Why are there differences in measurement quality across countries?

Daniel Oberski
Willem E. Saris

ESADE
Barcelona

Jacques Hagenaars

University of Tilburg

Abstract

The European Social Survey (ESS) has the unique characteristic that in more than 20 countries the same questions are asked and that within each round of the ESS Multitrait-Multimethod (MTMM) experiments are built in to evaluate the quality of a limited number of questions. This gives us an exceptional opportunity to observe the differences in quality of questions over a large number of countries. The MTMM experiments make it possible to estimate the reliability, validity and method effects of single questions (Andrews 1984, Saris and Andrews 1991, Saris et al 2004). The product of the reliability and the validity can be interpreted as the explained variance in the observed variable by the variable one would like to measure. It is a measure of the total quality of a question.

These MTMM experiments showed that there are considerable differences in measurement quality across countries. Because these differences in quality can cause wrong conclusions with respect to differences in relationships across countries this paper studies several reasons for these differences. The following explanations are considered: differences in translation, differences in the MTMM design and difference in complexity of the question formulation in the different countries.

It turned out that the main reasons for the quality differences are differences in the formulation of the questions in the different countries and differences in the MTMM design. The complexity of the question formulation did not contribute much to the explanation.

1. Introduction

In the ESS a lot of time, money, and effort is spent to make the questions as functionally equivalent across countries as possible (Harkness 2002, 2007) and to make the samples as comparable as possible (Haeder and Lynn 2007). Nevertheless, considerable differences in quality of the questions can be observed across countries (see Table 4). In round 2 of the ESS the largest difference found was between questions in Sweden with a quality of .4 and in Portugal with a quality above .9. The Scandinavian countries had an average quality around .5 over 54 questions while other countries such as Greece, Portugal and Estonia had an average quality of .8. To study these differences is important because these differences can cause differences in relationships between variables in different countries which have no substantive meaning but are just caused by differences in quality in the measurement (Saris and Gallhofer 2007). In order to avoid such differences it is also important to study reasons for these differences in quality. In this paper we want to explore several possible explanations for these differences. But before we can do this we have to show what we mean by quality of measures.

In Figure 1 we show the basic response model (Saris and Gallhofer 2007) we use as our starting point.

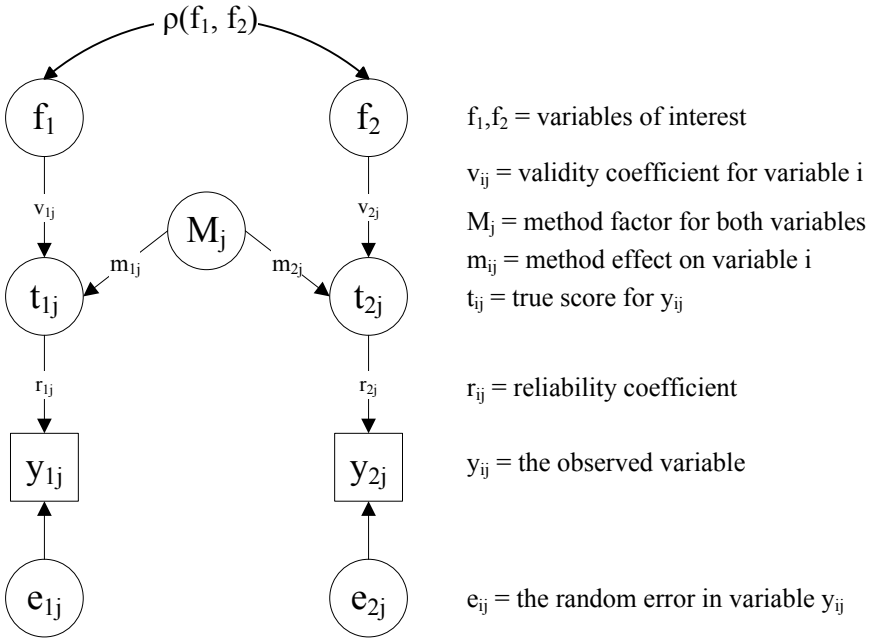

Figure 1: The response model used in the MTMM experiments.

The difference between the observed response (y) and the so called "true score"(t) is random measurement error (e). So the coefficient r represents the reliability coefficient and r^2 is the reliability.

The difference between the true score and the concept by intuition (f_1) are the respondents' systematic reactions to the method (m). So the coefficient v represents the true score validity coefficient and v^2 is the true score validity. The quality of a measure (q^2) is defined as $q^2 = r^2.v^2$ and q is the quality coefficient. The correlation between the unobserved variables of interest is denoted by $\rho(f_1,f_2)$.

