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 15, 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, 2007a). 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 (Muthén 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, 2007a). 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., frth). 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, 2007a) 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
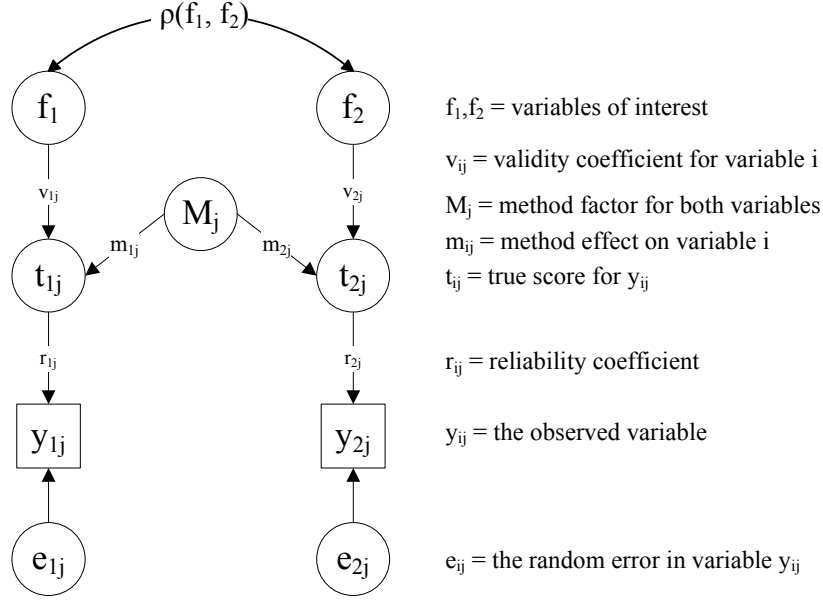
Figure 1: The response model used in the MTMM experiments.

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:

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

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., 2004). 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
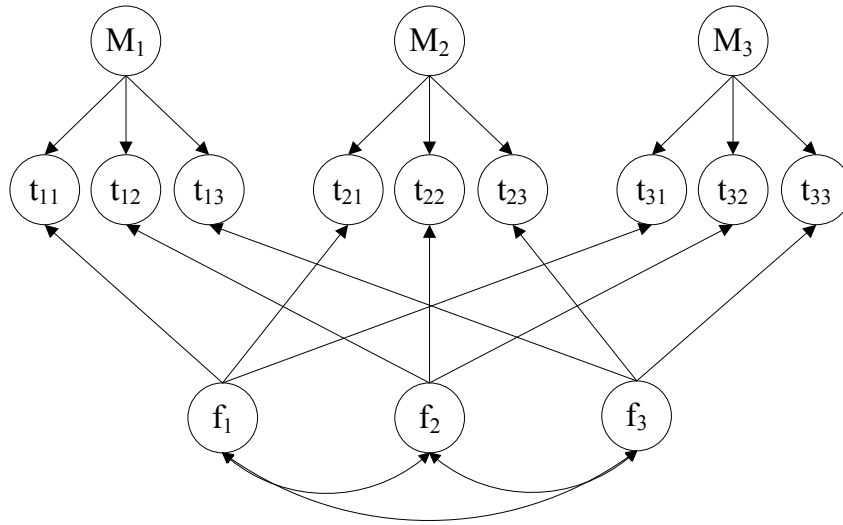

Figure 2: MTMM model illustrating the true scores and their factors of interest.

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 measured on 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 (1984). This is the model we will use in our subsequent analysis of t2 he 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 above. The true score and classic MTMM models are, however, equivalent (Coenders

---

[2]It can be shown that analysing polychoric correlations in an MTMM model is a special case of the model we use (Muthén and 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. The model we use is equivalent to a multi-dimensional two parameter graded response model in item response theory (Muthén and Asparouhov, 2002).

and Saris, 2000).

The second stage at which errors arise 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 distances between the means of the latent response variable in those categories. This happens when the thresholds are not equally spaced, 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 result which are due purely to differences in measurement errors.

