UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 10

2009

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Categorization errors and differences in the quality of questions across countries

Daniel Oberski[1,2] *
Willem E. Saris[1]
Jacques Hagenaars[2]

[1]ESADE, Universitat Ramon Llull. [2]Tilburg University.

August 18, 2009

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 the quality of the measures from the viewpoint of categorization. We assume that each category represents a range of scores on a latent continuous variable that have been grouped together, causing grouping errors. It depends on the distribution of values of the latent response variable in each category whether the intervals between the categories are equally far apart. If they are not, there is also transformation error. Both grouping and transformation are sources of measurement error due to categorization and therefore possible explanations for differences in the quality of questions. The results show that this effect is quite strong.

*The authors would like to thank the editors, as well as the participants of the 2007 IOPS winter meeting for their comments. For questions, remarks, or requests, please contact daniel.oberski at gmail

Introduction

Measurement error can invalidate conclusions drawn from cross-country comparisons if the errors differ from country to country. For this reason, when different groups such as countries are 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 functioning 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. To study these differences is important because they 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 the reasons behind them.

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., 2007). Because these factors did not explain much of the differences we now consider differences in categorization errors as a source of differences between countries.

Categorization errors are part of the discrepancy between an unobserved continuous variable and a discrete observed variable that measures the unobserved continuous variable. Specifically, categorization errors are the differences between the score on the latent variable and the observed category that are due solely to the categorization process.

For example, suppose a person's age is known only to belong in one out of three categories, which are assigned the scores one, two, and three, but there are never any mistakes in this categorization. In spite of the absence of mistakes, there is still a discrepancy between the age of the person and the category she is assigned to; first, because people of different ages have been lumped together. And second, the distance between the categories in terms of average age may not be equal to the distances of unity between the numbers one, two, and three, assigned to the categories. This means that if one treats the observed variable as an interval level measure, the result of calculations such as correlations will differ also from what would have been obtained if the original age variable had been used.

In general, one can say that categorization errors arise when a continuous latent response variable is split up into different categories. This leads to two types of errors: grouping and transformation errors (Johnson and Creech, 1983). Grouping errors occur when different opinions are grouped together in the same category. Transformation errors occur when the differences between the numerical values of adjacent categories do not correspond to equal distances between the means of the latent response variables in those categories. If, for instance, the distances between categories are not the same in two different countries, this can lead to larger categorization errors in one country than another, leading in turn to lower question quality. This is why the distance between

categories is a possible explanation for differences in question quality across countries.

The first section will discuss the models we use to estimate the measurement error coefficients of survey questions starting from a basic response model. We will then present the data from the European Social Survey that will be used. A short discussion of previous results follows. First the estimates from our previous research are shown. In a previous study, we already examined some possible explanations for the large differences in these estimates found across countries. These will be shortly reviewed. We then go on to present the model that will be the focus of this study, which accounts for categorization errors. It will be shown what we mean by such errors and how we compare the results we get from categorical models with those from continuous models. The statistical method of estimation is presented, after which we discuss our results. Since we have many such results, they are followed by a meta-analysis of the results. Finally, we discuss our general conclusions from this meta-analysis.

1 Theory

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

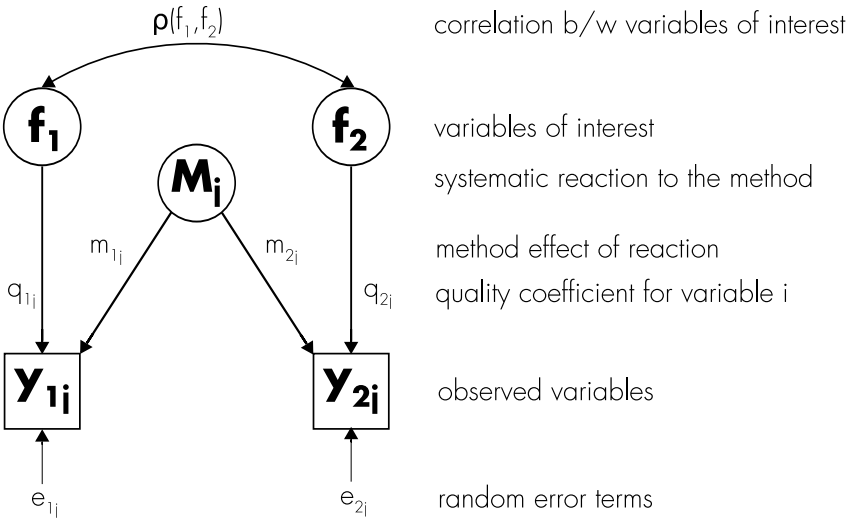

Figure 1: The continuous response model used in the MTMM experiments.

The difference between the observed response (y) and the variable of interest or concept by intuition (f) is both random measurement error (e) and systematic error due to the respondent's reaction to the method (M). This method effect is the only systematic error considered in the model.