Several remarks should be made. The first is that the correlation $\rho(y_i, y_j)$ between two observed variables is:

$$\rho(y_i, y_j) = \overbrace{\rho(f_i,f_j)}^{\text{Correlation of interest}} \cdot \overbrace{q_i \cdot q_j}^{\text{Attenuation factor}} + \overbrace{r_i \cdot r_j \cdot m_i \cdot m_j}^{\text{Correlation due To method}} \qquad (1)$$

This means that the correlation between the observed variables is normally smaller than the correlation between the variables of interest, but can be larger if the method effects are considerable. A second remark to make is that one can not compare correlations across countries without correction for measurement error if the measurement quality coefficients are very different across countries: this follows directly from the above equation (1). A third point is that one can not estimate these quality indicators from this simple design with two observed variables. In this model there are two reliability coefficients, two validity coefficients, two method effects and one correlation between the two latent traits, leaving us with seven unknown parameters, while only one correlation can be obtained from the data. It is impossible to estimate these seven parameters from just one correlation.

There are two different approaches to estimate these coefficients. The first is the use of MTMM experiments. The second is the use of the prediction program SQP that is based on a large number of MTMM experiments. Because both approaches will be used below, we will briefly introduce both procedures. A more elaborate introduction can be found in Saris and Gallhofer 2007.

1.1 MTMM experiments

Campbell and Fiske (1959) suggested using multiple traits and multiple methods (MTMM). The classical MTMM approach recommends the use of a minimum of three traits that are measured with three different methods leading to nine different observed variables. An example of such a design is given in Table 1.

Table 1. The classic MTMM design used in the ESS pilot study

The three traits were presented by the following three requests:
- *On the whole, how satisfied are you with the present state of the economy in Britain?*
- *Now think about the national 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 Britain?*

The three methods are specified by the following response scales:

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

Very dissatisfied *Very satisfied*

0 1 2 3 4 5 6 7 8 9 10

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

---

Collecting data using this MTMM design, data for nine variables are obtained and from that data a correlation matrix of 9×9 is obtained. The model formulated to estimate the reliability, validity, and method effects is an extension of the model

presented in Figure 1. This Figure illustrates the relationships between the true scores and their 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. To reduce the complexity of the figure, it is not indicated 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.

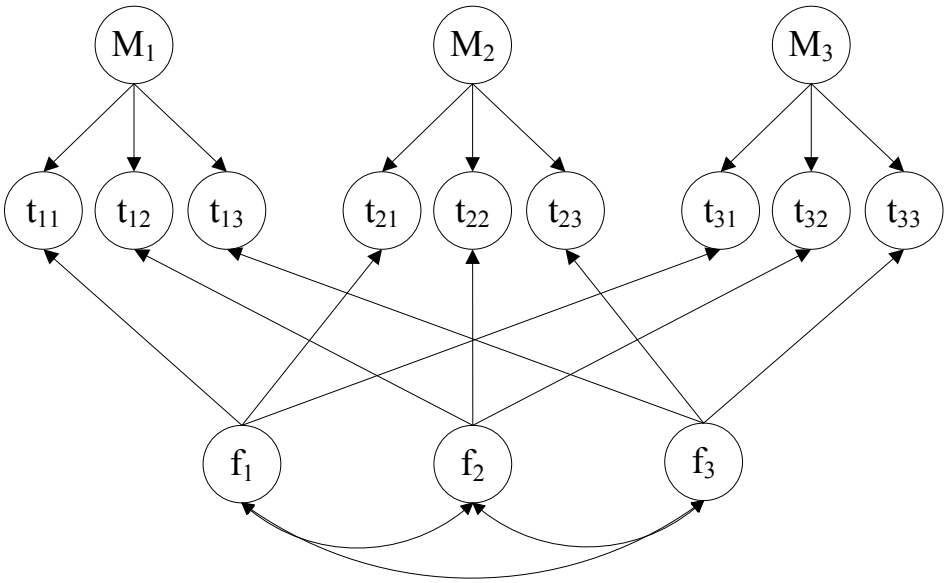

Figure 2: MTMM model illustrating the true scores and their factors of interest.

The MTMM design of 3 traits and 3 methods generates 45 covariances 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. In total there are 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.

Table 2 presents the correlations that we derived between the 9 measures obtained from a sample of 481 people in the British population. Using the specifications of the model indicated above and the ML estimator to estimate the quality indicators, the results presented in Table 3 are obtained[1].