The model we use allows to a certain extent for 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 parameters (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. For example, 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 relative to the observed correlations 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 be larger in the categorical model than in the continuous model, 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 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 and artificially inflate estimates of quality, categorisation error is an important possible cause of finding 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 $q^2_{cont.}$, the quality from the continuous analyis, and $Var(f)/Var(c)$, the quality from the categorical MTMM analysis ($q^2_{cat.}$), have been obtained. So the effect of categorisation on the quality estimate can be derived. We attach no specific significance to this number other than that it is a useful index of the relative differences between the quality estimates of the continuous and categorical models.

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 completed 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 et al. (2002).

Despite these efforts to make the data as comparable as possible, large 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 ($F = 3.19$, $df = 16$, $p < .001$)

---

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

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

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

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.

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 a few possible explanations of these discrepancies (Oberski et al., frth).

The first explanation we studied were errors in the translation. Although in the ESS a lot of care has been taken to ensure the correct translation of the questions, we found that a few questions in the supplementary questionnaire had not been translated as intended. In particular, one item in the 'doctors' experiment had been translated in all French questionnaires as 'Doctors rarely tell their patients the whole truth' rather than 'Doctors rarely keep the whole truth from their patients'. Since these sentences have opposite meanings, it is unsurprising that we should find a different relationship with the trait of

---

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

interest.

Another alternative explanation for differences across countries is differences in the implementation of the experimental design. Here one difference existed between the implementations in Norway, Sweden, and Finland and the other countries: in these countries respondents could send in the supplementary questionnaire containing the repetitions at a time chosen by themselves, while the general design used in other countries was that the supplementary questionnaire was administered directly after the main interview. Since in some cases quite some time passed before the second questionnaire was completed, it is possible that the opinion has in this time changed or that other unique considerations are taken into account at the second moment than at the first moment. An MTMM analysis of a sample split according to whether the questionnaire was returned within two days or later provided strong evidence that this was indeed the case. In fact, the sample which returned the questionnaire on the same day by itself was very similar with respect to the quality to other countries.

The third alternative we considered was that the language of the questions could be more complex in one language than in another. Previous meta-analyses found that language complexity can have an effect on the quality (Saris and Gallhofer, 2007b). However, we found no strong evidence that the complexity of the questions caused differences in question quality in this case.

Thus, in some cases we found artificial differences in quality which are likely to be due to an erroneous translation or different implementation of the experimental design–notably in the Scandivian countries except Denmark and for one item in the French-speaking countries. However, these cases are not so numerous that they can explain the large overall variations in question quality that were found in the ESS. Therefore we now turn to the possibility that the use of the categories in the categorical questions differs from country to country, and proceed to investigate the influence of categorisation errors on the quality in different countries and experiments.

4   Methods

In almost every country of the ESS, respondents were asked to complete a supplementary questionnaire containing the repetitions used in the experiments. Not all respondents completed the same questionnaire. The sample was randomly divided into subgroups, so that half of the people answered the first and second form of the questions, and the other half answered the first and third form.

This split-ballot MTMM approach lightens the response burden by presenting fewer questions and fewer repetitions. Saris et al. (2004) showed that the different parameters of the MTMM model can still be estimated using this planned missing data design. If the covariances between the questions belonging to the first and second, between those from the first and third methods and the covariances between questions using the same method are identified, then so are the different parameters of the MTMM model. In the categorical MTMM model, these covariances are indeed also identified, and therefore (Millsap and Yun-Tein, 2004, appendix A) the entire model is as well.

For each experiment, two different models were estimated. The continuous analysis was conducted using the covariance matrices as input, and estimated using the maxi-

mum likelihood estimator in LISREL 8. The results presented in the tables below were standardised after the estimation.

The categorical model can in principle also be estimated using maximum likelihood. However, in order to deal with the planned missing data (split-ballot) a procedure such as full-information maximum likelihood would be necessary. This requires numerical integration in the software we used (Mplus 4), making the procedure prohibitively slow and imprecise. We therefore developed an alternative two step approach, whereby in the first step the covariance matrices of the scales were estimated, and in the second step the MTMM model is fitted to the estimated matrices. The estimation in the first step was done using the weighted least squares approach described by Flora and Curran (2004), and the second step again employed the maximum likelihood estimator.

