Categorisation errors and differences in the quality of questions across countries

Willem E. Saris
Daniel Oberski
ESADE, Universitat Ramon Llull

Jacques Hagenaars
Tilburg University

September 12, 2007

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 et al., 2004; Saris and Andrews, 1991). 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.

Introduction

Measurement error can invalidate conclusions drawn from cross-country comparisons if the errors differ from country to country (Saris and Gallhofer, 2007). For this reason, when different groups such as countries as compared with one another, attention should not only be given to absolute levels of errors, but also to the differences between the groups. Different strategies have been developed to deal with the problem, for example within the context of invariance testing in the social sciences (Jöreskog, 1971), differential item function in psychology (Muthen and Lehman, 1985), and differential measurement error models in epidemiology and biostatistics (Carroll et al., 1995).

In the ESS a lot of time, money, and effort is spent to make the questions as functionally equivalent across countries as possible (Harkness et al., 2002) and to make the samples as comparable as possible (Häder 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 an earlier study, we investigated differences in translations, differences in the experiments' design, and differences in the complexity of the question as possible reasons for differences in question quality across countries (Oberski et al., a). Here we consider differences in categorisation errors or use of the answer scale as a source of differences between countries.

First we provide the necessary background: the response model we use as our starting point, and explanations of the multitrait-multimethod (MTMM) experiments we use to assess question quality and of categorisation error. Next we give a short overview of our limited previous findings on the origins of cross-country differences in measurement error. We then present the data from the European Social Survey and the structure of the experiments, and discuss the methods used to analyse these data. The subsequent section describes our analysis of categorisation errors for four experiments using categorical scales, and ends with a meta-analysis of these results. We then summarise and discuss our findings. Finally, some general conclusions are drawn and suggestions for further research are indicated.

1 Background

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

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

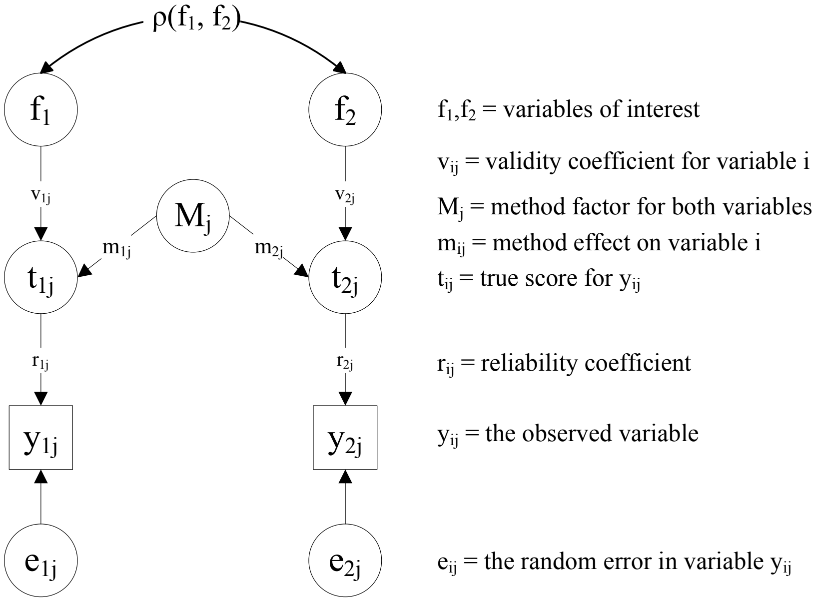

Figure 1: The response model used in the MTMM experiments.

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. This quality–sometimes also called the reliability ratio–equals $\frac{Var(f)}{Var(y)}$: it can be interpreted as the proportion of variation in the observed variable that is due to the unobserved trait of interest. 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:

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

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 (Oberski et al., b). Because we use the MTMM approach here, it is briefly introduced below. A more elaborate introduction to MTMM and SQP can be found in Saris and Gallhofer (2007).