[1] In this case the ML estimator is used. The estimation is done using the covariance matrix as the input matrix and not the correlation matrix . Thereafter, the estimates are standardized to obtain the requested coefficients. A result of this is that the standardized method effects are not exactly equal to each other.

Table 2: The correlations between the 9 variables of the MTMM experiment with respect to satisfaction with political outcome

	Method 1			**Method 2**			**Method 3**		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Method 1									
Q1	1.00								
Q2	.481	1.00							
Q3	.373	.552	1.00						
Method 2									
Q1	-.626	-.422	-.410	1.00					
Q2	-.429	-.663	-.532	.642	1.00				
Q3	-.453	-.495	-.669	.612	.693	1.00			
Method 3									
Q1	-.502	-.374	-.332	.584	.436	.438	1.00		
Q2	-.370	-.608	-.399	.429	.653	.466	.556	1.00	
Q3	-.336	-.406	-.566	.406	.471	.638	.514	.558	1.00
Means	2.42	2.71	2.45	5.26	4.37	5.13	2.01	1.75	2.01
Standard Deviation	.77	.76	.84	2.29	2.37	2.44	.72	.71	.77

The estimated parameter values presented in Table 3 point to method 2 having the highest reliability for these traits. With respect to validity, the first two methods have the highest scores and are approximately equal. When considering all estimates method 2 is preferable to the other methods.

Table 3: Standardized estimates of the MTMM model specified for the ESS data of Table 2

	Validity coefficients			Method effects			Reliability coefficients
	F_1	F_2	F_3	M_1	M_2	M_3	
T_{11}	.93			.36			.79
T_{21}		.94		.35			.85
T_{31}			.95	.33			.81
T_{12}	.91				.41		.91
T_{22}		.92			.39		.94
T_{32}			.93		.38		.93
T_{13}	.85					.52	.82
T_{23}		.87				.50	.87
T_{33}			.88			.48	.84

Note that the validity and the method effects do not have to be evaluated separately because they complement each other, as was mentioned previously: $v_{ij}^2 = 1-m_{1j}^2$. With this example we have shown how the MTMM approach can be used to evaluate the quality of several survey items with respect to validity and reliability using MTMM experiments.

1.2. Use of the program SQP

The experiment presented in the previous section is typical of the MTMM experiments of the last 30 years. Such studies have been conducted by Andrews (1984) and Rodgers et al. (1992) in the United States. Költringer (1995) has conducted a similar study for German questionnaires, while Scherpenzeel and Saris (1996) in the Netherlands and Billiet and Waege in Belgium have conducted similar studies regarding Dutch questionnaires. In total, 87 MTMM studies are available containing 1023 survey items. All of these studies are based on at least regional samples of the general population. In the United States, the Detroit area was studied, in Austria and the Netherlands national samples were used, while in Belgium random samples of the Flemish-speaking part of the population were taken. The topics in the different experiments are highly diverse. In general, the MTMM experiments are integrated into normal survey research where three or more questions of the survey are used for further experimentation. This approach guarantees that questions that are common to survey research are used. The same is true for the variation in the choices made in the design of survey items. The experiments are designed for the most commonly used methods (choices). More details on the studies are presented in Saris and Gallhofer (2007).

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 a list of 50 characteristics. 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 website[2]. 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.

The following equation presents the approach used to estimate the effects of the question characteristics on the quality criteria.

$$C = a + b_{11}D_{11} + b_{21}D_{21} + \ldots + b_{12}D_{12} + b_{22}D_{22} + \ldots + b_3N_{cat} + \ldots + e \qquad (2)$$

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 results of the analysis with this model based on more than 1000 questions have been reported in Saris and Gallhofer (2007).

Suppose that a survey designer would like to conduct a survey and, before going into the field, would like to evaluate the quality of the proposed survey items of the questionnaire using the information from this meta-analysis. This would necessitate coding all the items on the variables in the classification system, and after that applying the prediction from this study on all these items, in order to determine the total score for reliability and validity. This clearly is a great deal of work. It would, therefore, be advantageous to have a computer program that could help in this task. This is what the

[2] Details of the codebook can be found at http://www.sqp.nl.

program Survey Quality Prediction (SQP) does. The user codes characteristics of the questions, and the program subsequently predicts the values of the reliability, validity, method effect and the total quality of the questions.

In the next section we will show how large the differences in quality of questions can be across countries. These results have been obtained using the same MTMM experiments in 18 different countries. Both procedures mentioned above will be used to explain the differences in quality of these questions.