This approach has the advantage that consistent and numerically precise estimates can be obtained within seconds rather than days (Muthén and Asparouhov, 2002). The disadvantages are that the standard errors of the estimates of the categorical MTMM model are incorrect, and that the chi-square statistic and modification indices may be inflated. Although the problem could in principle be remedied by using the asymptotic covariance matrix of the covariances as weights in the estimation (Jöreskog, 1990), in the present paper we compare only the consistent point estimates of this model.

We note here that the categorical MTMM model is equivalent to the 'graded response model' in item response theory. There is a simple relationship between the threshold and quality coefficients of our model and difficulty and discrimination parameters in IRT models: the quality coefficients are scaled discrimination parameters, while a scaled difficulty for each category can be obtained by dividing each threshold by the corresponding quality coefficient (Muthén and Asparouhov, 2002).

The two models are the same with respect to the covariance structure of the response variables (the 'MTMM part' of the model). However, while in the continuous model it is necessary to assume that the variables are not categorical, in the categorical model it is necessary to assume that the unobserved response variables follow a multivariate normal distribution. While in principle the continuous model makes the same assumption, it has been shown that the continuous model is in many cases highly robust to violations of the normality assumption (Satorra, 1990). At the same time, the categorical model is known to be more sensitive to violations of this assumption, particularly when the variables are skewed in opposite directions (Coenders, 1996). Thus, it depends on the structure of the data whether one model or the other provides a more adequate estimate of the quality of the questions. This should be kept in mind in the interpretations of the results.

5 Results

We estimated the quality of the measures based on the continuous model and based on the categorical model for four experiments which used an answer scale of five categories or less in the main questionnaire. For each of these questions we took the ratio of the two different quality measures as an index of the effect that categorisation has on the continuous quality estimates. The first experiment's results will be described in some detail, while we provide the results of the other experiments more briefly.

5.1 Experiment 1

The first experiment concerned opinions on the role of women in society (see table 5). We first turn to the hypothesis that all thresholds are equal across different countries. If this hypothesis cannot be rejected then there is also little reason to think that the categorisation is causing differences in the quality coefficients.

   We selected the two countries with the highest and the country with the lowest quality coefficients. In this experiment, the 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 and Greece. The lowest coefficients for this experiment were found in Estonia. To be able to separately study misspecifications in the categorisation part of the model, we imposed no restrictions on the covariance matrix of the latent response variables at this stage.

Table 5: The 'women' experiment: questions and threshold estimates (in z-scores).

'A woman should be prepared 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
	Agree strongly			*Agree*		*Neither disagree nor agree*		*Disagree*	*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	

   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

Table 6: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorisation factors for the experiment on opinions about the role of men and women in society.

		CutDown	'Women' Respnsib.	MenRight
Continuous analysis				
q^2	Greece	0.71	0.66	0.71
	Slovenia	0.54	0.25	0.68
m	Greece	0.15	0.15	0.15
	Slovenia	0.17	0.24	0.15
Categorical analysis				
q^2	Greece	0.51	0.35	0.48
	Slovenia	0.69	0.29	0.65
m	Greece	0.49	0.14	0.32
	Slovenia	0.33	0.75	0.19
Categorisation factor				
	Greece	1.4	1.9	1.5
	Slovenia	0.8	0.9	1.0

standardised values are very large with some values close to .9 while in other countries the highest obtained and exceptional value is .6. For some reason, the equality constraint on the Portuguese thresholds appears to be a particularly gross misspecification.

As it turns out, this particular 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. 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. In addition, our analysis of the measurement model has revealed an item that was incorrectly translated in the supplementary questionnaire but had (as far as we know) gone undetected so far.

Continuing our analysis with two countries, the model where all thresholds are constrained to be equal yields a likelihood ratio of 351 and 36 degrees of freedom ($p < .00001$). This model should therefore be rejected: the thresholds are significantly different across countries.

We formulated a new model in which some thresholds were constrained to be equal, while others were freed to vary. The resulting model has an approximate likelihood ratio of 2.8 on 2 degrees of freedom ($p = .24$)[6]. The resulting estimates of the threshold parameters are presented in table 5. These estimates have been expressed as z-scores in order to make them comparable.

Table 5 presents three different traits, each asked in two different forms. The first form of each trait is the form asked in the main questionnaire, while the second form was asked in the supplementary questionnaire (the third form has been omitted for brevity).