An example of a method effect is when each respondent chooses her own reference points for all 11 point scales. For instance, an 11 point agree-disagree scale might label the highest category with the text 'disagree'. But in principle one can also disagree 'strongly' or even 'completely', although such opinions are not marked with a number on the answer scale. Thus it is up to the respondent to choose a location for these most extreme reference points. This choice will influence which category is finally chosen, given any opinion. Different reference points are generally chosen by different people if

these points are not fixed by the question, causing non-substantive random variation. If the same reference points are chosen by the same people given the same answer scale, then there will also be a correlation between the answers to all 11 point scales that has nothing to do with the respondents' opinions (Saris, 1988). This systematic variation can be considered method variance.

The coefficient q represents the quality coefficient and we call q^2 the total quality[1]. 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 $\rho(y_{ij}, y_{kj})$ between two observed variables measured with the same method is:

$$\rho(y_{ij}, y_{kj}) = \underbrace{\rho(f_i, f_k)}_{\textit{Correlation of interest}} \cdot \underbrace{q_{ij} \cdot q_{kj}}_{\textit{Attenuation factor}} + \underbrace{m_{ij} \cdot m_{kj}}_{\textit{Correlation due to method}} \tag{1}$$

where $i \neq k$ index the concepts by intuition and j a method.

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 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 quality coefficients, two method effects, and one correlation between the two latent traits, leaving us with five unknown parameters, while only one correlation can be obtained from the data. It is impossible to estimate these five parameters from just one correlation.

There are two different approaches to estimate these coefficients. The first is direct estimation from MTMM experiments. The second is the use of the prediction program SQP. SQP predicts the quality coefficient and method effect of a single question from many of its characteristics such as the topic, the number of categories, etc[2]. It is currently based on a meta-analysis of 87 MTMM experiments and 1028 different questions, while many more experiments are soon to be added (Oberski et al., 2004). In this study we use the MTMM approach.

Campbell and Fiske (1959) suggested using multiple traits and multiple methods to evaluate the quality of measurement instruments (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. Given the responses on all the variables, the coefficients described above can be estimated. A more elaborate introduction to MTMM and SQP can be found in Saris and Gallhofer (2007).

[1]One can also separate the reliability and method variance. This response model is known as the true score model and is more easily interpreted in terms of classical test theory, but mathematically equivalent to the classic MTMM model used here. For more details of the different models we refer to (Saris and Andrews, 1991)

[2]See the website `http://www.sqp.nl/`

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

The three traits were presented by the following three items:

- *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

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, collected in 2004.

The questionnaires were administered by face to face interviewing in all countries. In Finland, France, Ireland, Italy, the Netherlands, Norway, and Sweden, the supplementary questionnaire with the repetition questions was self-completed with the interviewer present rather than asked face to face. This confounds mode effects with country effects for these countries. The countries we compare in subsequent sections all used face to face interviewing for both questionnaires, however. Therefore mode effects are not an issue in this particular study.

The topics of the four MTMM experiments from the ESS we will study were the following:

1. The social distance between the doctor and patients;
2. Opinions about job;
3. The role of men and women in society;
4. Political efficacy.

Concerning each of these topics three questions were asked and these three questions were presented in three different forms following the discussed MTMM designs. 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

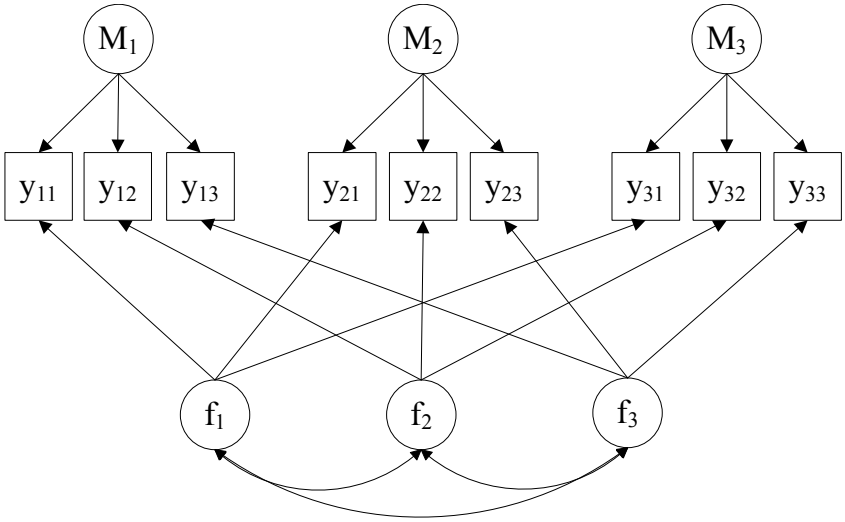

Figure 2: MTMM model illustrating the observed scores and their factors of interest.

but different groups got different version of the same questions (Saris et al., 2004). For the specific questions in the 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 'social distance' experiment examines the effect of choosing arbitrary scale positions as a starting point for agreement-disagreement with a statement. The 'job' experiment compares a four point true-false scale with direct questions using 4 and 11 point scales. In the 'role of 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. Finally, the political efficacy experiment pitted agree-disagree scales against direct questions.