1.1 MTMM models

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

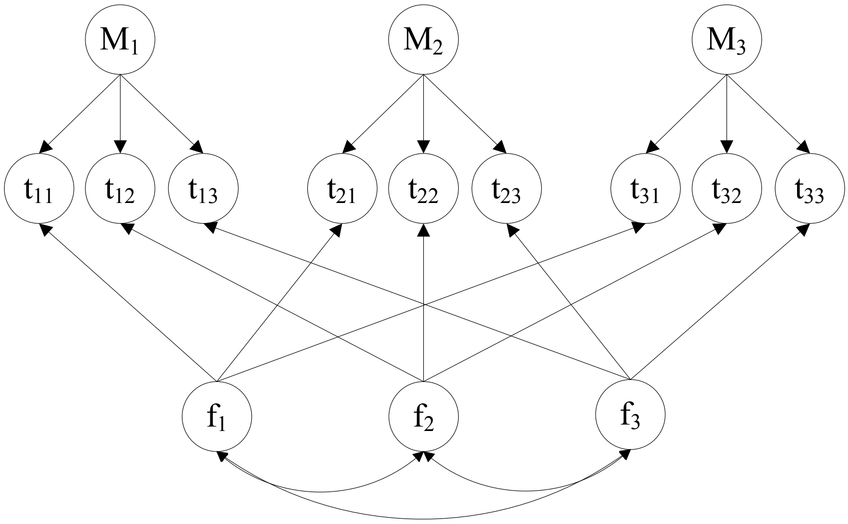

Figure 2: MTMM model illustrating the true scores and their factors of interest.

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.

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].

Table 2: Correlations between nine measures obtained from a sample of 481 people from the British population. Correlations between a repetition of a trait using a different method ('monotrait-heteromethod' correlations) are indicated in boldface.

	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
Mean	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

[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 3: The standardised estimates of the MTMM parameters (figure 2), which are obtained when analysing the sample covariance matrix above.

	Validity coeffs			Method effects			Reliability coeff.
	F1	F2	F3	M1	M2	M3	
T11	.93			.36			.79
T21		.94		.35			.85
T31			.95	.33			.81
T12	.91				.41		.91
T22		.92			.39		.94
T32			.93		.38		.93
T13	.85					.52	.82
T23		.87				.50	.87
T33			.88			.48	.84

1.2 The categorical response model

The response model discussed so far makes no mention of the fact that many of the measures we use are in fact ordinal–that is, they are ordered categories rather than a continuous scale. Broadly speaking, two types of measurement models have been proposed for this situation. The first assumes there is an unobserved categorical variable, and that errors arise from the conditional chances of choosing a category on the survey question given the unobserved score. Such models are often referred to as latent class models (Lazarsfeld and Henry, 1968; Hagenaars and McCutcheon, 2002).

The second approach deals with the case where a continuous scale or 'latent response variable' is thought to underly the observed categorical item. Such models are sometimes called latent trait models. Several possible extensions are possible, but we focus on a special case described by Muthén (1978). This is the model we will use in our subsequent analysis of the data[2].

Errors may arise at two stages. The first is the connection between the latent response variable and its latent trait. This part of the error model is completely analagous to factor analysis or MTMM models for continuous data: the scale is modelled as a linear combination of a latent trait, a reaction to the particular method used to measure the trait, and a random error, and interest then focuses on the connection between the trait and the scale, which we again term the 'quality coefficient' (see figures 1 and 2). It should be noted that, because we are interested only in the quality coefficient and not in the separation of the reliability and validity coefficients, we do not make a distinction between the true score and the latent trait as is done in the true score model explained

[2]It can be shown that analysing polychoric correlations in an MTMM model is a special case of the model we use (Muthén & Asparouhov, 2002). However, we do not use polychoric correlations because it would be necessary to assume that the variances of the latent response variables are equal across countries. Since we try to separate categorisation errors from differences in the continuous part of the model, this is not a desirable assumption. Last, it should be noted that the model we use is equivalent to a multi-dimensional two parameter graded response model in item response theory (Muthén & Asparouhov, 2002).

above. The two models are, however, mathematically equivalent (Coenders and Saris, 2000).