2. The data

The European Social Survey (ESS) has the unique characteristic that in more than 20 countries the same questions were asked and that within each round of the ESS Multitrait-Multimethod (MTMM) experiments are built in to evaluate the quality of a limited number of questions. This gives us an exceptional opportunity to observe the differences in quality of questions over a large number of countries. In this paper we have used the MTMM experiments of round 2 of the ESS. The topics of the 6 MTMM experiments in the second round of the ESS were the following:

1. Time spent on housework
2. The social distance between the doctor and patients
3. Opinions about job
4. The role of men and women in society
5. Satisfaction with the political situation
6. Political trust

Concerning each of these topics 3 questions were asked and these three questions were presented in 3 different forms following the discussed MTMM designs (Campbell & Fiske, 1959). The first form, used for all respondents, was presented in the main questionnaire. The two alternative forms were presented in a supplementary questionnaire which was filled in after the main questionnaire. All respondents were only asked to reply to one alternative form but different groups got different version of the same questions (Saris et al. 2004). For the specific questions for the 6 experiments we refer to the ESS website where English source version of all questions are presented[3], and for the different translations we refer to the ESS archive[4].

Each experiment varies a different aspect of the method by which questions can be asked in questionnaires. The 'housework' experiment compares numeric estimates by respondents with other scales. The 'doctors' experiment examines the effect of choosing arbitrary scale positions as a starting point for agreement-disagreement with a statement. The 'job' experiment compares a 4 point with an 11 point scale and a true-false scale with a direct question. In the 'women' experiment agree-disagree scales are reversed, there is one negative item, and a 'don't know' category is omitted in one of the methods. The 'satisfaction' experiment varies the extremeness and number of fixed reference points of the scale. And finally, the experiment on political trust was meant to investigate the effect of repeating the same question in the same format.

A special group took care that the samples in the different countries where proper probability samples and as comparable as possible (Haeder and Lynn 2007).

[3] http://europeansocialsurvey.org/

[4] http://ess.nsd.uib.no/

The questions asked in the different countries have been translated from the English source questionnaire. An optimal effort has been made to make these questions as equivalent as possible and to avoid errors. In order to reach this goal two translators independently translated the source questionnaire and a third person was involved to choose the optimal translation by consensus if differences were found. For details of this procedure we refer to the work of Harkness (2007).

Despite these efforts to make the data as comparable as possible, big differences in measurement quality were found across the different countries. Table 4 shows the mean and median standardised quality of the questions in the main questionnaire across the experiments for the different countries.

Table 4: the quality of all 18 questions included in the experiments in the main questionnaire.

Country	Mean	Median	Minimum	Maximum
Portugal	0.79	0.81	0.63	0.91
Switzerland	0.79	0.84	0.56	0.90
Greece	0.78	0.79	0.64	0.90
Estonia	0.78	0.85	0.58	0.90
Poland	0.73	0.85	0.51	0.90
Luxembourg	0.72	0.73	0.53	0.88
United Kingdom	0.70	0.71	0.56	0.82
Denmark	0.70	0.70	0.52	0.80
Belgium	0.70	0.73	0.46	0.90
Germany	0.69	0.70	0.53	0.83
Spain	0.69	0.64	0.54	0.90
Austria	0.68	0.68	0.51	0.85
Czech Republic	0.65	0.60	0.52	0.87
Slovenia	0.63	0.60	0.46	0.82
Norway	0.59	0.59	0.35	0.83
Sweden	0.58	0.58	0.43	0.68
Finland	0.57	0.54	0.42	0.78
All	0.69	0.69	0.35	0.91

A remarkable phenomenon in this table is that the Scandinavian countries have the lowest quality of all while the highest quality has been obtained in Portugal, Switzerland, Greece, and Estonia. The others countries are in between these two groups.

The differences are considerable and statistically significant across countries (F=3.19, df=16, p<.001) and experiments (F=92.65, df=5, p<.0001)[5]. The highest mean quality is .79 in Portugal while the lowest is .57 in Finland. If the correlation between the constructs of interest is .6 in both countries and the measures for these variables have the above quality then the observed correlation in Portugal would be .474 while the observed correlation in Finland would be .342. Most people would say that this is a large difference in correlations which requires a substantive explanation. But this difference can be expected because of differences in data quality and has no substantive meaning at all.

[5] The significance of the differences in the quality coefficients was determined using their observed distribution.

3. Possible explanations

Why the quality could vary so much for different countries remains an open question. We will investigate three possible explanations.

The first possibility is errors in the translation. The questions are in principle the same in all countries and an effort was made to make the translations as comparable as possible but, of course, there may be a difference due to errors which have been made. It is a problem in such a multi-languages project that not all languages are so well known that one can control all forms.