A special group took care that the samples in the different countries were 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 2 shows the mean and median standardized 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

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

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

Table 2: 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

are considerable and statistically significant across countries ($F = 3.19$, $df = 16$, $p < 0.001$) and experiments ($F = 92.65$, $df = 5$, $p < 0.0001$). The highest mean quality is 0.79 in Portugal while the lowest is 0.57 in Finland. If the correlation between the constructs of interest is 0.60 in both countries and the measures for these variables have the above quality then the observed correlation in Portugal would be 0.47 while the observed correlation in Finland would be 0.34. 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. Not all of these differences are necessarily due to categorization, however. Below we discuss other possible explanations for some the differences.

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., 2007).

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 in the way intended. In particular, one item in the 'social distance' 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 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 should be administered directly after the main interview. Some respondents waited quite some time before answering the supplementary questions. In the time between the two interviews their opinions may have changed, or have been influenced by new considerations unique to that 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 of people who had returned the questionnaire on the same day was by itself very similar in the quality to other countries.

The third alternative we considered was that the language of the questions might 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 could explain the 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 Scandinavian 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 found in the ESS. Therefore we now turn to the possibility that the distance between the categories in the categorical questions differs from country to country. Before we proceed to investigate the influence of categorization errors on the quality in different countries and experiments, we explain in more detail the model used to estimate the distances between the categories.

3.1 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 most likely ordered categories rather than measured on an interval scale. Broadly speaking, two types of measurement models have been proposed for this situation. The first assumes that there is an unobserved discrete variable, and that errors arise because the probability of choosing a category on the observed variable given a score on the unobserved variable is not equal to one. That is, the errors are modelled by 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' (LRV) is thought to underly the observed categorical item. Such models are sometimes called latent trait models. Several 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 the data (figure 3)[5].

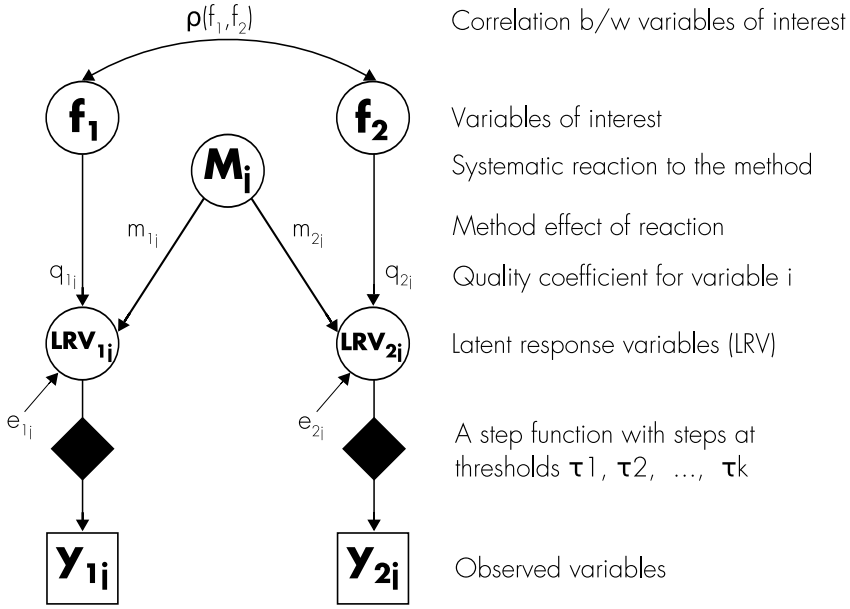

Figure 3: The categorical response model used in the MTMM experiments.

Errors may arise at two stages. The first is the connection between the latent response variable (LRV_{ij} in figure 3) and its latent trait (f_i). This part of the error model is completely analogous to factor analysis or MTMM models for continuous data: the scale is modeled as a linear combination of a latent trait (f_i), a reaction to the particular method used to measure the trait (M_j), and a random error (e_{ij}), and interest then focuses on the connection between the trait and the scale (q_{ij}), which we again term the 'quality coefficient' (see also figures 1 and 2).

The second stage at which errors arise differs from the continuous case. This is the connection between the variables LRV_{ij} and y_{ij} in figure 3. 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. In figure 3 this step function has been represented by a black triangle. Examples of step functions are illustrated in figure 4.

In figure 4, the steps (solid line) show the relationship between the LRV and the observed variable, while the straight (dotted) line plots the expectation of the LRV given the latent trait. In the step function on the left-hand side, the LRV has been categorized using equal intervals. The error that is added by the categorization is the vertical distance between the dotted line and the step. That is, the distance between the dotted line and the horizontal segments of the solid line. It can be seen that the error is zero when the

[5]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 categorization 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).