Looking at the first question, it can be seen that the endpoints are somewhat distant, especially in Slovenia: there the category 'disagree strongly' is 1.8 standard deviations

---

[6]It is also possible to free more parameters and put no restrictions at all on the model. This might lead us to find differences between countries more easily, since the parameters are allowed to vary. However, we prefer to aid our estimation by imposing these restrictions: if they do not hold in the population, this leads us to be conservative in ascribing differences between countries to the categorisation.

or more away from the mean, reducing the number of scale points that are available for some people. The remaining categories, however, are reasonably similar, not too far away from the distribution of the opinion, and not very unequally spaced.

The second form of the same question is similar to the first form in this respect, except that here both of the endpoints are rather distant in both countries, again reducing the number of scale points. As noted above, a reduction in scale points can be expected to increase grouping errors.

The second trait ('responsibility') presents a radically different picture. In both countries the 'disagree' and 'disagree strongly' categories are quite far away from the mean. This again reduces the number scale points, while, at the same time, the scale is cut off in this manner only from one side. Large transformation errors can be expected. Moreover, in Slovenia this effect is much worse than in Greece: the category 'neither disagree nor agree' is already 1.3 standard deviations or more away from the mean, reducing the amount of information provided by this variable in Slovenia even further.

The second phrasing of this question seems to provide a better coverage of the prevailing opinions on women and men's responsibility for the home and children.

For the third and last trait–the right to a job–the most striking feature of the thresholds is that in Slovenia, the first three categories represent opinions below the mean, while in Greece only the first category does. Beyond this, it is difficult to say which scale might produce fewer categorisation errors. Surprising, however, is that the second form of the same question seems to produce much more comparable scales with respect to the thresholds than the first one.

The thresholds provide some insight into the nature of categorisation errors. However, the final quality of the measure depends also on other parameters of the categorical response model such as the quality coefficients and the error variances. Since the quality is a complex of different parameters, the discrepancy between the quality obtained in the categorical model and the quality in the continuous model is more informative on the effects of categorisation. As explained above, we operationalise this discrepancy by taking the ratio $q^2_{continuous}/q^2_{categorical}$. Table 6 contains the quality estimates and standardised method effects under both models together with this categorisation factor for the three questions in the main questionnaire.

The top two rows of table 6 show that the quality in Greece was higher than in Slovenia using the continuous model; this is, indeed, the reason we chose these particular countries to compare. The quality in Slovenia is lower for the first question, dramatically lower for the second question, and very similar for the third question. This is in principle in line with the descriptions given above of our expectations of categorisation errors.

However, table 6 also shows that such interpretations of the possible influence of the thresholds are not as straightforward as they might seem. We fitted the MTMM model to the estimated covariance matrix of the latent response variables, and obtained a model which seemed to fit reasonably well ($\chi^2 = 20$, $df = 10$, $p = .02$). While for the first and second questions the low qualities are indeed corrected upwards somewhat after the categorisation has been taken into account, the opposite happens in Greece. In that country all of the quality coefficients are lower using the categorical analysis than they are in the continuous analysis.

A consequence of this is that, using the continuous model, a higher quality is obtained in Greece than in Slovenia, while the reverse is true in the categorical model for the first

and last items. This is rather striking given that, taken over all questions in the main questionnaire, Greece had a substantially higher quality estimate than Slovenia (see table 4).

Having indicated our procedure and points of interest in the results, we now discuss the remaining three experiments in less detail; they have been analysed in the same manner.

5.2   Experiment 2

The second experiment concerned some aspects of the respondent's job. Table 7 again shows the threshold estimates after acceptable models for the threshold structure had been found while leaving the covariance matrices free to vary (the test statistics for the two split-ballot groups were 2.4 with 4 df, $p = .66$, and 2.8 with 2 df, $p = .24$, respectively.).

Table 7: The 'job' experiment: questions and threshold estimates (in z-scores).

'There is a lot of variety in my work.'

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all true	*A little true*	*Quite true*	*Very true*
Belgium	-1.4	-0.7	0.2	
Slovenia	-1.7	-0.8	0.4	

'Please choose one of the following to describe how varied your work is.'