The second possibility is that the differences are an artefact of a difference in the execution of the MTMM experiments in the Scandinavian countries. A subgroup of the respondents did not fill out the questionnaire containing the repeated questions immediately, but sent it in later. Since this difference in the execution of the experiment exists only in the Scandinavian countries, where the quality estimates were low, we have reason to suspect these low estimates are a consequence of the time between the repetitions.

The third possibility is that the languages differ in complexity of the sentences. With complex we mean that one language uses longer sentences or longer words or more subordinate clauses than another.

In the next section the different explanations will be evaluated insofar as this can be done.

4. The empirical evaluation of the differences

In this section we will present the results of three possible explanations for the differences in data quality between the different countries. We start with the explanation based on differences in the translation in different languages.

4.1. Translations

In our analyses of the data for the different countries we found that the data of Portugal were better than the data from all other countries except Switzerland. Therefore we looked at the questionnaires of Portugal in order to see if there was anything different in these questionnaires. In doing so we found that in the supplementary questionnaire the Portuguese team did not follow the text of the source questionnaire. Where the source questionnaire uses for the questionnaire on the doctors the following formulations:

Please say how much you agree or disagree with each of these statements:
CARD*: Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree*

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
...doctors seldom keep the whole truth from their patients?	1	2	3	4	5	8
...GPs seldom treat their patients as their equals?	1	2	3	4	5	8
Before doctors decide on a treatment, they seldom discuss it with their patient	1	2	3	4	5	8

In Portugal the word "seldom" was omitted in the translation. In other countries this was not done. In this way the questions in the main questionnaire and these questions in the supplementary questionnaire became more similar in Portugal than in the other countries and this leads to higher correlations between the questions and consequently to higher estimates of the quality of the questions.

Although this is an explanation of why the quality of the questions in Portugal is relatively high for some topics, it can not be a general explanation because as far as we could judge[6], the error occurred only in Portugal. So it can not be the explanation for other large differences.

4.2 Explanation by time between repetitions

Another explanation is the difference in time between the repetitions of the same questions. The two different forms of each question were not all asked at the same time. The main questionnaire contained a standard form, while the repetitions using the alternative forms were presented to respondents in a supplementary questionnaire. In most countries this supplementary questionnaire was administered directly after the main interview of an hour. In Finland, Norway, and Sweden, however, the supplementary questionnaire was left to be completed by the respondents themselves who were asked to send it in by mail. In some cases quite some time passed before the supplementary questionnaire was received: a few respondents even waited several months before sending it in.

We can expect that if time passes between the administration of the main and supplementary questionnaire in some countries but not in others, that differences will be observed in the reliability and validity coefficients because:

1. *Changes in the traits* occur when too much time passes between the repetitions. For example, respondent's home or job situations may change in between the time they fill in the main and supplementary questionnaires. This lowers the observed correlations and leads to an *underestimate* of the quality in countries which allow respondents to return the supplementary questionnaire by mail (i.e. in Finland, Norway, and Sweden).

[6] The French questionnaire for Belgium was not available at the moment we were analyzing these data

2. *Differences in the cognitive process* can occur when the repetition is presented much later. If the repetitions follow after more that 20 minutes they people do not remember what they have said before even for identical questions (Van Meurs and Saris 1991). On the other hand when the time is much longer more ideas are processed by the respondent and the processing of a earlier question may occur in a very different cognitive setting and so the process may be different and the resulting response as well (Zaller 1991, Van der Veld 2007).

These two effects or their combination can cause differences in the quality estimates between Finland, Norway and Sweden and the other countries of the ESS.

To investigate whether this could be the case, we first compared the observed difference between these countries and the rest in average reliability and validity coefficients with a prediction of the differences from the results of previous experiments. Taking the observed effect of the maximal distance between the questions as an approximation to waiting a day or more before answering the question, an expected difference according to the meta-analysis can be obtained. This expected difference has been summarised in Table 5 together with the actually observed deviations of each country from the overall average.

Table 5: Differences in average reliability and validity coefficients between the three countries and the grand mean: observed and expected according to the SQP meta-analysis.

Country	Mean reliability coefficient	Observed difference reliability coefficients	Max. expected difference rel. coef. (SQP)	Mean validity coefficient	Observed difference validity coefficients	Max. expected difference val. coef. (SQP)
Finland	.79	.04	.042	.92	.02	.062
Norway	.78	.05	.042	.92	.02	.062
Sweden	.81	.02	.042	.91	.03	.062
All countries	.83			.94		
All w/o FNS	.84			.95		

Table 5 shows that the observed differences in average reliability coefficients between the three countries and the overall mean is very close to what can be predicted from the SQP meta-analysis. The observed differences in validity coefficients over these 18 questions in the main questionnaire are somewhat smaller than could be maximally expected from the meta-analysis.