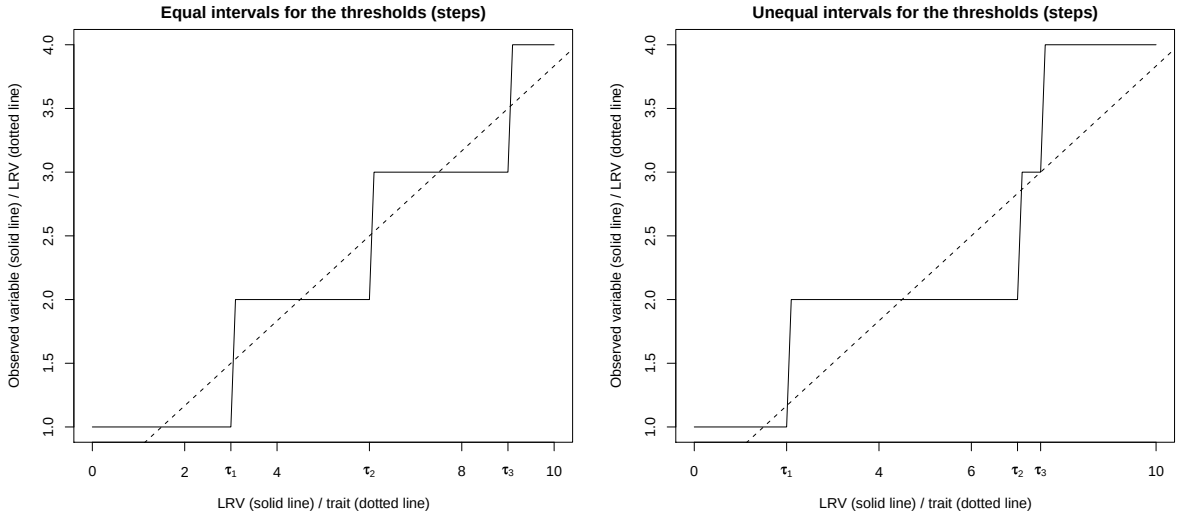

Figure 4: Two hypothetical step functions which result from categorization. The solid lines plot the observed categorical variable as a function of the latent response variable (LRV). The diagonal dotted lines plot the expectation of the LRV as a function of the latent trait on the same scale. The thresholds used for categorization are denoted by the symbols τ_1, τ_2, and τ_3.

straight line crosses the steps, and that at each step, the error is the same (at 3, 6, and 9). The expectations within the categories have the same interval as the thresholds of unity, and so if the values 1, 2, 3, and 4 are assigned to the categories, no transformation occurs. Errors still occur, because the values along the dotted line have been grouped into the four categories formed by the solid line. Relationships of the observed categorical variable with other variables will therefore be attenuated.

Conversely, the right hand side shows a latent response variable that has been categorized with unequal steps. The figure shows that the distances between the thresholds τ_1, τ_2, and τ_3 are very different from each other. The consequence is that at the second step, i.e. in between τ_2 and τ_3, there is almost no extra error, while at the first and third steps the errors are much larger. Here a transformation occurs. Suppose that the categories are given the numerical values 1, 2, 3, and 4, as is often done. Then the distances between the expectations of the LRV in each of the categories do not equal unity, which is the distance between the values chosen for categories.

To sum up, two types of errors can be distinguished at this stage (Johnson and Creech, 1983):

1. *Grouping errors* occur because the infinite possible values of the latent response variable are collapsed into a fixed number of categories (the vertical distances between the diagonal line and the steps in figure 4). 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.

We have described the categorization process here. Is is important to note, however, that normally this process is not observed and one only observes a discrete variable, which we then assume is the result of this process.

Categorization, 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.

Thus, the model we use allows to a certain extent for the separation of errors due to the categorization, 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 categorization introduced across countries.

3.2 Categorization 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 variable model (hereafter referred to as CV 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 categorization. For this reason Coenders (1996) argued that the linear MTMM model assuming continuous variables does not ignore categorization errors, but absorbs them to a certain extent in the estimates of the random error and method correlations. How this absorption functions exactly will depend on the model in use and is not extensively studied. The extent to which it holds in general is thus a topic that is still under discussion.

However, since the quality coefficient is estimated from the covariance matrix of the measures, it can be both reduced and increased by categorization errors. In general all correlations between measures increase after correction for categorization, but they need not all increase equally. For example, consider again table **??**. If categorization errors are higher using the first method, 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 response variable due to the method will be larger in the categorical model than in the CV 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. Categorization 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.

As noted before, the quality 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

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

The scale of LRV, the latent response variable, is arbitrary, except that it may vary across countries due to relative differences in variance (Muthén & Asparouhov, 2002). However, the ratio $Var(LRV)/Var(y)$ can easily be calculated once q^2_{con}, the quality from the continuous analysis, and $Var(f)/Var(LRV)$, the quality from the categorical MTMM analysis ($q^2_{cat.}$), have been obtained. So equation (3) shows that $q^2_{con} = q^2_{cat} \cdot c$ and

$$c = \frac{q^2_{con}}{q^2_{cat}},$$

where c is the categorization effect, or

$$\ln(q^2_{con}) = \ln(q^2_{cat}) + \ln(c).$$

This correction factor 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 'categorization factor' for different countries and experiments, and examine to what extent it can explain the differences in quality across countries.

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 so-called 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 different parts of the model are identified, so is the entire model. Since we can identify the necessary covariances in the categorical model, this is identified as well (Millsap and Yun-Tein, 2004).

For each experiment, two different models were estimated. The continuous analysis was conducted using the covariance matrices as input, and estimated using the maximum likelihood estimator in LISREL 8. The results presented in the tables below were standardized 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 used an alternative two step approach, whereby in the first step the covariance matrices of the latent response variables 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[6].