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all varied	*A little varied*	*Quite varied*	*Very varied*
Belgium	-1.5	-0.7	0.3	
Slovenia	-1.7	-0.6	1.0	

'My job is secure.'

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all true	*A little true*	*Quite true*	*Very true*
Belgium	-1.2	-0.7	0.1	
Slovenia	-1.5	-0.6	0.7	

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

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all secure	*A little secure*	*Quite secure*	*Very secure*
Belgium	-1.6	-0.8	0.3	
Slovenia	-1.6	-0.4	1.3	

'My health or safety is at risk because of my work.'

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all true	*A little true*	*Quite true*	*Very true*
Belgium	0.0	0.9	1.3	
Slovenia	-0.3	0.6	1.6	

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

	1 τ_1	2 τ_2	3 τ_3	4
	Not at all at risk	*A little at risk*	*Quite a lot at risk*	*Very much at risk*
Belgium	-0.1	0.9	1.8	
Slovenia	-0.5	0.8	1.9	

It can be seen that in all cases, the endpoints for Slovenia are further away than the same endpoints for Belgium. Transformation errors are also present, as the thresholds

are sometimes one-sided or not equidistant.

Taking the estimated covariances of the latent response variables as our input, we estimated the quality in order to compare the results of this analysis with the results from the continuous analysis. This model resulted in a test statistic of 40 on 23 degrees of freedom ($p = .01$). The different estimates, including the method effects, for the main questionnaire items are presented in table 8.

Table 8: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorisation factors for the experiment on certain aspects of the job of the respondent.

		'Job'		
		Varied	Secure	Risky
Continuous analysis				
q^2	Belgium	0.88	0.88	0.92
	Slovenia	0.61	0.21	0.55
m	Belgium	0.06	0.06	0.06
	Slovenia	0.17	0.29	0.18
Categorical analysis				
q^2	Belgium	0.92	0.58	0.66
	Slovenia	0.50	0.29	0.67
m	Belgium	0.19	0.56	0.51
	Slovenia	0.00	0.00	0.00
Categorisation factor				
	Belgium	0.96	1.52	1.39
	Slovenia	1.22	0.72	0.82

The low quality of .21 for the second item in Slovenia catches the eye in table 8. Interestingly, there is nothing in the thresholds to suggest where such a low quality should come from. The change in this coefficient when using the categorical model is relatively small, going from .21 to .29, suggesting the low quality is not so much tied to the thresholds as to other parameters of the model.

For the second and third question, the country with the higher quality estimates in the continuous analysis again finds its estimates in the categorical model corrected substantially downwards. Belgium retains its position as the country with higher quality measurements according to this model, however.

5.3   Experiment 3

The third experiment concerned the respondent's perception of the behaviour of doctors in general. Table 9 displays our estimates of the threshold parameters for the different questions.

Table 9 shows that the thresholds appear to be very similar across the two countries, except for threshold τ_3 for the very first question[7]. Considerably more information is

---

[7]The test statistics for the first and second split-ballot groups were 6.3 with 7 df, $p = .50$, and 4.5 with 3 df, $p = .22$, respectively.

Table 9: The 'doctors' experiment: questions and threshold estimates (in z-scores).

'Doctors keep the whole truth from their patients.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Never or almost never		Some of the time		About half of the time		Most of the time		Always or almost always
Denmark		-0.6		0.6		0.8		1.4	
Estonia		-0.5		1.0		1.4		2.1	

'Doctors rarely keep the whole truth from their patients'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Agree strongly		Agree		Neither disagree nor agree		Disagree		Disagree strongly
Denmark		-1.4		0.0		0.6		1.8	
Estonia		-1.4		0.0		0.6		1.8	

'GPs treat their patients as their equals.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Denmark		-1.8		-0.9		-0.7		0.3	
Estonia		-1.3		-0.6		-0.2		0.8	

'GPs rarely treat their patients as their equals'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Denmark		-2.3		-0.9		-0.4		1.2	
Estonia		-1.9		-0.6		-0.1		1.4	

'Before doctors decide on a treatment, they discuss it with their patient.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Denmark		-1.6		-0.7		-0.4		0.5	
Estonia		-1.3		-0.5		-0.1		0.9	

'Before doctors decide on a treatment, they rarely discuss it with their patient'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
Denmark		-1.9		-0.7		-0.3		1.4	
Estonia		-1.8		-0.6		-0.1		1.5	

provided by this question in Denmark than in Estonia, since in the latter country only the first two categories are less than one standard deviation away from the mean. If we examine the distribution of this variable, this corresponds to a strong difference in skew: in Estonia 85% of the observations fall in the first two categories, while in Denmark this percentage is a more modest, though still skewed, 72%. Whether this is due to the thresholds themselves or to a skew in the underlying distribution, we can not say.