These results show that the aggregate differences can be at least partly predicted from the meta-analysis' observation. A longer time interval between repetitions lowers the reliability and validity coefficient estimates. The possible explanation for these differences have been given above.

Next to this test on aggregate level the ESS data can be used to see whether this phenomenon can also be observed in the individual data. To this end we re-estimated the data for one of the experiments in Norway[7], this time splitting the sample into two groups. One group returned the questionnaire the same day, while the other group waited two or more days to send it back. We then fit the MTMM model to each group to observe the differences in estimated reliability and validity coefficients. Tables 6 and 7 show the results of this analysis.

[7] The experiment analysed was number 5: the effect of varying fixed reference points on answers to satisfaction items.

Table 6. A test of the model where the groups are constrained to be equal and the same model but with coefficients that may vary

Model	No. free parameters	Chi^2	Difference *df*	p-value
Two group free	55	32.37	35	0.60
Two group equal	32	53.88	23	0.0001

First the MTMM model was estimated without restrictions on the parameters across the two groups. Row 1 of the table shows that this model fits the data Next we also tested whether we could just as well fit the model with all coefficients equal across the two groups. Table 6 shows that this last hypothesis should be rejected.

Given this result we present the results of the first model without restrictions across the groups in Table 7.

Table 7: The reliability (r) and validity (v) coefficients estimated for the satisfaction experiment in Norway when the sample is split according to whether the supplementary questionnaire was filled in on the same day or not.

	Group	
Completely standardised coefficient	Same day (n=769)	Other day (n=734)
r11	0.852	0.855
r12	0.964	0.990
r13	0.866	0.846
r21	0.905	0.812
r22	0.912	0.768
r23	0.944	0.815
r31	0.762	0.777
r32	0.862	0.795
r33	0.919	0.791
v11	1*	1*
v12	0.911	0.852
v13	0.884	0.847
v21	1*	1*
v22	0.911	0.829
v23	0.904	0.801
v31	1*	1*
v32	0.918	0.848
v33	0.918	0.813

*Coefficient fixed in analysis.

Large differences between the same-day and other-day groups are observed. The largest two differences between the standardized coefficients are 0.14 and 0.13. In all cases the group which filled in the questionnaire on the same day has higher estimates for the reliability and validity coefficients[8].

[8] The only exceptions are the non-significant differences between the reliability coefficients *r11*, *r31* and *r12* in group 1 and 2. These may be due to rounding errors or sampling fluctuations and can be constrained to be equal across the groups (χ^2 = 1.42, df=3, p=0.70).

Note also that the average reliability in the "same day"-group equals 0.89, while the average across all countries found earlier was 0.83. Thus, while the questions in this experiment in Norway averaged over both groups have one of the lowest reliabilities, the group which completed the supplementary questionnaire immediately has a slightly higher than average reliability. Similarly, the average validity coefficient in this group almost equals the overall average 0.94.

This analysis is only a case study but provides strong evidence that the estimates of reliability and validity and therefore the quality of the measures in the Scandinavian countries is underestimated due to the fact that the data collections of the main questionnaire and the supplementary questionnaire were farther separated from each other than in the other countries. Due to this structural factor opinion change or differences in the cognitive processes can be the reason for the underestimation of the data quality. In the next analysis we will therefore ignore the data of the Scandinavian countries.

4.3 Explanation by language differences

As was said above one can imagine that in the translation from English to other languages differences in the structure of the sentences occur; some languages/countries use more complex sentences and longer words or sentences than others. In the meta-analysis of 87 MTMM experiments it was found that these factors played a role in explaining the differences in quality between questions (Saris and Gallhofer 2007). The results of this analysis are presented in table 8.

Table 8: The effect of the complexity of the formulation of the questions on the reliability and validity of questions when all other characters of the questions remain the same. These effects are multiplied by 1000 for legibility.