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 model categorization errors using threshold parameters. These thresholds are the theoretical cutting points where the continuous latent response variable (LRV) has been discretized into the observed categories. If the thresholds are different across countries, the questions are not directly comparable, since differences in the frequency distribution are partly due to differences in the way the LRV was discretized. If the thresholds are the same across countries the questions may still not be comparable due to differences in linear transformations (loadings) and random errors. But in that case it is not categorization error that causes incomparability. A final possibility is that loadings, random errors, and thresholds are all the same across countries. In that case the frequency distributions can be directly compared.

In this paper we will perform only a basic invariance test on thresholds. If the thresholds are equal, categorization error is not a likely cause of differences in quality. However, we do not continue with tests for invariance on loadings and error variance, but will compare the results of the two different models.

The two models are the same with respect to the covariance structure of the response variables (the 'MTMM part' of the model). However, they differ in their basic assumptions about the 'observation part' of the model: the CV model assumes that the continuous response variables have been directly observed, while the categorical model assumes a threshold connection between the response variables and the observed ones.

Both models assume normality of the response variables, but the differences in basic assumptions cause the categorical model to be more sensitive to departures from normality. While in the CV model, under quite general conditions, violation of normality will not affect the consistency of the estimates (Satorra, 1990), this is not so in the categorical model. There, the threshold estimates are derived directly from quantiles of the normal distribution which the latent response variable is assumed to follow. Therefore, if the LRV's are not normally distributed, the threshold estimates will be biased. The MTMM estimates depend on the thresholds and can also change, though the precise conditions

[6]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).

under which such estimates would change significantly have, to our knowledge, not been investigated analytically. It has been found in several different simulation studies that bias may occur especially when the latent response variables are skewed in opposite directions (Coenders, 1996).

Thus, while the categorical model may be more realistic in modelling the observed variables as ordinal rather than interval level measures, the CV model may be more realistic in that it is robust to violations of normality[7]. In any particular analysis, whether one or the other model provides a more adequate estimate of the quality of the questions therefore depends on the degree to which these assumptions are violated[8]. This should be kept in mind in the interpretations of the results.

We estimated the quality of the measures based on the CV 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 experiment, the countries with the highest and the lowest qualities in the CV model were analysed. For each of the questions we took the ratio, called 'categorization factor', of the two different quality measures as an index of the effect that categorization has on the continuous quality estimates. The next section presents the results.

5 Results

5.1 Results of the experiments

The first experiment's results will be described in some detail, while we provide the results of the other experiments in the appendix.

The first experiment concerned opinions on the role of women in society (see table 3). We first turn to the hypothesis that all thresholds are equal across different countries. If this hypothesis cannot be rejected there is also little reason to think that the categorization 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 was 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 Slovenia. 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.

[7] One important point to make here is that even when univariate distributions such as histograms and tables of the observed categorical variables are highly non-normal, this does not necessarily imply that the normality assumption of the categorical model is violated. The reason is that a very non-normally distributed observed variable may be the consequence of a perfectly normally distributed variable that has been categorized in a very uneven way.

[8] In principle the normality assumption on the latent response variables is testable. However, the question then still remains what impact any non-normality would have on the estimates. This question is beyond the scope of the present paper.

Table 3: The 'role of 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 standardized values are very large with some values close to 0.9 while in other countries the highest obtained and exceptional value is 0.6. 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 two countries.

The model where all thresholds are constrained to be equal yields a likelihood ratio of 351 and 36 degrees of freedom ($p < 0.00001$). This model should therefore be rejected: the thresholds are significantly different across countries.

We use the procedure of Saris et al. (frth) to determine whether misspecifications are present in the model. For this test we need the Expected Parameter Change (EPC), Modification Index (MI) and the power of the test. The EPC gives direct estimates of the

Table 4: Quality (q^2) and method effects (m) according to the continuous and categorical models, with categorization 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
Categorization factor				
	Greece	1.4	1.9	1.5
	Slovenia	0.8	0.9	1.0

size of the misspecification for all fixed parameters, while the MI provides a significance test for the estimated misspecification (Saris et al., 1987).

However, these two indices are not sufficient for determining misspecifications because the MI depends on other characteristics of the model. For this reason, the power of the MI test must be known in order to determine whether a restriction is misspecified. We use these quantities to incrementally free parameters that were indicated to be misspecified.

Using the modification indices and power as guides, we formulated a new model in which some thresholds were constrained to be equal, while others were freed to vary. Equality of thresholds is not required to estimate the relationships, but it is useful because the equality of thresholds allows for differences in variances of the response variables across the groups. This is in contrast with the use of polychoric correlations where the variances are constrained to be equal across the groups.

The resulting model has an approximate likelihood ratio of 2.8 on 2 degrees of freedom ($p = 0.24$)[9]. The resulting estimates of the threshold parameters are presented in table 3. These estimates have been expressed as z-scores in order to make them comparable.

Table 3 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).

The thresholds in this model represent how extreme the 'agreement' has to be before the next category is chosen rather than the previous one. This strength is expressed in z-scores, i.e. standard deviations from the mean. Take, for instance, the third statement in the table: "Men should take as much responsibility as women for the home and children". Slovenians need to have an agreement differing from the country mean 2.6 times more than the standard deviation, before they will respond 'disagree strongly'.