In addition, there is a difference between the countries for threshold τ_1 for the third question, 'GPs treat their patients as equals'. In Denmark this extreme cutting point is more distant than in Estonia.

Not everything about categorisation error can directly deduced by looking only at the thresholds. The categorical model produced results which might seem surprising if this is not taken into account. The likelihood ratio in this case was 26 with 21 degrees of freedom ($p = .19$). Table 10 shows the results we obtained for this experiment.

An extremely low quality for the first question was obtained in Denmark; this variable had very low correlations with its repetitions–and every other variable in the model. These low correlations were corrected upwards somewhat after taking the categorisation into account, but they remained low, producing an also low quality of .05 in the continuous

Table 10: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorisation factors for the experiment on perception of the behaviour of doctors.

| | | 'Doctors' | | |
		Truth	Equals	Discuss
Continuous analysis				
q^2	Denmark	0.05	0.74	0.77
	Estonia	0.42	0.85	0.83
m	Denmark	0.00	0.00	0.00
	Estonia	0.00	0.00	0.00
Categorical analysis				
q^2	Denmark	0.12	0.83	0.82
	Estonia	0.75	0.79	0.75
m	Denmark	0.49	0.17	0.17
	Estonia	0.48	0.37	0.42
Categorisation factor				
	Denmark	0.41	0.89	0.94
	Estonia	0.56	1.07	1.10

and .12 in the categorical analysis[8]. As suggested before, an explanation for this is that the random errors are very large for this question.

In Estonia the quality using the continuous model is also estimated to be rather low. However, the categorical model estimates the quality of the underlying scale at a level similar to that of the other two questions.

For the other two questions, it is worth noting that while the estimates are slightly lower in Denmark in the continuous model, these differences are reduced in the categorical model.

5.4 Experiment 4

The topic of the last experiment was the subjective feeling of compentence to participate in politics, also called 'efficacy'. This was the only experiment we analysed from the first round of the ESS. The results of the threshold model[9] are presented in table 11.

For the first two questions belonging to the first trait, no relevant differences were found between the thresholds. The third questions exhibits differences in the sense that the skew of the thresholds in Switzerland and Denmark are in different directions. At the same time, the skew of the fourth question in Switzerland is in the opposite direction as for the third question. This is what might be expected given the answer scales. However, one consequence is that the continuous analysis may contain two variables that are skewed in opposite directions. It has been shown that this situation in particular causes differences between categorical and continuous analyses (Johnson and Creech, 1983).

The same does not happen in Denmark though; here the fourth question exhibits

---

[8]The data were subjected to a search for mistakes, but no coding or other such errors became apparent.

[9]The test statistics for the first and second split-ballot groups for this eperiment were 3.3 with 6 *df* ($p = .78$) and 5.0 with 6 *df* ($p = .54$), respectively.

Table 11: The 'efficacy' experiment: questions and threshold estimates (in z-scores).

'How often does politics seem so complicated that you cant really understand what is going on?'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Never		Seldom		Occasion-ally		Regularly		Frequently
Switzerland		-1.4		-0.5		0.6		1.3	
Denmark		-1.3		-0.5		0.6		1.3	

'Sometimes politics seems so complicated that I cant really understand what is going on.'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Agree strongly		Agree		Neither disagree nor agree		Disagree		Disagree strongly
Switzerland		-1.5		-0.1		0.4		1.5	
Denmark		-1.4		-0.1		0.4		1.4	

'Do you think that you could take an active role in a group involved with political issues?'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Definitely not		Probably not		Not sure either way		Probably		Definitely
Switzerland		-0.4		0.3		0.7		1.3	
Denmark		-1.0		-0.6		-0.3		0.4	

