Users can also get predictions of the quality of all questions, already existing questions in the data base or new questions in many European languages. However this prediction requires that the user codes the characteristics of the question. The program will than provide the quality estimates and suggestions for possible improvements

This overview summarizes the work our research group has done to make it possible for users of the program SQP 2.0 to get a estimate and/or a prediction of the quality of any questions that can be formulated in all languages used in Europe. This does not mean that the estimates and predictions are equally good for all questions. We will discuss the limitations of the program below.

8.1 Limits and future developments

In chapter 6 we have indicated how frequently different questions are asked in MTMM experiments studied. It is impossible to determine how represent the distributions with respect to the domains, concepts and the methods are. What we have done in our research is always choosing questions which were included in ongoing surveys. So, at least the questions selected are questions used in survey research. With respect to the prediction of the quality of questions the result may depend on the selection of the questions. So in this respect there is an uncertainty in our approach. However, it is unclear how this situation can be improved given that there does not exist something like a 'population of questions, let alone a sampling for drawing questions.

With respect to the ESS studies we have concentrated our selection of questions on the questions which will be repeated over the years i.e. we selected especially questions from the core questionnaire. So for these questions we have the quality estimates available in our data base. So for the most commonly used question in the ESS there is no problem. The information about the quality is available.

Another point on which the approach so far is limited is that it is not easy to look for a specific question. At the moment the system can be used in combination of the questionnaires of the ESS. One can find in the questionnaires the number and the name of the question and use this information to look up that question in SQP2.0. At this moment one can not search on words or combinations of words in the question text or on names of concepts. This possibility will be a next step in the process.

The next limitation we should mention is that we concentrate on single questions and not concepts measured by several questions together. In our publications (Saris and Gallhofer 2007) we have indicated how the information presented here can be used to evaluate such concepts but there is no automatic procedure available to do so at this moment. In the context of an extension of the programs we will take this issue also into account. For the moment we can only refer to the publication mentioned above.

The same is true for a simple procedure to take the measurement errors into account in the analysis. We have indicated in Chapter 7 that this can be done relatively simply. Therefore we are planning to include this option also in the next program we are going to develop in this context.

REFERENCES

- Allison P. D. 1987. Estimation of linear models with incomplete data. In C. C. Clogg (ed.), *Sociological Methodology*, Washington DC: American Sociological Association, 71–103.
- Althauser R. P., T. A. Heberlein, and R. A. Scott 1971. A causal assessment of validity: The augmented multitrait-multimethod matrix. In H. M. Blalock Jr. (ed.), *Causal Models in the Social Sciences*. Chicago: Aldine, 151–169.
- Alwin D. F. 1974. An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (ed.), *Sociological Methodology*, San Francisco: Jossey Bass, 79–105.
- Alwin D. F., and I. A. Krosnick 1991. The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, 20, 139–181.
- Alwin D. F. 1997. Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research*, 25, 318–341.
- Alwin D.F. 2007. Margins of error: A study of reliability in survey measurement. Hoboken, Wiley.
- Andrews F. M. 1984. Construct validity and error components of survey measures: A structural equation approach. *Public Opinion Quarterly*, 48, 409–442.
- Arminger G., and M. E. Sobel 1991. Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, 85, 195–203.
- Bagozzi R.P. and 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. and Waterplas L (1986) *Het survey-interview onderzocht*. Leuven Sociologisch Onderzoeksinstituut, KU Leuven.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Browne M. W. 1984. The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37, 1–21.
- Campbell, D. T., and D. W. Fiske 1959. Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin*, 56, 81–105.
- Campbell D. T., and E. I. O'Connell 1967. Method factors in multitrait-multimethod matrices: multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409–426.

- Coenders G., and W. E. Saris 1998. Relationship between a restricted correlated uniqueness model and a direct product model for multitrait-multimethod data. In A. Ferligoi (ed.), *Advances in Methodology, Data Analysis and Statistic., Metodološki Zvezki* 14. Ljubljana: FDV, 151–172.
- Coenders, G., and W. E. Saris 2000. Testing nested additive, multiplicative and general multitrait-multimethod models. *Structural Equation Modeling* , 7, 219–250.
- Corten I., W. E. Saris, G. Coenders, W. van der Veld, C. Albers, and C. Cornelis 2002. The fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, 9, 213–232 .
- Cote J.A. , Buckley M.R. (1987) Estimating trait, method and error variance; generalizing across 70 construct validity studies. *Journal of Marketing Research*, 11, 535-559
- Cox E. (1980) The optimal number of response alternatives for a scale: A review. *Journal of Marketing research*, 17,407-422
- Cudeck, R. 1988. Multiplicative models and MTMM matrices. *Journal of Educational Statistics*, 13, 131–147.
- Dillman D. A. 1978. *Mail and Telephone Survey: The Total Design Method*. New York: Wiley.
- Dillman D. A. 2000. *Mail and Internet Surveys. The Tailored Design Method*. New York: Wiley.
- Eid M. (2000).Multitrait-multimethod model with minimal assumptions.*Psychometrika*, 65, 241–261.
- Eid M. and Diener E. (2006) *Handbook of multimethod measurement in psychology*. Washington DC: American Psychological Association.
- Esposito J., Campanelli P., Rothgeb J. and Polivka A. (1991) Determining which questions are best: Methodologies for evaluating survey questions.
- European Social Survey 2002. *European Social Survey Round 1: Report of the First Round*
- Forsman G. And I.Schreiner (1991) The design and analysis of Reinterview: An Overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley,279-303
- Forsyth B., Lessler J., and Hubbard M (1992) Cognitive evaluation of the questionnaire. In Tanur and R. Tourangeau (Eds) *Cognition and survey research*. NewYork; Wiley, 183-198.
- Graesser A. Wiener-Hastings K., Wiemer –Hastings P., and Kreuz R. (2000) *The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices*.
- Groves, R. M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.
- Heise D. R. 1969. Separating reliability and stability in test-retest-correlation.