Variables	*Number of measures*	*Effect on reliability*			*Effect on validity*		
		Effect	*se*	*sign*	*effect*	*se*	*sign*
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

We selected 16 questions from the ESS questionnaire to investigate the differences in validity and reliability coefficients one might expect due to differing complexity. These questions were the standard versions of the items used in all six

MTMM experiments[9]. We also selected five different countries—Austria, Belgium, Germany, Spain, and Switzerland—where in total six different languages are spoken with which we were sufficiently familiar to be able to code the complexity of the questions. Based on the results of the meta-analysis of the MTMM experiments with respect to the complexity of the questions we have estimated the effects of the complexity of the questions in the different languages for the countries mentioned above. What became clear in the coding that the complexity does not differ so much across languages like Dutch, German, French, Italian, Spanish and Catalan. The average number of words per sentence over these 16 questions varied between 18.38 in Swiss-German and 22.25 in Catalan. The mean number of words per sentence varied between 10.77 in Austrian German and 13.35 in Catalan. Finally the average number of syllables per word varied between 1.66 in Belgian-French and 2.16 in Swiss-Italian.

Table 7 shows the total effects of the complexity of the questions in the different languages and countries.

Table 7. The effect of the complexity of questions based on the number of words, the mean number of words per sentence and the number of syllables per word on the reliability and validity estimates of SQP

	Predicted effect on reliability coefficient	**Predicted effect on validity coefficient**
Austria	-0.07	-0.03
Belgium:Dutch	-0.07	-0.04
Belgium:French	-0.06	-0.04
Switzerland:Italian	-0.08	-0.03
Switzerland:German	-0.08	-0.03
Switzerland:French	-0.05	-0.04
Germany	-0.06	-0.04
Spain:Castilian	-0.08	-0.04
Spain:Catalan	-0.07	-0.04

It will be clear that the complexity has a non-negligible effect, particularly on the reliability coefficient. But at the same time, the differences between the different languages in average predicted effects are very small. This means that this table provides little support for the idea that the considerable quality differences, as between Switzerland and Austria (.11) could be due to differing complexities across languages and translations.

However, table 7 only provides averages, and it may still be that the coefficients are predicted to be differentially influenced for some items and not for others. The question is whether the predicted differences based on language are also actually found within the same country. The multi-lingual countries offer a good opportunity for comparison, because they allow us to discount—to a certain extent—differences between countries[10]. Therefore we split the Belgian sample into a Dutch-speaking (Flemish) and a French-speaking (Walloon) group, and compared the two groups on their reliability and validity coefficients for the three questions about doctors. These

[9] By standard version we mean the version of the question asked in the main questionnaire. Two questions (B6 and B7) were left out because they are so similar in terms of complexity to the already included item B5 that no gain could be expected from coding them.

[10] Here we have to recognise that although part of the same country, French and Dutch-speaking Belgium might still differ with regard to some important characteristics; notably, the education level and age distribution. We therefore compared the two samples on these variables and indeed found some differences. However, these differences were small and reduced by adding the Brussels region to the French-speaking group. The characteristics might affect the results slightly but not so much as to invalidate the comparison.

questions were chosen because they showed relatively large differences in the predictions from the meta-study.

Table 8 presents the results of the tests of model with the parameters free over the two-group and over two-group with equal coefficients and a model with equal coefficients except for the "truth" item models .

Table 8: A comparison between models where the groups are constrained to be equal and models where the coefficients may vary for the French and Dutch Belgian questionnaire. The groups differ significantly on three coefficients.

Model	*Chi*^{2}	*df*	*p*	*SB diff*[11]	*p diff*
All loadings equal	72.89	28	<.0001	-	-
All equal except 3 loadings	37.93	25	0.047	40.63	<.0001
All free	34.55	20	0.023	3.59	0.611

Table 8 shows that complete equality must be rejected but that the model with only certain parameters free to vary is just as good as the two-group free model. The differences in coefficients across the groups under this model are shown in Table 9.

Most of the parameters could be constrained to be equal, with three exceptions (shown in bold in table 9). These exceptions, however, are not all in the direction predicted by the observation that the language of the Dutch questionnaire is slightly more complex than that of the French one for these questions, and that therefore the Flemish sample should exhibit slightly lower reliability and validity coefficients. The Walloon sample shows a lower coefficient in one out of the three cases.

The other two differences concern the coefficients for the statement "Doctors do not tell their patients the whole truth". Here the Walloon estimates are higher. However, an alternative explanation of this difference is that it is due to a difference in the translation. In French the above phrase was translated literally for the main questionnaire version[12], while in Dutch the phrase was translated as the much stronger "Doctors *conceal* the whole truth from their patients"[13]. Incidentally, this difference also exists between different German versions of the questionnaire.