'I think I could take an active role in a group involved with political issues'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Agree strongly		Agree		Neith.d.n.a.		Disagree		Disagree strongly
Switzerland		-1.8		-0.6		-0.3		0.6	
Denmark		-1.2		-0.0		0.2		1.2	

'How difficult or easy do you find it to make your mind up about political issues?'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Very difficult		Difficult		Neither difficult nor easy		Easy		Very easy
Switzerland		-1.5		-0.6		0.3		1.5	
Denmark		-1.8		-0.9		-0.2		1.0	

'I find it easy to make my mind up about political issues'

	1	τ_1	2	τ_2	3	τ_3	4	τ_4	5
	Agree strongly		Agree		Neith.d.n.a.		Disagree		Disagree strongly
Switzerland		-1.2		0.0		0.7		1.9	
Denmark		-1.2		0.1		0.6		1.7	

more or less symmetrical thresholds. This is somewhat surprising given the thresholds of the same question using a direct question rather than an agree-disagree scale; if the scales were equally good but simply a reverse way of phrasing then one would expect also a reversal of the thresholds. The fact that this does not occur suggests that other factors play a part besides the categorisation errors, such as the quality coefficients, random errors and method variance.

The comparison between the categorical and continuous models for this experiment is shown in table 12.

A seemingly puzzling phenomenon in table 12 is that the estimates for the first question should be so different in the categorical model[10] after we have seen in table 11 that for this question the differences between the thresholds are minimal. This stresses again the fact the thresholds alone cannot provide all the necessary information. It can be seen,

---

[10]The chi-square test statistic of our final categorical model was 28 with 2 degrees of freedom ($p = .10$).

Table 12: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorisation factors for the experiment on the respondent's subjective political competence (political efficacy).

		'Efficacy'		
		Complex	Active	Mind
Continuous analysis				
q^2	Switzerland	0.49	0.81	0.50
	Denmark	0.77	0.83	0.79
m	Switzerland	0.00	0.00	0.00
	Denmark	0.00	0.00	0.00
Categorical analysis				
q^2	Switzerland	0.62	0.94	0.62
	Denmark	0.63	0.70	0.63
m	Switzerland	0.00	0.00	0.00
	Denmark	0.11	0.08	0.11
Categorisation factor				
	Switzerland	0.79	0.86	0.81
	Denmark	1.23	1.18	1.25

however, that the categorisation factors and the differences between them across the two countries are very similar for all three questions.

Now that we have presented and discussed the results of the different experiments, the question remains whether there are differences between the countries with regard to the categorisation factors, and what the influence on this factor of the topic and the scale used is. The next section therefore presents the results of a meta-analysis we conducted on the categorisation factors.

5.5 A meta-analysis of the results

Using the results presented in the previous sections, we constructed a data set consisting of the categorisation factor for all questions–including those from the supplementary questionnaire not shown above–in the four different experiments for which this index was available. We also added data on questions from Switzerland and the Czech Republic for the 'women' experiment. This yielded 78 cases in total. We also recorded the country, topic of the experiment, the type of scale used (agree-disagree, true-false or direct question) and whether the question was phrased as a negative statement or not. The categorisation factors analysed are presented as a histogram in figure 3. Figure 4 shows a plot of the quality obtained in the continuous analysis and the categorisation factor.

The categorisation factor is of course calculated using $q^2_{cont.}$, but a relationship is only to be expected when the qualities under the categorical and continuous models are not constantly proportional to each other. Figure 4 shows that with one exception, categorisation factors above one occur only when the quality is relatively high. It can also be seen that the absolute level of the quality and the discrepancies between the quality under the two models are related. If the sample is split according to whether the quality was 'high' or 'low' (above or below the median 0.56), the means of the two groups are
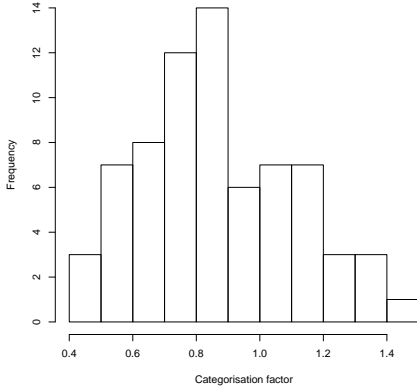
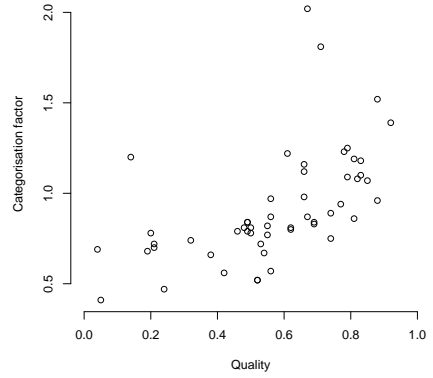Figure 3: Histogram of the categorisation factor.