American Sociological Review, 34, 93–101.

a.

- Heise D. R., and G. W. Bohrnstedt 1970. Validity, invalidity and reliability. *Sociological Methodology*, 2, 104–129.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.
- Jöreskog K. G. 1971. Simultaneous factor analysis in several populations, *Psychometrika* 34, 409–426.
- Jöreskog K. G., and D. Sörbom 1989). *LISREL 7. A Guide to the Program and Applications*. Chicago: SPSS Inc.
- Jöreskog, K.G. and Sörbom, D. (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.
- Kenny D. A., and D. A. Kashy 1992. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112 , 165–172.
- Kolenikov, S., and K.A. Bollen. (2008). “Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?” *University of North Carolina*
- Költringer R. 1995. Measurement quality in Austrian personal interview surveys. In W. E. Saris, and A. Münnich (eds.), *The Multitrait-Multimethod Approach to evaluate measurement instruments*, Budapest: Eötvös University Press, 207–225.
- Krosnick J. A., and R. P. Abelson 1991 . The case for measuring attitude strength in surveys. In J. M. Tanur (ed.), *Questions about Questions. Inquiries into the Cognitive Bases of Surveys*, New York: Russel Sage Foundation, 177–203.
- Krosnick J. A. and L.R. Fabrigar (forthcoming). *Designing Good Questionnaires: Insights from Cognitive Psychology*.
- Lance C.E., B. Dawson, D.Birkelbach and B.J.Hoffman (2010) Method effects, measurement error and substantive conclusions. *Organizational Research Methods*, 13, 435-455
- Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. *R News* 2(3), 18--22.
- Marsh, H. W., and L. Bailey 1991. Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 47–70.
- Marsh H. W. 1989. Confirmatory factor analysis of multitrait-multimethod data: many problems and few solutions. *Applied Psychological Measurement* , 13, 335–361.
- Molenaar. N. I. 1986. *Formuleringseffecten in Survey-Interviews. PhD thesis*, Amsterdam: Free University.
- Muthén, L.K. and Muthén, B.O. (1998-2007). *Mplus User’s Guide*. Fifth Edition. Los Angeles, CA: Muthén & Muthén

- Németh, László and hyphenation file contributors (2005). *Hunspell* 1.3.1. [Budapest Technical University Media Research Centre \(BME MOKK\)](#).
- Nonyane, Bareng A. S. and Foulkes, Andrea S. (2007) "Multiple Imputation and Random Forests (MIRF) for Unobservable, High-Dimensional Data," *The International Journal of Biostatistics*: Vol. 3: Iss. 1, Article 12.
- Oberski D., L. Kuipers, and W.E. Saris 2005. *SQP Survey Quality Predictor*. www.sqp.nl
- Presser S. and Blair J. (1994) Survey pretesting: Do different methods produce different results? In Marsden (Ed) *Sociological Methodology*. Oxford, Basil Blackwell, 73-104
- R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Revilla M. and W.Saris (2011) The split-ballot MTMM approach: implementation and problems. Barcelona, *RECSM working paper 19*.
- Rindskopf, D. (1984). "Structural Equation Models." *Sociological Methods & Research* 13 (1):109.
- Rodgers W. L., F. M. Andrews, and A. R. Herzog 1992. Quality of survey measures: A structural modelling approach. *Journal of Official Statistics*, 8, 251—275.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
- Saris W. (1988) *Variation in response functions: a source of measurement error in attitude research*. Amsterdam: SRF.
- Saris W. E. 1990. The choice of a model for evaluation of measurement instruments. In W. E. Saris, and A. van Meurs (eds.), *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod studies*, Amsterdam: North Holland, 118—133.
- Saris W. E., and F. M. Andrews 1991. Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, 575—599.
- Saris W. E., and I. N. Gallhofer 1998. Classificatie van survey-vragen. *Tijdschrift voor Communicatie wetenschap*, 2, 96—122.
- Saris W. E., and C. Aalberts 2003. Different explanations for correlated errors in MTMM studies. *Structural Equation Modeling*, 10, 193—214.
- Saris W. E., A. Satorra, and G. Coenders 2004b. A new approach for evaluating quality of measurement instruments. *Sociological Methodology*, 3, 311—347.

- Saris W. E., and I. N. Gallhofer 2004. Operationalization of social science concepts by intuition. *Quality and Quantity*, 38, 235-258.
- Saris W. E., and I. N. Gallhofer 2006. *The results of the MTMM experiments in round 2. Report for the ESS*. London, ESS.
- Saris W. E., and I. N. Gallhofer 2007. Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1, 31–46.
- Saris W.E. and I.N. Gallhofer 2007 design, evaluation and analysis of questionnaires for Survey research, Hoboken, Wiley.
- Saris W.E., M.Revilla, J. Krosnick and E.M.Schaefer, (2009) Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods 4(1): 61-79*
- Saris, W.E, Satorra, A., Van der Veld, W.M. (2009). “Testing Structural Equation Models or Detection of Misspecifications?” *Structural equation modeling: A multidisciplinary Journal*, 16 (4): 561-582
- Satorra A. 1992. Asymptotic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (ed.), *Sociological Methodology 1992*. Oxford: Basil Blackwell, 249–278.
- Scherpenzeel A. C. 1995. *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies*. KPN Research: Leidschendam.
- Scherpenzeel A. C., and W. E. Saris 1997. The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*. 25, 341–383.
- Scherpenzeel A. C., and W. E. Saris 2006. Multitrait-Multimethod models for longitudinal research. In K.van Montford, H. Oud and A. Satorra (eds.), *Longitudinal Models in Behavioral and Related Sciences*, London: Lawrence Erlbaum, 381–403.
- Schmid, Helmut (1994): [Probabilistic Part-of-Speech Tagging Using Decision Trees](#). *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Schuman H., and S. Presser 1981. *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York: Academic Press.
- Subman S and Brandburn N. (1982) *Asking questions: a practical guide to questionnaire design*. San Francisco, Jossey Bass
- Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1), 307.
- Sudman S., and N. M. Bradburn, and N. Schwarz 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