Note that the threshold part of the relationship between LRV and observed response

[9]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 categorization.

is deterministic. However, not all Slovenians with an *opinion* on the indicator of 2.6 standard deviations or more away from the mean will necessarily answer 'disagree completely'. This is so because the latent response variable is also affected by random measurement error. The combination of the threshold model and normally distributed random measurement error gives rise to a familiar probit relationship between indicator and response. Because the random error plays an important role in this relationship, not only the thresholds should be discussed here, but also the quality coefficients.

Looking at the first question, it can be seen that the distances between the thresholds are unequal for these two countries and different from one. One can also see that the endpoints are somewhat distant, especially in Slovenia: there the category 'disagree strongly' is 1.8 standard deviations or more away from the mean, reducing the number of scale points that are available for some people.

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 categorization 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.

It is also clear from the table that the two forms of phrasing are not exactly opposite in the way they are understood and/or answered. This is especially true for the 'right to a job' item. However, the choice for one phrasing or the other seems arbitrary. This particular way of phrasing a question is therefore inadvisable, because a decision that seems arbitrary is not arbitrary in its consequences. The key problem in this case may be the complex sentence structure in which men are compared to women, given an attribute (right to a job) under a certain condition (when jobs are scarce), and then a 'degree of agreement' with a norm ('should have') is asked. A more accurate way of measurement that may be less sensitive to such arbitrary shifts in response behavior might be to ask questions about the rights men and women should have according to the respondent directly.

The thresholds provide some insight into the nature of differences in categorization. However, the quality of the measure in the continuous model depends also on parameters of the categorical response model such as the method effects and the error variances, and on the latent response variable distribution.

Besides the thresholds also the correlations between the LRVs are estimated. Based on these correlations the MTMM model mentioned before has been estimated and so estimates of the quality and method effects of the measures corrected for categorization are estimated for all questions. The quality and method effects of the CV model have also been estimated. The results are presented in table 4. Based on these results the categorization effect can be derived because it is the ratio of the two coefficients. This result, too, is presented in table 4.

The top two rows of table 4 show that the quality in Greece was higher than in Slovenia using the CV 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 categorization errors.

However, table 4 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 ($\chi^2 = 20$, $df = 10$, $p = 0.02$). While for the first and second questions the low qualities are indeed corrected upwards somewhat after the categorization 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 CV 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 2).

The analyses of the other three experiments show that sometimes no large differences between the countries are found, while in others the thresholds are rather different. In particular we found several cases where the same question did not cover the distribution of the opinion in one country, but provided more information in another. We also found both examples of cases where differences in the quality do not go together with differences in the thresholds, and examples of cases where they do. A more detailed discussion of the results for the other three experiments can be found in the appendix.

Now that we have presented and discussed the results of one experiment in detail, the question remains whether there is a connection between the categorization factor and the quality of the question. The next section therefore presents the results of a meta-analysis we conducted on the categorization factors.

5.2 A meta-analysis of the results

The question remains whether the categorization factor affects the quality or not. Using the results presented in the previous sections, we constructed a data set consisting of the categorization factor for all questions–including those from the supplementary questionnaire not shown above–in the four different experiments for which this index was available. This yielded 72 cases in total.

As shown before, the categorization factor equals $c = q^2_{con}/q^2_{cat}$, and so $q^2_{con} = q^2_{cat}(c)$. If there were no effect of the categorization, then there would be no relationship between c

and q^2_{con}, since q^2_{cat} would be higher or lower by a constant factor. If c and q^2_{con} are plotted against one another, one would then expect to find the points randomly distributed along a horizontal line. Figure 5 shows the scatter plot of these two quantities. Estimates from different experiments have been indicated with different symbols.