The second stage at which errors arise, however, differs from the continuous case. Here the continuous latent response variable is split up into the different categories, such that each category of the observed variable corresponds to a certain range on the unobserved continuous scale. The sizes of these ranges are determined by threshold parameters. Several types of errors can be distinguished at this stage (Coenders, 1996):

1. *Grouping errors* occur because the infinite possible values of the latent response variable are collapsed into a fixed number of categories. These errors will be higher when there are fewer categories;

2. *Transformation errors* occur when the distances between the numerical scores assigned to each category are not the same as the mean of the latent response variable in that category. This happens when the thresholds are not at equal distances, or when the available categories do not cover the unobserved opinions adequately.

Categorisation, then, can be expected to be another source of measurement error besides random errors and method variance. If these errors differ across countries, then so will the overall measurement quality, and differences in means, correlations, regression coefficients, and cross-tables across countries may be found which are due purely to differences in measurement errors.

The model we use allows to a certain extent the separation of errors due to the categorisation, errors due to the reaction to the method and random errors. In this paper we take advantage of this separation to compare the amount of error due to categorisation introduced across countries.

1.3 Categorisation errors in survey questions

The previous sections showed that, using the MTMM design, it is possible to obtain a measure (q^2) of the total quality of a question. If a continuous model is used, this quality is influenced by errors in both stages of the categorical response model: not only random errors and method effects are included, but also errors due to the categorisation. Consequently the linear MTMM model, assuming continuous variables, does not discount categorisation errors, but includes them to a certain extent in the estimates of the quality and other estimates (Coenders, 1996).

However, since the quality coefficient is estimated from the covariance matrix of the measures, it can be both reduced and increased by categorisation errors. In general all correlations between measures increase after correction for categorisation, but they need not all increase equally. As an illustration, consider again table 2. If categorisation errors are higher using the first method then the correlations between the latent response variables using this method (the upper-left triangle of the matrix) will increase more than the correlations of each variable with its repetition using a different method (in bold). In this case the amount of variance in the reponse variable due to the method will go up, and the estimated quality of the measure in the categorical response model can become lower than the estimated quality in the continuous MTMM model. This is because there there are method effects (correlated errors) on the level of the continuous latent response variables which do not manifest themselves in the observed (Pearson) correlations between the categorical variables. Categorisation can therefore in some cases inflate estimates of the quality of categorical observed variables, even though, at the same time, it causes errors which reduce the quality. There are thus two processes at work, which have opposite effects on the estimates of the quality.

Because it can both reduce the quality and artificially inflate estimates of quality, categorisation error is an important possible source of differences in question quality across countries. If either of these two processes differs from one country to another, the estimates of the quality can be expected to differ as well. Therefore, we focus here on the changes in the quality coefficient one observes in a continuous reponse model relative to a categorical model. This allows us to study the attentuation or inflation of quality estimates across countries.

As noted before, the quality (or reliability ratio) of a variable is defined as the ratio of the true trait variance to the observed variance (see also figure 1 in the first section):

$$q^2 = \frac{Var(f)}{Var(y)}. \tag{2}$$

However, we have now seen that y is itself a categorization of an unobserved continuous variable (c), and therefore the above equation 2 can be decomposed into

$$q^2 = \frac{Var(f)}{Var(c)} \cdot \frac{Var(c)}{Var(y)}. \tag{3}$$

The scale of c is arbitrary, except that it may vary across countries due to relative differences in variance (Muthén & Asparouhov, 2002). However, the ratio $Var(c)/Var(y)$ can easily be calculated once q^2, the quality from the continuous analysi, and $Var(f)/Var(c)$, the quality from the categorical MTMM analysis, have been obtained. So the effect of categorisation on the quality estimate can be derived.

In the present study, we estimate this 'categorisation factor' for different countries and experiments, and examine to what extent it can explain the differences in quality across countries. First, however, we give an overview of the data we use.

2 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 and 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 the 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 (Häder and Lynn 2007).

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.

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 other 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

[3]http://www.europeansocialsurvey.org

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

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

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

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.

3 Explanations for cross-country differences in question quality

The previous section showed that in some cases large differences were found in question quality across the countries of the ESS. In a previous study, we examined different explanations of these discrepancies.

4 Methods

5 Results

[this part needs to be re-written...]