- Torvalds, Linus (2009). *Git*. <http://git-scm.com/>
- Tourangeau R., L. J. Rips, and K. Rasinski 2000. *The Psychology of Survey Response*. Cambridge MA: Cambridge University Press.
- Van Buuren, Stef, and Karin Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, forthcoming.
- Van der Veld, W.M., Saris, W.E., Satorra, A. (2008) Judgment Aid Rule Software
- Van der Zouwen J. 2000. An assesment of the difficulty of questions used in the ISSP questionnaires, the clarity of their wording and the comparability of the responses. *ZA-information* 45, 96–114.
- Van Driel, O. P. (1978). “On various Causes of Improper Solutions in Maximum Likelihood Factor Analysis.” *Psychometrika* 43 (2):225-243.
- Van Meurs A., and W. E. Saris 1990. Memory effects in MTMM studies. In W. E. Saris and A. van Meurs (eds.), *Evaluations of Measurement Instruments by Metaanalysis of Multitrait-Multimethod Studies*, Amsterdam: North Holland, 134–146.
- Werts C. E., and R. L. Linn 1970. Path analysis. Psychological examples. *Psychological Bulletin*, 74, 193–212.
- Wiley D. E., and I. A. Wiley 1970. The estimation of measurement error in panel data. *American Sociological Review*, 35, 112–117.
- Wothke W. 1996. Models for multitrait-multimethod matrix analysis. In G. C. Marcoulides, and R. E. Schumacker (eds.), *Advanced Structural Equation Modeling: Issues and Techniques*, Mahwah NJ: L. Erlbaum, 7–56.

Appendix: The MTMM questions in the ESS

Round1 Experiment 1 Media

Questions in the Main questionnaire

A1 TvTot

CARD 1 On an average weekday, how much time, in total, do you spend watching television? Please use this card to answer.

No time at all	00	GO TO A3
Less than ½ hour	01	
½ hour to 1 hour	02	
More than 1 hour, up to 1½ hours	03	
More than 1½ hours, up to 2 hours	04	ASK A2
More than 2 hours, up to 2½ hours	05	
More than 2½ hours, up to 3 hours	06	
More than 3 hours	07	
(Don't know)	88	

ASK ALL

A3 RdTot

STILL CARD 1 On an average weekday, how much time, in total, do you spend listening to the radio? Use the same card.

No time at all	00	GO TO A5
Less than ½ hour	01	
½ hour to 1 hour	02	
More than 1 hour, up to 1½ hours	03	
More than 1½ hours, up to 2 hours	04	
More than 2 hours, up to 2½ hours	05	ASK A4
More than 2½ hours, up to 3 hours	06	
More than 3 hours	07	
(Don't know)	88	

ASK ALL

A5 NwspTot

STILL CARD 1 On an average weekday, how much time, in total, do you spend reading the newspapers? Use this card again

No time at all	00	GO TO A7
Less than ½ hour	01	
½ hour to 1 hour	02	
More than 1 hour, up to 1½ hours	03	
More than 1½ hours, up to 2 hours	04	
More than 2 hours, up to 2½ hours	05	ASK A6
More than 2½ hours, up to 3 hours	06	
More than 3 hours	07	
(Don't know)	88	

Questions in the supplementary questionnaire: group1

ALL RESPONDENTS ANSWER

HS1 On an average weekday, how much time, in total, do you spend watching television²⁰?

WRITE IN HOURS: AND MINUTES:

HS2 On an average weekday, how much time, in total, do you spend listening to the radio²¹?

WRITE IN HOURS: AND MINUTES:

HS3 On an average weekday, how much time, in total, do you spend reading the newspapers²²?

WRITE IN HOURS: AND MINUTES:

Questions in the supplementary questionnaire: group2

HS19 On an average weekday, how much time, in total, do you spend watching television³⁶? **Please tick one box.**

No time at all 01

Very little time 02

A little time 03

Some time 04

Quite a lot of time 05

A lot of time 06

A great deal of time 07

HS20 On an average weekday, how much time, in total, do you spend listening to the radio³⁷? **Please tick one box.**

No time at all 01

Very little time 02

A little time 03

Some time 04

Quite a lot of time 05

A lot of time 06

A great deal of time 07

HS21 On an average weekday, how much time, in total, do you spend reading the newspapers³⁸? **Please tick one box.**

No time at all 01

Very little time 02

A little time 03

Some time 04

Quite a lot of time 05

A lot of time 06

A great deal of time 07

Round1 Experiment 2 Political efficacy

Questions in the Main questionnaire

B2 PolCmpl

CARD 6 How often does politics seem so complicated that you can't really understand what is going on? Please use this card.

- | | |
|--------------|---|
| Never | 1 |
| Seldom | 2 |
| Occasionally | 3 |
| Regularly | 4 |
| Frequently | 5 |
| (Don't know) | 8 |

B3 PolActiv

CARD 7 Do you think that you could take an active role⁶ in a group involved with political issues? Please use this card.

- | | |
|---------------------|---|
| Definitely not | 1 |
| Probably not | 2 |
| Not sure either way | 3 |
| Probably | 4 |
| Definitely | 5 |
| (Don't know) | 8 |

B4 PolDcs.

CARD 8 How difficult or easy do you find it to make your mind up⁷ about political issues⁸? Please use this card.

- | | |
|----------------------------|---|
| Very difficult | 1 |
| Difficult | 2 |
| Neither difficult nor easy | 3 |
| Easy | 4 |
| Very easy | 5 |
| (Don't know) | 8 |

Questions in the supplementary questionnaire: group1

Please indicate to what extent you agree or disagree with each of the following statements.

HS4 "Sometimes politics seems so complicated that I can't really understand what is going on²³."

Please tick one box.