At the same time, in the supplementary questionnaire item 2 was wrongly translated in the French version as 'Les médecins disent rarement toute la vérité à leurs patients' ('Doctors rarely tell their patients the whole truth'), rather than the intended 'Doctors rarely keep the whole truth from their patients'. Since the meaning in the French translation is the opposite of the intended meaning, the coefficients' sign is also the opposite from that in the Flemish sample.

It is likely that such differences in the translation are the cause of differing coefficients rather than the complexity of the language. This is corroborated by the fact that none of the other coefficients are significantly different, except for one in the direction opposite of the direction one would expect if the complexity were the reason.

Table 9: The three differing parameters. The differences are in opposite directions for the (wrongly translated) Truth item and the equal item.

	Unstandardized coefficient		Completely standardized coefficient	
	Dutch (n=1026)	French (n=748)	Dutch (n=1026)	French (n=748)
Truth-item 2	0.30	-1.48	0.70	-0.89
Truth-item 3	-0.67	-2.04	-0.99	-1.00
Equal-item 2	0.54	0.29	0.75	0.41

[11] The differences between the scaled chi-square statistics were obtained using the procedure outlined in Satorra & Bentler (2001), by the SBDIFF.exe program (Crawford, http://www.abdn.ac.uk/~psy086/dept/sbdiff.htm).

[12] "Les médecins ne disent pas toute la vérité à leurs patients."

[13] "Dokters verzwijgen de volledige waarheid voor hun patiënten"

5. Conclusions

All in all, we conclude that the idea that the differing complexity of languages causes the observed differences in reliability and validity across countries is not convincingly supported. The predictions from the meta-analysis for five different countries differ little from one another. Where the predictions do differ, some actually observed differences in the estimated reliability and validity coefficients can be found between two groups in the same country using a different language questionnaire with different complexities. However, these are not all in the expected direction and are likely due to a difference in the translation. We therefore think that the differing complexity of language does not explain the differences across countries.

For the high quality results in Portugal, we detected that they are at least partially caused by differences in the formulation of the questions in that country compared with other countries. These two findings indicate the importance of the equivalence of the formulation of the questions for the comparability of the responses.

We also found that the relatively low measurement quality of the questions in the Scandinavian countries was a consequence of a difference in the procedure in the data collection in these countries with respect to the supplementary questionnaire. Due to the delay in filling in these questionnaires the measurement quality was estimated to be lower. This is an unfortunate situation because now we do not know how good the data in the Scandinavian countries are compared with those in the other countries. In order to obtain this information the Scandinavian countries should ask the respondents to fill in supplementary questionnaire immediately after the main questionnaire as this is done in the other countries. For the present data a possible alternative would be to analyse only the group which filled in the supplementary and main questionnaires on the same day.

Given the large effects we detected of formulations of the questions on the quality estimates it would have been interesting to study as well the effect of the response categories used in the different countries. However, this requires a rather different approach which is outside the scope of this paper but will be addressed in future research.

References

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

Campbell, D. T. & Fiske, D. W. (1959). "Convergent and Discriminant Validation by the MultitraitMultimethod Matrix". *Psychological Bulletin, 56(2)*, 81-105.

Harkness J.A. (2002) Questionnaire translation. In Harkness J.A.; Van De Vijver F.J.R.; Mohler P. (Eds) (2002). *Cross-Cultural Survey Methods.* New York: Wiley, 35-57

Harkness J.A. (2007) Improving the comparability of translations. In R.Jowel, C.Roberst, R.Fitzgerald and G.Eva (eds) Measuring attitudes cross –nationally. London, Sage, 79-95

Költringer R. (1995). Measurement quality in Austrian personal interview surveys. In W.E. Saris and A. Muennich eds., The Multitrait-Multimethod Approach to evaluate measurment instruments. Budapest , Eötvös University Press, 207-225.

Oberski, D. and W.E. Saris and S. Kuipers (2005). *SQP: survey quality predictor*. Computer application programme. http://www.sqp.nl.

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.

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: John Wiley & Sons, 575-599.

Saris W.E., A.Satorra and G.Coenders (2004) A new approach for evaluating quality of measurement instruments, Sociological Methodology 2004, 311-347

Saris, W.E. and I. Gallhofer (2007): *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley.

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. Vol.25. 341-383.

Van Meurs, A. and W.E. Saris (1990). Memory effects in MTMM studies. In W.E. Saris and A. van Meurs eds., Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies, Amsterdam: North Holland, 134-146.

Van der Veld W. M. (2006) The survey response dissected: A new theory about the survey response process. PhD Thesis of the University of Amsterdam.

Zaller J.R. (1992). The Nature and Origins of Mass Opinion. Cambridge: Cambridge University Press.