We first turn to the hypothesis that all thresholds are equal across different countries. We selected the three countries with the highest and the two countries with the lowest quality coefficients for the 'women' experiment. In this experiment wording of the question reversed in the second method. For example, the statement 'When jobs are scarce, men should have more right to a job than women' from the main questionnaire was changed to 'When jobs are scarce, women should have the same right to a job as men' in the supplementary questionnaire. The countries with high quality coefficients were, in this case, Portugal, Greece, and Switzerland. The lowest coefficients for this experiment were found in Estonia and the Czech Republic. To be able to separately study misspecifications in the categorization part of the model, we imposed no restrictions on the covariance matrix of the latent response variables at this stage. The estimation was done using the weighted least squares approach described by Flora & Curran (2004)*.

In the first analysis, all thresholds were constrained to be equal across the five countries. This yields a likelihood ratio statistic of 507 on 48 degrees of freedom. The country with the highest (128) contribution to this chi-square statistic is Portugal. When we examine the expected parameter changes, it also turns out that in this country these standardised values are very large with some values close to .9 while in other countries the highest obtained and exceptional value is .6. At the same time, there is nothing about this group that makes the power particularly high. There are a few very high power values of .97 for large expected changes, but the average power is .37. For some reason, the equality constraint on the Portuguese thresholds appears to be a particularly gross misspecification.

This misspecification is very likely due to a translation error. The intention of the experiment was to reverse the wording of the question in the second method. But in Portugal the reverse wording was not used, and the same version was presented as in the main questionnaire. The supplementary questions mean something else in Portugal and therefore it makes sense that the thresholds should differ from those found in the other countries. To prevent incomparability when the MTMM model is estimated, we omit Portugal from our further analyses and continue with four countries. We should note, however, that the resulting experiment does provide a good opportunity to examine the effect of a repetition on the threshold structure.

Continuing our analysis with four countries, the model where all thresholds are constrained to be equal yields a likelihood ratio of 351 and 36 degrees of freedom. We approximate the power of the score test (modification index) to detect a standardised difference of at least .1, and examine the modification indices and expected parameter changes.

Using the above criteria, it is clear that the equality model contains severe misspecifications. We formulated a new model in which some thresholds were constrained to be equal, while others were freed to vary. However, we found that too many parameters were freed for Greece and some parameters could again be constrained to be equal: the expected parameter change which led us to free these parameters is an estimate, which need not equal the actual parameter change. For some thresholds the actual change was negligible (although still within a 95% confidence interval for the expected change) and we constrained these four parameters again. The resulting model has an approximate likelihood ratio of 2.8 on 2 degrees of freedom ($p = .24$)3.

Table 5:

		CutDown	'Women' Respnsib.	MenRight
Continuous analysis				
q^2	Greece	0.71	0.66	0.71
	Slovenia	0.54	0.25	0.68
Categorical analysis				
q^2	Greece	0.51	0.35	0.48
	Slovenia	0.69	0.29	0.65
Categorisation factor				
	Greece	1.4	1.9	1.5
	Slovenia	0.8	0.9	1.0

5.1 Experiment 1

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

	1 *Agree strongly*	τ_1	2 *Agree*	τ_2	3 *Neither agree nor disagree*	τ_3	4 *Disagree*	τ_4	5 *Disagree strongly*
Slovenia		-1.4		-0.1		0.6		1.8	
Greece		-1.1		-0.2		0.5		1.4	

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

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.5		-0.0		0.6		2.0	
Greece		-1.5		-0.3		0.4		1.5	

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

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-0.5		1.3		1.9		2.6	
Greece		-0.6		0.7		1.6		2.3	

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

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.7		-0.7		-0.2		1.2	
Greece		-1.6		-0.5		0.0		1.4	

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

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-1.8		-0.8		-0.3		0.9	
Greece		-0.9		0.1		0.6		1.4	

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

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Slovenia		-0.8		0.7		1.1		1.9	
Greece		-1.1		-0.1		0.7		2.0	

Table 6:

		'Job'		
		Varied	Secure	Risky
Continuous analysis				
q^2	Belgium	0.88	0.88	0.92
	Slovenia	0.61	0.21	0.55
Categorical analysis				
q^2	Belgium	0.92	0.58	0.66
	Slovenia	0.50	0.29	0.67
Categorisation factor				
	Belgium	0.96	1.52	1.39
	Slovenia	1.22	0.72	0.82