- | | | |
|----------------------------|--------------------------|---|
| Disagree strongly | <input type="checkbox"/> | 1 |
| Disagree | <input type="checkbox"/> | 2 |
| Neither disagree nor agree | <input type="checkbox"/> | 3 |
| Agree | <input type="checkbox"/> | 4 |
| Agree strongly | <input type="checkbox"/> | 5 |

HS5 "I think I could take an active role in a group involved with political issues²⁴."

Please tick one box.

- | | | |
|----------------------------|--------------------------|---|
| Disagree strongly | <input type="checkbox"/> | 1 |
| Disagree | <input type="checkbox"/> | 2 |
| Neither disagree nor agree | <input type="checkbox"/> | 3 |
| Agree | <input type="checkbox"/> | 4 |
| Agree strongly | <input type="checkbox"/> | 5 |

HS6 "I find it easy to make my mind up about political issues²⁵."
Please tick one box.

Disagree strongly 1

Disagree 2

Neither disagree nor agree 3

Agree 4

Agree strongly 5

Questions in the supplementary questionnaire: group2

Please indicate to what extent you agree or disagree with each of the following statements.

HS22 "Sometimes politics seems so complicated that I can't really understand what is going on³⁹"
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

HS23 "I think I could take an active role in a group involved with political issues⁴⁰"
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

HS24 "I find it easy to make my mind up about political issues⁴¹"
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

Round1 Experiment 3 Political orientation

Questions in the Main questionnaire

CARD 16 Using this card, please say to what extent you agree or disagree with each of the following statements. **READ OUT EACH STATEMENT AND CODE IN GRID**

		Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
B43 GinvEco	The less that government intervenes ¹⁶ in the economy, the better it is for [country]	1	2	3	4	5	8
B44 GincDif	The government should take measures to reduce differences in income levels	1	2	3	4	5	8
B45 NeedTrU	Employees need strong trade unions to protect their working conditions and wages	1	2	3	4	5	8

Questions in the supplementary questionnaire: group1

Please indicate to what extent you agree or disagree with each of the following statements.

HS16 "The less that government intervenes in the economy, the better it is for [country]³³"
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

HS17 "The government should take measures to reduce differences in income levels³⁴".
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

HS18 "Employees need strong trade unions to protect their working conditions and wages³⁵".
Please tick one box.

Agree strongly 1

Agree 2

Neither agree nor disagree 3

Disagree 4

Disagree strongly 5

Questions in the supplementary questionnaire: group2

HS34 Is it generally good for [country] if government intervenes less in the economy⁵¹? **Please tick one box.**

- Definitely 1
Probably 2
Not sure either way 3
Probably not 4
Definitely not 5

HS35 Should the government take measures to reduce differences in income levels⁵²? **Please tick one box.**

- Definitely 1
Probably 2
Not sure either way 3
Probably not 4
Definitely not 5

HS36 Do employees need strong trade unions to protect their working conditions and wages⁵³? **Please tick one box.**

- Definitely 1
Probably 2
Not sure either way 3
Probably not 4
Definitely not 5

Round1 Experiment 4 Satisfaction

Questions in the Main questionnaire

B30 StfEco

STILL CARD 13: On the whole how satisfied are you with the present state of the economy in [country]? Still use this card.

**Extremely
Dissatisfied**

**Extremely
satisfied (Don't
know)**

00 01 02 03 04 05 06 07 08 09 10 88

B31 StfGov

STILL CARD 13 Now thinking about the [country] government¹³, how satisfied are you with the way it is doing its job? Still use this card.

**Extremely
Dissatisfied**

**Extremely
satisfied (Don't
know)**

00 01 02 03 04 05 06 07 08 09 10 88

B32 StfDem

STILL CARD 13 And on the whole, how satisfied are you with the way democracy¹⁴ works in [country]? Still use this card.

**Extremely
Dissatisfied**

**Extremely
satisfied (Don't
know)**

00 01 02 03 04 05 06 07 08 09 10 88

Questions in the supplementary questionnaire: group1

HS7 On the whole how satisfied are you with the present state of the economy in [country]²⁶?

Please tick one box.

Very dissatisfied 1

Fairly dissatisfied 2

Fairly satisfied 3

Very satisfied 4

HS8 Now thinking about the [country] government, how satisfied are you with the way it is doing its job²⁷?

Please tick one box.

Very dissatisfied 1

Fairly dissatisfied 2

Fairly satisfied 3

Very satisfied 4

HS9 And on the whole, how satisfied are you with the way democracy works in [country]²⁸?
Please tick one box.

- Very dissatisfied 1
Fairly dissatisfied 2
Fairly satisfied 3
Very satisfied 4

Questions in the supplementary questionnaire: group2

HS25 On the whole, how satisfied are you with the present state of the economy in [country]⁴²? Please tick the box that is closest to your opinion, where 0 means extremely dissatisfied and 5 means extremely satisfied.

Extremely
dissatisfied

Extremely
satisfied

- 0 1 2 3 4 5

HS26 Now thinking about the [country] government, how satisfied are you with the way it is doing its job⁴³? **Please tick one box.**

Extremely
dissatisfied

Extremely
satisfied

- 0 1 2 3 4 5

HS27 And on the whole, how satisfied are you with the way democracy works in [country]⁴⁴? **Please tick one box.**

Extremely
dissatisfied

Extremely
satisfied

- 0 1 2 3 4 5

Round1 Experiment 5 Social trust

Questions in the Main questionnaire

A8 PplTrst

CARD 3: Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful³ in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.

<i>You can't be too careful</i>	01	02	03	04	05	06	07	08	09	<i>Most people can be trusted</i>	<i>(Don't know)</i>
00										10	88

A9 PplFair

CARD 4: Using this card, do you think that most people would try to take advantage⁴ of you if they got the chance, or would they try to be fair?

<i>Most people would try to take advantage of me</i>	01	02	03	04	05	06	07	08	09	<i>Most people would try to be fair</i>	<i>(Don't know)</i>
00										10	88

A10* PplHlp

CARD 5: Would you say that most of the time people try to be helpful⁵ or that they are mostly looking out for themselves? Please use this card.