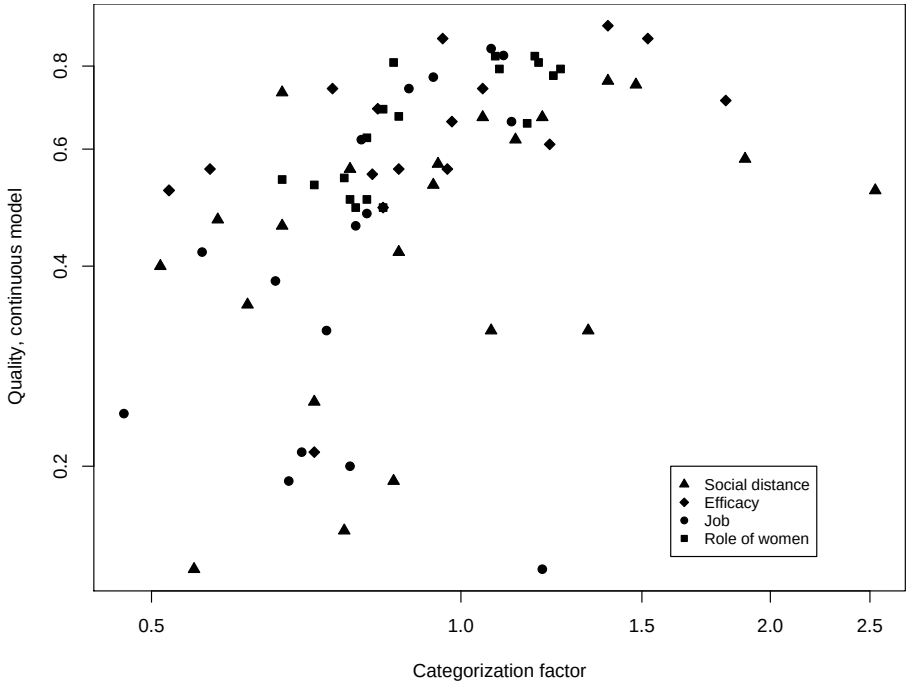

Figure 5: Scatterplot of the categorization factor (c) and the total quality of a measure (q^2_{con}) across the experiments. Note the log-log scales.

The clear relationship that can be seen in the figure indicates that high quality coefficients from the continuous model tend to be lower in the categorical model, and vice versa. Figure 5 shows that categorization factors above unity were mostly found for questions with a high quality. We can estimate the relationship between the quality from the continuous model for each experiment easily by the transformation $\ln(q^2_{con}) = \alpha_k + \beta_k \ln(c)$. Here k indexes the four different experiments. Note that the base level of q^2_{con} is $\exp(\alpha_k)$.We then fit a linear regression to the transformed variables. The resulting predictions for each experiment are shown in figure 6 on the original scales.

Figure 6 shows that both the intercepts and slopes for the 'efficacy' and 'job' experiments are rather similar, while the coefficients for the 'role of women' and 'social distance' experiments are completely different. The effect of the categorization factor is strongest in the 'social distance' experiment, where also some large differences between the threshold distances were found (see appendix). The experiment with the smaller number of categories, 'job', does not have a high coefficient.

We now turn to the question if these factors are also different between the countries with 'high' and 'low' quality coefficients. If the sample is split according to whether the quality was 'high' or 'low', the means of the categorization factors of the two groups are 1.25 and 0.85, respectively, for the questions in the main questionnaire ($t = 3.7$, $df \approx 18$, $p = 0.002$. For the questions in the supplementary questionnaire, the difference is in the opposite direction, but not statistically significant ($t = -1.70$, $df = 28$, $p = 0.10$).

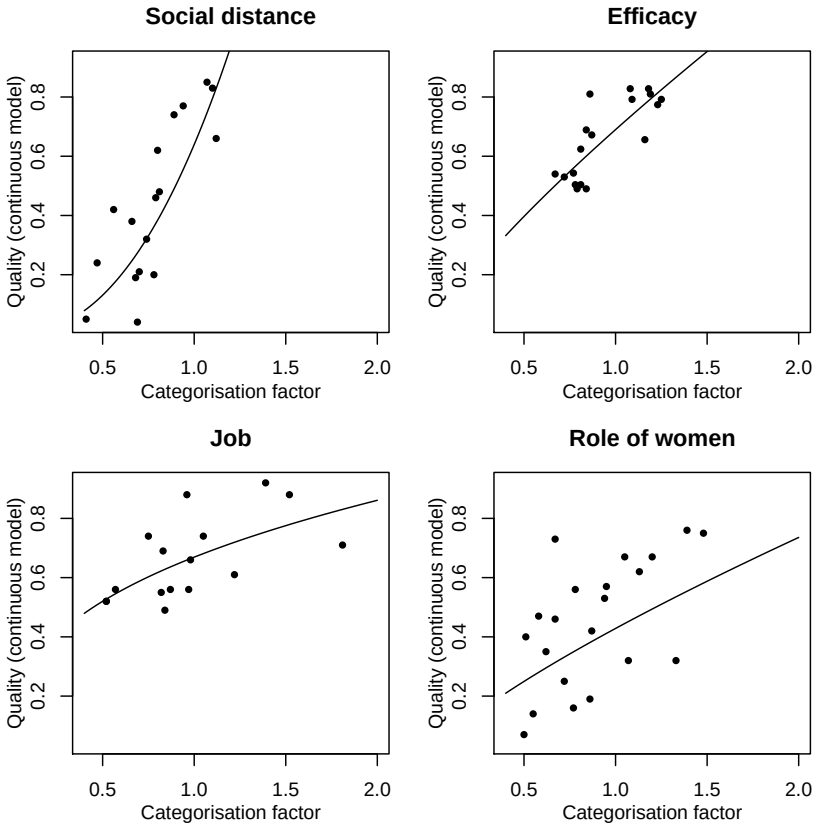

Figure 6: Scatterplot of the categorization factor (c) and the total quality of a measure (q^2_{con}) by experiment. The prediction line of the model $\ln(q^2) = \alpha + \beta \ln(c)$, as estimated for each experiment separately, is also given.. For the numerical estimates of these coefficients, please see the appendix.

This suggests there is a considerable effect of the categorization, at least in the main questionnaire.

One possible explanation for the interaction effect found here is that the method factors were often constrained to zero for the main questionnaire. The questions in the main questionnaire were selected especially because they were expected to have the highest quality and lowest method effects. After the initial continuous analysis the model often indicated that the questions in the main questionnaire indeed had zero method correlation. Since the categorical model tends to increase the correlations, if the monomethod correlations for the main questionnaire go up more than the other correlations, then it can happen that in the categorical model a method factor is found where none was found before. This will then lower the quality estimates.