Figure 4: Scatterplot of $q^2_{continuous}$ and the categorisation factor.

1.11 and 0.73, respectively ($t = 5.7$, $df \approx 38$, $p < .00001$).

In order to see what differences in the categorisation factors could be detected, we estimated a linear regression of the categorisation factors on country, topic, scale and whether the question was negatively phrased or not. An examination of the residuals suggests that there are no large deviations from the assumptions of linearity of the relationships and normality of the residuals[11]. The overall amount of variance in explained is 55%. The estimates can be found in table 13. This analysis is necessarily limited by the design of the experiments. For example, true-false scales were only used for the 'job' items–although this was of course not the only scale used in that experiment.

The categorisation factors did not differ very much across the different topics, except for the experiment on the role of women in society. In general, controlling for country and scale[12], the lowest categorisation factors were found in the 'doctors' experiment.

The country explains little of the variation in the categorisation factor. The exception however, is Estonia where a large coefficient was found which is likely to be different from zero. The average categorisation factor for this country keeping all other variables at zero is above one, indicating that the quality under the categorical model in this country was lower than the same quality estimate under the continuous model, scale, negative phrasing, and topic being equal. We should also note the higher standard errors for these coefficients. Given the relationship between the categorisation factor and the quality and the fact that the quality differs across the countries, it is likely that the coefficients are in fact higher than a statistical test would indicate at this point[13].

6   Discussion and conclusion

interpret meta analysis

---

[11]The largest standardised residual equals .34.

[12]No negatively phrased questions were used in this experiment; consequently, we cannot control for this variable.

[13]A bayesian analysis of this model using uniform priors on the coefficients suggests that the probability that the country effects are larger than 0.1 is around .8.

Table 13: 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)		0.53	0.19	0.16	0.90
Topic					
Doctors	(reference category)				
Efficacy		0.20	0.09	0.03	0.37
Job		0.23	0.19	-0.13	0.58
Women		0.51	0.13	0.25	0.78
Country					
Belgium	(reference category)				
Slovenia		0.09	0.10	-0.11	0.28
Czech Republic		0.28	0.15	-0.02	0.59
Switzerland		0.31	0.15	0.01	0.60
Greece		0.32	0.15	0.01	0.62
Denmark		0.33	0.17	-0.01	0.67
Estonia		0.53	0.19	0.15	0.91
Scale					
Agree-disagree		-0.16	0.07	-0.29	-0.03
Direct	(reference category)				
True-false		0.30	0.11	0.08	0.51
Scale position					
Negative		-0.50	0.07	-0.62	-0.33

Multiple R-Squared: 0.55; Adjusted R-squared: 0.46; F-test for factor sums of squares: Topic: $p = .15$; Country: $p = .15$; Scale: $p = .01$; Position: $p < .00001$.

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.

Flora, D. B. and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9:466–491.

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.

Johnson, D. R. and Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48:398–407.

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

Jöreskog, K. G. (1990). New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24:387–404.

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.

Millsap, R. E. and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39:479–515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49:115–132.

Muthén, B. and Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. *Mplus Web Notes*.

Muthén, 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. (frth). *Festschrift for Jaak Billiet*, chapter Differences in the quality of questions across countries.

Oberski, D., Saris, W. E., and Kuipers, S. (2004). 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. (2007a). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley Series in Survey Methodology. John Wiley & Sons.

Saris, W. E. and Gallhofer, I. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1.

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.

Satorra (1990). Robustness issues in structural equation modeling: a review of recent developments. *Quality and Quantity*, 24:367–386.