<i>People mostly look out for themselves</i>	01	02	03	04	05	06	07	08	09	<i>People mostly try to be helpful</i>	<i>(Don't know)</i>
00										10	88

Questions in the supplementary questionnaire: group1

HS10 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people²⁹? Please tick the box that is closest to your opinion, where 0 means you can't be too careful and 5 means that most people can be trusted.

<i>You can't be too careful</i>						<i>Most people can be trusted</i>
0	1	2	3	4	5	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

HS11 Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair³⁰? **Please tick one box.**

<i>Most people would try to take advantage of me</i>						<i>Most people would try to be fair</i>
0	1	2	3	4	5	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

HS12 Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves³¹? **Please tick one box.**

<i>People mostly look out for themselves</i>						<i>People mostly try to be helpful</i>
0	1	2	3	4	5	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Questions in the supplementary questionnaire: group2

HS28 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people⁴⁵? **Please tick one box.**

You can't be too careful 1

Most people can be trusted 2

HS29 Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair⁴⁶? **Please tick one box.**

Most people would try to take advantage of me 1

Most people would try to be fair 2

HS30 Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves⁴⁷? **Please tick one box.**

People mostly look out for themselves 1

People mostly try to be helpful 2

Round1 Experiment 6 Political trust

Questions in the Main questionnaire

CARD 11: Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. 0 means you do not trust an institution at all, and 10 means you have complete trust. Firstly...**READ OUT**

		No trust At all										Complete trust	(Don't know)
B7 TrstPri	... [country]'s parliament?	00	01	02	03	04	05	06	07	08	09	10	88
B8 TrstLgl	... the legal system?	00	01	02	03	04	05	06	07	08	09	10	88
B9 TrstPlc	... the police?	00	01	02	03	04	05	06	07	08	09	10	88

Questions in the supplementary questionnaire: group1

Please indicate on a score of 0 to 10 how much you personally trust each of the institutions below. 0 means you do not trust an institution at all, and 10 means you have complete trust³².

Please tick the box that is closest to your opinion.

		No trust at all										Complete trust
HS13	[Country]'s parliament	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HS14	The legal system	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
HS15	The police	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions in the supplementary questionnaire: group2

HS31 Please say on a scale of 0 to 10 how much you trust **[country]'s parliament**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the parliament, the higher the score should be⁴⁸.

Your score:

HS32 Please say on a scale of 0 to 10 how much you trust the **legal system**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the legal system, the higher the score should be⁴⁹.

Your score:

HS33 Please say on a scale of 0 to 10 how much you trust the **police**. If you have no trust at all give a score of 0. If you have complete trust, give a score of 10. The more you trust the police, the higher the score should be⁵⁰.

Your score:

Round 2 Experiment 1 Work in the house

Questions in the Main questionnaire

G22 CARD 64 I'd now like to talk about housework, as described on the card. By housework, we mean things done around the home, such as cooking, washing, cleaning, care of clothes, shopping, maintenance of property, but not including childcare and leisure activities. On a typical weekday about how many hours, in total, do people in your household spend on housework for your home?

NOTE TO INTERVIEWER: CODE TO NEAREST HOUR. ACCEPT ESTIMATE.

WRITE IN:

(Don't know) 88

G23 CARD 65 And about how much of this time do you spend yourself? Please use this card.

None or almost none	01
Up to a quarter of the time	02
More than a quarter, up to a half of the time	03
More than a half, up to three quarters of the time	04
More than three quarters, less than all of the time	05
All or nearly all of the time	06
(Don't know)	88

G24 STILL CARD 65 And about how much of this time does your husband/wife/partner spend on housework?
Please use this card.

None or almost none	01
Up to a quarter of the time	02
More than a quarter, up to a half of the time	03
More than a half, up to three quarters of the time	04
More than three quarters, less than all of the time	05
All or nearly all of the time	06
(Don't know)	88

Questions in the supplementary questionnaire: group1

iS2²⁰ We'd now like to ask you about housework

By housework, we mean things done around the home, such as cooking, washing, cleaning, care of clothes, shopping, maintenance of property, but not including childcare, looking after other people and leisure activities. On a typical weekday about how many hours, in total, do people in your household spend on housework for your home?

WRITE IN HOURS:

iS3²¹ And about how many hours of these hours do you spend yourself?

WRITE IN HOURS:

iS4²² And about how many does your husband / wife / partner spend?

WRITE IN HOURS:

Questions in the supplementary questionnaire: group2

iS15⁵¹ We'd now like to ask you about housework

By housework, we mean things done around the home, such as cooking, washing, cleaning, care of clothes, shopping, maintenance of property, but not including childcare, looking after other people and leisure activities. On a typical weekday about how many hours, in total, do people in your household spend on housework for your home?

WRITE IN HOURS:

iS16⁵² And what percentage of this time do you spend yourself? 0% means 'absolutely none' and 100% means 'absolutely all'.

WRITE IN PERCENTAGE:

iS17⁵³ And what percentage of this time does you husband / wife / partner spend? 0% means 'absolutely none' and 100% means 'absolutely all'.

WRITE IN PERCENTAGE:

Round 2 Experiment 2 Contact with doctor

Questions in the Main questionnaire

CARD 31 Using this card, please indicate how often you think the following applies to doctors in general:

	Never or almost never	Some of the time	About half of the time	Most of the time	Always or almost always	(Don't know)
D25 Doctors keep the whole truth ⁴¹ from their patients.	1	2	3	4	5	8
D26 GPs ⁴² treat their patients as their equals.	1	2	3	4	5	8
D27 Before doctors decide on a treatment, they discuss it with their patient.	1	2	3	4	5	8

Questions in the supplementary questionnaire: group1

ALL

Please indicate how much you agree or disagree with each of the following statements about doctors in general.

iS5²³ "Doctors rarely keep the whole truth from their patients"

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS6²⁴ "GPs rarely treat their patients as their equals"

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS7²⁵ "Before doctors decide on a treatment, they rarely discuss it with their patient ."