Table 7:

		'Doctors'		
		Truth	Equals	Discuss
Continuous analysis				
q^2	Denmark	0.05	0.74	0.77
	Estonia	0.42	0.85	0.83
Categorical analysis				
q^2	Denmark	0.12	0.83	0.82
	Estonia	0.75	0.79	0.75
Categorisation factor				
	Denmark	0.41	0.89	0.94
	Estonia	0.56	1.07	1.10

Table 8:

		'Efficacy'		
		Complex	Active	Mind
Continuous analysis				
q^2	Switzerland	0.49	0.81	0.50
	Denmark	0.77	0.83	0.79
Categorical analysis				
q^2	Switzerland	0.62	0.94	0.62
	Denmark	0.63	0.70	0.63
Categorisation factor				
	Switzerland	0.79	0.86	0.81
	Denmark	1.23	1.18	1.25

Table 9: A meta-analysis of the categorisation error studies. The table presents a linear regression of the categorisation factor on topic, country, scale and scale position.

	Estimate	S.E.	lower	upper
(Intercept)	-0.04	0.22	-0.49	0.40
Topic				
Doctors	(reference category)			
Efficacy	0.14	0.12	-0.11	0.39
Job	0.20	0.19	-0.18	0.57
Women	0.45	0.17	0.11	0.79
Country				
Belgium	(reference category)			
Slovenia	0.09	0.10	-0.11	0.29
Czech Republic	0.28	0.15	-0.02	0.59
Switzerland	0.31	0.15	0.01	0.61
Greece	0.32	0.15	0.01	0.63
Denmark	0.33	0.17	-0.02	0.67
Estonia	0.53	0.19	0.15	0.92
Scale				
Agree-disagree	(reference category)			
Direct	0.12	0.09	-0.07	0.30
True-false	0.42	0.14	0.13	0.70
Scale position				
Negative	(reference category)			
'Usually'	0.39	0.15	0.09	0.69
'Rarely'	0.41	0.15	0.11	0.71
Positive	0.48	0.07	0.34	0.63

Multiple R-Squared: 0.55; Adjusted R-squared: 0.44; F-test for factor sums of squares:

Topic: $p = .16$; Country: $p = .17$; Scale: $p = .02$; Position: $p < .00001$.

5.2 Experiment 2

5.3 Experiment 3

5.4 Experiment 4

6 Discussion and conclusion

interpret meta analysis

advice based on this study

future research: o.a. specific causes for differing quality such as education, age, or interviewer effects?

References

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *The Public Opinion Quarterly*, 48:409–442.

Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*, 56:81–105.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). Nonlinear measurement error models. *Monographs on Statistics and Applied Probability.(Chapman and Hall, New York) Volume*, 63.

Coenders, G. (1996). *Structural Equation Modeling of Ordinally Measured Survey Data*. PhD thesis, Universitat Ramon Llull.

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

Häder, S. and Lynn, P. (2007). How representative can a multi-nation survey be? In Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G., editors, *Measuring Attitudes Cross-nationally: Lessons from the European Social Survey*. SAGE.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Harkness, J. A., van de Vijver, F. J. R., and Mohler, P. P. (2002). *Cross-cultural survey methods*. Wiley-Interscience.

Jöreskog (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36:409–426.

Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G. (2007). *Measuring Attitudes Cross-nationally: Lessons from the European Social Survey*. SAGE.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.

Muthen, B. and Lehman, J. (1985). Multiple group irt modeling: Applications to item bias analysis. *Journal of Educational Statistics*, 10:133–142.

Oberski, D., Saris, W. E., and Hagenaars, J. *Differences in the quality of questions across countries*.

Oberski, D., Saris, W. E., and Kuipers, S. Sqp: survey quality predictor.

Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. *Measurement errors in surveys*, page 575?599.

Saris, W. E. and Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley Series in Survey Methodology. John Wiley & Sons.

Saris, W. E., Satorra, A., and Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot mtmm design. *Sociological Methodology*, 34:311–347.