A test was done of the hypothesis that questions for which the method effect was constrained to zero in the continuous model have the same categorization factor as other questions, controlling for country effects. This hypothesis was rejected ($p = 0.02$)[10]. The explanation that constraining the method factors to zero causes the interaction found

[10]Result of a hierarchical linear model fit using `R 2.6.1` with fixed effects of country and restricting the method to zero or not (0/1), and a random intercept across topics to account for the dependency among the observations.

above therefore seems plausible.

6 Discussion and conclusion

Using the multitrait-multimethod design and model in the ESS, we found large differences between countries in the quality of survey questions. Because such differences can have important implications for cross-country research and survey design, we set out to discover whether these differences could not be attributable to errors due to the use of a small number of categories.

Overall, we found that categorization errors do occur besides random errors and method effects. These errors have two types of effects on the quality of the questions, which can work in opposing directions. The first is that the quality is lower when there is more categorization error. The second, that the categorization attenuates the relationships between different variables in the model differently, affecting not only the quality, but also the method effects and other parameters of the model. This in turn has as its consequence that the quality parameter under the CV model is not always smaller than the quality under the categorical model, as evidenced by the many 'categorization factors' above unity which we found.

A caveat should be added to the interpretation of this result, because a violation of the assumptions of the models (no categorization error versus bivariate normality) can have different consequences for the estimates. It is therefore not necessarily true that a categorization factor above unity indicates overestimation of the quality in the CV model. Several studies of the robustness of factor analysis models to categorization errors exist (see Olsson 1979). However, we found that their results do not necessarily apply in the MTMM model, which also includes method factors. Given the ubiquity of correlated errors in survey questions, it would be useful to study more closely the robustness of this particular type of measurement error model to categorization error. This, however, is beyond the scope of the present paper.

In a meta-analysis, we gathered the results from our four different experiments and analysed the relationship between the categorization factor and the quality in the continuous model. Effects were found for all four experiments.

If the categorization factors were equal for countries with the highest and lowest quality coefficients, they could not explain the differences in quality which we found earlier. The meta-analysis suggested that there is a considerable difference in the categorization factor between countries where the highest and the lowest quality coefficients were found given whether the question was part of the main or supplementary questionnaire.

The methods in the main questionnaire were chosen beforehand based on other experiments as the ones least likely to cause method effects. For example, direct questions rather than batteries were used. After re-examining the experiments on which the meta-analysis was based, it appears this is closely related to the interaction effect found there.

The main reason for the interaction effect we found in the meta-analysis appears to be that the method variance for the main questionnaire method was often close zero. The general rise in correlations that results from correction for categorization seems to have 'pushed' the monomethod correlations of the main questionnaire variable to the point where the method variance could not anymore be constrained to zero. And as the

method variance rises, the quality must decrease in our model.

In other words, the correction for categorization has a negative influence on the quality. When the method factors were constrained to zero in the first instance, the effect was that the quality was in general lower in the categorical model than in the continuous model. This is contrary to what one might expect considering that all of the polychoric correlations are higher than their Pearson counterparts.

In this study we have shown that it is possible to split the measurement error model into three parts:

- A part due to random errors;
- A part due to systematic errors;
- A part due to splitting the variable into just a few categories: 'categorization error'.

This study has been largely descriptive of the effects of categorization error. Given our findings, it seems important to better judge the relative merits of the continuous and categorical models, and the effects that different question characteristics have, not only on quality and method effects, but also on the categorization errors.

Our study also has some limitations due to the assumptions made to attain the above separation. These are: normality of the latent response variables, linearity of the relationship between the latent traits and latent response variables, and interval measurement of the latent traits. In another paper these issues will be addressed by examining ways to relax the assumptions. Future research might also focus on finding other explanations for differences in quality across countries.

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. T. and O'Connell, E. J. (1967). Method factors in multitrait-multimethod matrices: multiplicative rather than additive? *Multivariate Behavioral Research*, 2:pp. 409–426.

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. (1998). Relationship between a restricted correlated uniqueness model and a direct product model for multitrait-multimethod data. In Ferligoj, A., editor, *Advances in methodology, data analysis and statistics: Metodoloski Zvezki*, pages 151–172. FDV, Ljubljana, Slovenia.

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

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65:pp. 241–261.

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.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12:pp. 247–252.

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. (2007). Why are there differences in measurement quality across countries? In Loosveldt, G., Swyngedouw, M., and Cambré, B., editors, *Measuring Meaningful Data in Social Research*. Acco, Leuven.

Oberski, D., Saris, W. E., and Kuipers, S. (2004). SQP: survey quality predictor.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, 14:p485–500.

Saris, W. E. (1988). *Variation in Response Functions: A Source of Measurement Error in Attitude Research*. Sociometric Research Foundation.

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.

Saris, W. E., Satorra, A., and Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, 17:105–129.

Saris, W. E., Satorra, A., and Veld, W. V. D. (2008, frth). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: an interdisciplinary journal*.

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