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

Questions in the supplementary questionnaire: group2

Please indicate how much you agree or disagree with each of the following statements about doctors in general.

iS28⁸² "Doctors usually keep the whole truth from their patients"

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS29⁸³ "GPs usually treat their patients as their equals"

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS30⁸⁴ "Before doctors decide on a treatment, they usually discuss it with their patient."

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

Round 2 Experiment 3 Job evaluation

Questions in the Main questionnaire

CARD 73 Using this card, please tell me how true each of the following statements is about your current job.

	Not at all true 1	A little true 2	Quite true 3	Very true 4	(Don't know) 8
G64 There is a lot of variety in my work.					
G66 My job is secure ⁹³	1	2	3	4	8
G70 My health or safety is at risk because of my work.	1	2	3	4	8

Questions in the supplementary questionnaire: group1

iS19⁵⁴ The next 3 questions are about your current job. Please choose one of the following to describe how varied your work is.

Please tick one box.

Not at all varied 1

A little varied 2

Quite varied 3

Very varied 4

iS20⁵⁵ Please choose one of the following to describe how secure your job is.

Please tick one box.

Not at all secure 1

A little secure 2

Quite secure 3

Very secure 4

iS21⁵⁶ Please choose one of the following to say how much, if at all, your work puts your health and safety at risk.

Please tick one box.

Not at all at risk 1

A little at risk 2

Quite a lot at risk 3

Very much at risk 4

Questions in the supplementary questionnaire: group2

iS32⁸⁵ Please indicate, on a scale of 0 to 10, how varied your work is, where 0 is not at all varied and 10 is very varied.

Please tick the box that is closest to your opinion

**Not at
all varied**

**Very
varied**

0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

iS33⁸⁶ Now please indicate, on a scale of 0 to 10, how secure your job is, where 0 is not at all secure and 10 is very secure.

Please tick the box that is closest to your opinion

**Not at
all secure**

**Very
secure**

0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

iS34⁸⁷ Please indicate, on a scale of 0 to 10, how much your health and safety is at risk from your work, where 0 is not at all at risk and 10 is very much at risk.

Please tick the box that is closest to your opinion

**Not at
all at risk**

**Very much
at risk**

0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Round 2 Experiment 4 Role of woman

Questions in the Main questionnaire

CARD 59 I am now going to read out some statements about men and women and their place⁸¹ in the family. Using this card, please tell me how much you agree or disagree with the following statements.

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
G6 A woman should be prepared to cut down on her paid work for the sake of her family. ⁸²	1	2	3	4	5	8
G7 Men should take as much responsibility as women for the home and children.	1	2	3	4	5	8
G8 When jobs are scarce, men should have more right ⁸³ to a job than	1	2	3	4	5	8

Questions in the supplementary questionnaire: group1

Please indicate how much you agree or disagree with each of the following statements about men and women and their place in the family.

iS8²⁶ "A women should not have to cut down on her paid work for the sake of her family."

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS9²⁷ "Women should take more responsibility for the home and children than men."

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

iS10²⁸ "When jobs are scarce, women should have the same right to a job as men."

Please tick one box.

Agree strongly 1

Agree 2

Neither disagree nor agree 3

Disagree 4

Disagree strongly 5

Questions in the supplementary questionnaire: group2

ALL

iS22⁵⁷ If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.

Please tick one box.

A woman should be prepared to cut down on her paid work for the sake of her family

1

2

3

4

5

A woman should not have to cut down on her paid work for the sake of her family

iS23⁵⁸ If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.

Please tick one box.

Men should take as much responsibility as women for the home and children

1

2

3

4

5

Women should take more responsibility for the home and children than men

iS24⁵⁹ If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.

Please tick one box.

When jobs are scarce, men should have more right to a job than women

1

2

3

4

5

When jobs are scarce, women should have the same right to a job as men

Round 2 Experiment 5 Satisfaction with the government

Questions in the Main questionnaire

B25 STILL CARD 10: On the whole how satisfied are you with the present state of the economy in [country]? Still use this card.

Extremely dissatisfied																		Extremely satisfied (Don't know)
		00	01	02	03	04	05	06	07	08	09	10	88					

B26 STILL CARD 10 Now thinking about the [country] government¹⁵, how satisfied are you with the way it is doing its job? Still use this card.

Extremely dissatisfied																		Extremely satisfied (Don't know)
		00	01	02	03	04	05	06	07	08	09	10	88					

B27 STILL CARD 10 And on the whole, how satisfied are you with the way democracy¹⁶ works in [country]? Still use this card.

Extremely dissatisfied																		Extremely satisfied (Don't know)
		00	01	02	03	04	05	06	07	08	09	10	88					

Questions in the supplementary questionnaire: group1

is11²⁹ On the whole how satisfied are you with the present state of the economy in [country]?
Please tick the box that is closest to your opinion.

Extremely dissatisfied											Extremely satisfied
	Neither satisfied nor dissatisfied										
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

is12³⁰ Now thinking about the [country] government, how satisfied are you with the way it is doing its job?
Please tick the box that is closest to your opinion.

Extremely dissatisfied											Extremely satisfied
	Neither satisfied nor dissatisfied										
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

is13³¹ And on the whole, how satisfied are you with the way democracy works in [country]?
Please tick the box that is closest to your opinion.

Extremely dissatisfied											Extremely satisfied
	Neither satisfied nor dissatisfied										
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions in the supplementary questionnaire: group2

iS35⁸⁸ On the whole how satisfied are you with the present state of the economy in [country]?

Please tick the box that is closest to your opinion.

Very dissatisfied											Very satisfied
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

iS36⁸⁹ Now thinking about the [country] government, how satisfied are you with the way it is doing its job?

Please tick the box that is closest to your opinion.

Very dissatisfied											Very satisfied
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

iS37⁹⁰ And on the whole, how satisfied are you with the way democracy works in [country]?

Please tick the box that is closest to your opinion.

Very dissatisfied											Very satisfied
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Round 2 Experiment 6 Political trust

Questions in the Main questionnaire

CARD 8: Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. 0 means you do not trust an institution at all, and 10 means you have complete trust. Firstly...**READ OUT...**

		No trust at all										Complete trust	(Don't know)
B4	... [country]'s parliament?	00	01	02	03	04	05	06	07	08	09	10	88
B5	... the legal system?	00	01	02	03	04	05	06	07	08	09	10	88
B7	... politicians?	00	01	02	03	04	05	06	07	08	09	10	88

Questions in the supplementary questionnaire: group1

Please indicate on a scale of 0 to 10 how much you personally trust each of the institutions below. 0 means you do not trust an institution at all, and 10 means you have complete trust.

Please tick the box that is closest to your opinion.

		No trust at all										Complete trust
iS25 ⁶⁰	[Country]'s parliament	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
iS26 ⁶¹	The legal system	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
iS27 ⁶²	politicians	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions in the supplementary questionnaire: group2

Please indicate on a scale of 0 to 10 how much you personally trust each of the institutions below. 0 means you do not trust an institution at all, and 10 means you have complete trust.

Please tick the box that is closest to your opinion.

		No trust at all										Complete trust
iS38 ⁹¹	[Country]'s parliament	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
iS39 ⁹²	The legal system	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
iS40 ⁹³	Politicians	0	1	2	3	4	5	6	7	8	9	10
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Round 3 Experiment 1 Well being

Questions in the Main questionnaire

CARD 39 Using this card, please say to what extent you agree or disagree with each of the following statements.

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
E40 I generally feel that what I do in my life is valuable and worthwhile ⁵⁵	1	2	3	4	5	8
E43 There are people in my life who really care about me	1	2	3	4	5	8
E45 I feel close to ⁵⁶ the people in my local area	1	2	3	4	5	8

Questions in the supplementary questionnaire: group1

Next some questions about how you feel about yourself and your life.

Please read each question and tick the box on each line that shows how much you agree or disagree with each statement.

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly
HS10²⁹ I generally feel that what I do in my life is valuable and worthwhile.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
HS11³⁰ There are people in my life who really care about me.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
HS12³¹ I feel close to the people in my local area.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5

Questions in the supplementary questionnaire: group 2

Please indicate to what extent you agree or disagree with each of the following statements

HS22⁶⁰ 'I generally feel that what I do in my life is valuable and worthwhile.'

Please tick one box.

- Agree strongly 1
- Agree 2
- Neither disagree nor agree 3
- Disagree 4
- Disagree strongly 5

HS23⁶¹ 'There are people in my life who really care about me.'

Please tick one box.

- Agree strongly 1
- Agree 2
- Neither disagree nor agree 3
- Disagree 4
- Disagree strongly 5

HS24⁶² 'I feel close to the people in my local area.'

Please tick one box.

- Agree strongly 1
- Agree 2
- Neither disagree nor agree 3
- Disagree 4
- Disagree strongly 5

Questions in the supplementary questionnaire: group3

Please indicate the extent to which you agree or disagree with the following statements.

HS34⁹¹ 'I generally feel that what I do in my life is valuable and worthwhile'.
Please tick one box.

Disagree strongly							Agree strongly
01	02	03	04	05	06	07	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS35⁹² 'There are people in my life who really care about me'.
Please tick one box.

Disagree strongly							Agree strongly
01	02	03	04	05	06	07	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS36⁹³ 'I feel close to the people in my local area'.
Please tick one box.

Disagree strongly							Agree strongly
01	02	03	04	05	06	07	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Round 3 Experiment 2 Consequences of immigration

Questions in the Main questionnaire

B38 CARD 15 Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? Please use this card.

Bad for the economy					Good for the economy					(Don't know)	
00	01	02	03	04	05	06	07	08	09	10	88

B39 CARD 16 And, using this card, would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?

Cultural life undermined					Cultural life enriched					(Don't know)	
00	01	02	03	04	05	06	07	08	09	10	88

B40 CARD 17 Is [country] made a worse or a better place to live by people coming to live here from other countries? Please use this card.

Worse place to live					Better place to live					(Don't know)	
00	01	02	03	04	05	06	07	08	09	10	88

Questions in the supplementary questionnaire: group1

- HS4²³** It is generally bad for [country's] economy that people come to live here from other countries 1 2 3 4 5
- HS5²⁴** [Country's] cultural life is generally undermined by people coming to live here from other countries 1 2 3 4 5
- HS6²⁵** [Country] is made a worse place to live by people coming to live here from other countries 1 2 3 4 5

Questions in the supplementary questionnaire: group2

HS16⁵⁴ How much do you agree or disagree that it is generally bad for [Country] 's economy that people come to live here from other countries?
Please tick one box.

Disagree strongly		Agree strongly								
0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS17⁵⁵ And how much do you agree or disagree that [Country] 's cultural life is generally undermined by people coming to live here from other countries?
Please tick one box.

Disagree strongly		Agree strongly								
0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS18⁵⁶How much do you agree or disagree that [Country] is made a worse place to live by people coming here from other countries?
Please tick one box.

Disagree strongly		Agree strongly								
0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions in the supplementary questionnaire: group3

**Now some questions about people from other countries coming to live in [country].
Please read each question and tick the box on each line that shows how much you agree or disagree with each of the following statements.**

	Disagree strongly						Agree strongly
IS28⁸⁵ It is generally bad for [country's] economy that people come to live here from other countries	<input type="checkbox"/> 01	<input type="checkbox"/> 02	<input type="checkbox"/> 03	<input type="checkbox"/> 04	<input type="checkbox"/> 05	<input type="checkbox"/> 06	<input type="checkbox"/> 07
IS29⁸⁶ [Country's] cultural life is generally undermined by people coming to live here from other countries	<input type="checkbox"/> 01	<input type="checkbox"/> 02	<input type="checkbox"/> 03	<input type="checkbox"/> 04	<input type="checkbox"/> 05	<input type="checkbox"/> 06	<input type="checkbox"/> 07
IS30⁸⁷ [Country] is made a worse place to live by people coming to live here from other countries	<input type="checkbox"/> 01	<input type="checkbox"/> 02	<input type="checkbox"/> 03	<input type="checkbox"/> 04	<input type="checkbox"/> 05	<input type="checkbox"/> 06	<input type="checkbox"/> 07

Round 3 Experiment 3 Allowing more immigrants

Questions in the Main questionnaire

Now some questions about people from other countries coming to live in [country].

B35 CARD 14 Now, using this card, to what extent do you think [country] should¹⁸ allow people of the same race or ethnic group as most [country's] people to come and live here¹⁹?

- Allow many to come and live here 1
- Allow some 2
- Allow a few 3
- Allow none 4
- (Don't know) 8

B36 STILL CARD 14 How about people of a different race or ethnic group from most [country] people? Still use this card.

- Allow many to come and live here 1
- Allow some 2
- Allow a few 3
- Allow none 4
- (Don't know) 8

B37 STILL CARD 14 How about people from the poorer countries outside Europe? Use the same card.

- Allow many to come and live here 1
- Allow some 2
- Allow a few 3
- Allow none 4
- (Don't know) 8

Questions in the supplementary questionnaire: group1

Now some questions about people from other countries coming to live in [country]. Please read each question and tick the box on each line that shows how much you agree or disagree with each statement.

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly
HS1²⁰ [Country] should allow more people of the <u>same race or ethnic group</u> as most [country's] people to come and live here	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
HS2²¹ [Country] should allow more people of a <u>different</u> race or ethnic group from most [country's] people to come and live here	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
HS3²² [Country] should allow more people from the <u>poorer countries outside Europe</u> to come and live here	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5

Questions in the supplementary questionnaire: group 2

Now some questions about people from other countries coming to live in [country].

HS13⁵¹ To what extent do you think [country] should allow people of the same race or ethnic group as most [country's] people to come and live here?
Please tick one box.

[country's] policy should be to...

Allow many to come and live here 1

Allow some 2

Allow a few 3

Allow none 4

HS14⁵² How about people of a different race or ethnic group from most [country] people?
Please tick one box.

[country's] policy should be to...

Allow many to come and live here 1

Allow some 2

Allow a few 3

Allow none 4

HS15⁵³ How about people from the poorer countries outside Europe?
Please tick one box.

[country's] policy should be to...

Allow many to come and live here 1

Allow some 2

Allow a few 3

Allow none 4

Questions in the supplementary questionnaire: group3

Now some questions about people from other countries coming to live in [country].
Please read each question and tick the box on each line that shows how much you agree or disagree with each of the following statements.

		Disagree strongly						Agree strongly							
IS25⁸²	[Country] should allow more people of the <u>same race or ethnic group</u> as most [country's] people to come and live here	<input type="checkbox"/>	01	<input type="checkbox"/>	02	<input type="checkbox"/>	03	<input type="checkbox"/>	04	<input type="checkbox"/>	05	<input type="checkbox"/>	06	<input type="checkbox"/>	07
IS26⁸³	[Country] should allow more people of a <u>different</u> race or ethnic group from most [country's] people to come and live here	<input type="checkbox"/>	01	<input type="checkbox"/>	02	<input type="checkbox"/>	03	<input type="checkbox"/>	04	<input type="checkbox"/>	05	<input type="checkbox"/>	06	<input type="checkbox"/>	07
IS27⁸⁴	[Country] should allow more people from the <u>poorer countries outside Europe</u> to come and live here	<input type="checkbox"/>	01	<input type="checkbox"/>	02	<input type="checkbox"/>	03	<input type="checkbox"/>	04	<input type="checkbox"/>	05	<input type="checkbox"/>	06	<input type="checkbox"/>	07

Round 3 Experiment 4 Learn new things

Questions in the Main questionnaire

CARD 35 Using this card, please tell me to what extent you agree or disagree with each of the following statements. **READ OUT EACH STATEMENT AND CODE IN GRID**

		Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
E26	I love learning new things.	1	2	3	4	5	8
E27	Most days I feel a sense of accomplishment from what I do.	1	2	3	4	5	8
E28	I like planning and preparing for the future.	1	2	3	4	5	8

Questions in the supplementary questionnaire: group1

Next some questions about how you feel about yourself and your life.

Please read each question and tick the box on each line that shows how much you agree or disagree with each statement.

		Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly
HS7²⁶	I love learning new things.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
HS8²⁷	Most days I feel a sense of accomplishment from what I do.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 0
HS9²⁸	I like planning and preparing for the future.	<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5

Questions in the supplementary questionnaire: group2

And now some questions about how you feel about yourself and your life. Please indicate to what extent each of the following statements applies to you.

HS19⁵⁷ 'I love learning new things'
Please tick one box.

Not at all											Very much
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS20⁵⁸ 'Most days I feel a sense of accomplishment from what I do'
Please tick one box.

Not at all											Very much
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS21⁵⁹ 'I like planning and preparing for the future'
Please tick one box.

Not at all											Very much
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions in the supplementary questionnaire: group3

And now some questions about how you feel about yourself and your life. Please indicate the extent to which you agree or disagree with each of the following statements.

HS31⁸⁸ 'I love learning new things.'
Please tick one box.

Disagree Strongly											Agree Strongly
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS32⁸⁹ 'Most days I feel a sense of accomplishment from what I do.'
Please tick one box.

Disagree Strongly											Agree Strongly
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

HS33⁹⁰ 'I like planning and preparing for the future'.
Please tick one box.

Disagree strongly											Agree strongly
0	1	2	3	4	5	6	7	8	9	10	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>