UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 11

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

Latent Class Multitrait-Multimethod Models *

Oberski, D.[1,2], J. Hagenaars[1], and W.E. Saris[2]
[1] Tilburg University, department of Methodology and Statistics, The Netherlands
[2] Universitat Pompeu Fabra, Spain

December 17, 2009

Abstract

The present paper suggests a statistical method, the latent class MTMM model, of estimating the quality of single questions while making fewer assumptions than have been made so far in such evaluations. The method is a combination of the multitrait-multimethod research design of Campbell and Fiske (1959), the basic response model for single questions of Saris and Andrews (1991), and the latent class factor model of Vermunt et al. (2004). The latent class MTMM model is thus not novel in itself, but combines an existing design, model, and method to improve the analysis of single questions in survey research.

A real experiment from the European Social Survey (ESS) is analyzed and the results are discussed at length, yielding valuable insights into the functioning of these questions.

Introduction

Since the late 19th century, psychometricians have studied the measurement quality of scales. With the advent of item response theory (IRT), the focus has shifted somewhat from scales *per se* to the quality of *indicators* as measurements of the scale (Hambleton et al., 1995). An IRT analysis of items provides more information about the functioning of the different indicators of the scale, separate from the properties of the scale as a predictor of behavior.

However, in some cases or disciplines, only one indicator may be available, an indicator may be used for different scales, or different countries must be compared with each other. Furthermore, a scientific interest exists among survey researchers in the effects of different design choices on the question quality, separate from the scaling properties of an indicator. In these cases, we argue, it is important to study the quality of *single questions* as a measurement of the indicator: the focus should then be shifted from indicators to single questions.

The present paper suggests a statistical method of estimating the quality of single questions as measurements of an indicator, while making fewer assumptions than have been made so far in the evaluation of single questions. The method is a combination of the multitrait-multimethod research design of Campbell and Fiske (1959), the basic response model for single questions of Saris and Andrews (1991), and the latent class factor model of Vermunt et al. (2004), originally formulated by Lazarsfeld and Henry (1968). The latent class MTMM model is thus not novel in itself, but combines an existing design, model, and method to improve the analysis of single questions in survey research.

The data obtained from multitrait-multimethod experiments (Campbell and Fiske, 1959) allow for a separation of systematic errors due to the method of asking a question and random measurement errors from the indicator of interest (Schmitt and Stults, 1986). By applying the latent class factor model, we obtain very precise information about the way responses are generated from underlying opinions on single indicators.

We discuss the method by applying it to a real dataset from a multitrait-multimethod experiment done in the European Social Survey (ESS). In an earlier study this experiment and several others were analyzed using the commonly applied confirmatory factor analysis and ordinal probit models (Oberski et al., 2007).

*Thanks are due to Jeroen Vermunt for his patient and invaluable explanations.

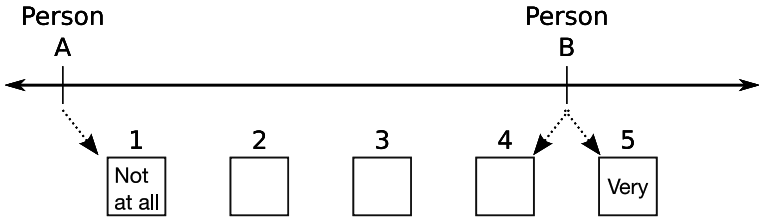


Figure 1: Theoretical true score and response options for the question 'How happy are you?'. The choices to be made when going from true score to response options are not always obvious.

The assumption of normally distributed latent response variables made in those analyses – that is, of an ordinal probit relationship between trait and indicator with parallel cumulative probability curves – may be false. The present application shows how this assumption can be relaxed. The latent class approach has the advantage that many assumptions that are usually made can be investigated. Among them are the measurement level (nominal, ordinal, or interval) of the observed variables, and the distribution of the latent variables. The model does not require the assumption of normally distributed latent variables, since the marginal distribution of the latent variables is left to be estimated.

The next section argues that it is essential to estimate the quality of single questions. We then explain the experimental design and the response model applied to analyze this quality. The rest of the paper applies this model to a real dataset, presented in the subsequent section. We then briefly note the software and methods used, after which the results of the analysis are discussed. Finally, conclusions are drawn from the analysis, showing the added value of our approach.

1 Measurement error in single questions

Answers to survey questions cannot be taken for granted. There are random and systematic components in the answers given by respondents that have nothing to do with the opinion the question was supposed to measure. Such components are therefore measurement errors.

Systematic components arise because different people have their own idiosyncratic way of answering questions given their opinion (Saris, 1988). Some give extreme answers on five-point scales while others tend to choose the middle point, for instance (Hui and Triandis, 1989). Some are more sensitive to social desirability than others, causing differences depending on how the question is phrased (Crowne and Marlowe, 1960). One may also say that respondents 'satisfice', using simplifying answering strategies to reduce cognitive burden (Krosnick, 1991). These processes are distinct but have in common that they may cause two people with the same underlying opinion to give different answers, and will cause two answers to unrelated questions answered by the same person to correlate.

Such systematic ways of answering the question vary across people, but may be stable across questions. They therefore cause both error variance and spurious relationships between answers to questions asked in the same way. If the way of answering a question is specific to both person and method, it is called a 'method effect'. An example would always choosing to agree or disagree 'completely' on agree-disagree scales, but not on other kinds of scales: this would be extreme response behaviour specific to the method. If the same respondent has a tendency to choose the extreme categories for *any* type of answer scale, the systematic error is called a 'style factor' (Jackson and Messick, 1958). Crucially, neither method effects nor style factors are related to the question content.

Random error is another source of measurement error; after the respondents have moulded their opinion into the form required by the question, some element of arbitrariness in choosing a response option may still remain. Consider the lines in figure 1. The possible opinions after correction for systematic effects or 'true scores' (Lord and Novick, 1968) of the respondent are represented as a line, while the response options below are categorical. Person A would presumably have no difficulty choosing 'not at all'. However he or she may make a mistake and accidentally mark option 2 rather than option 1. Person B, at the same time, could equally well choose options 4 or 5, and might do so at random from occasion to occasion. Both processes may occur at the same time and give rise to random measurement error.

This suggests that answers to survey questions contain random and systematic measurement errors.

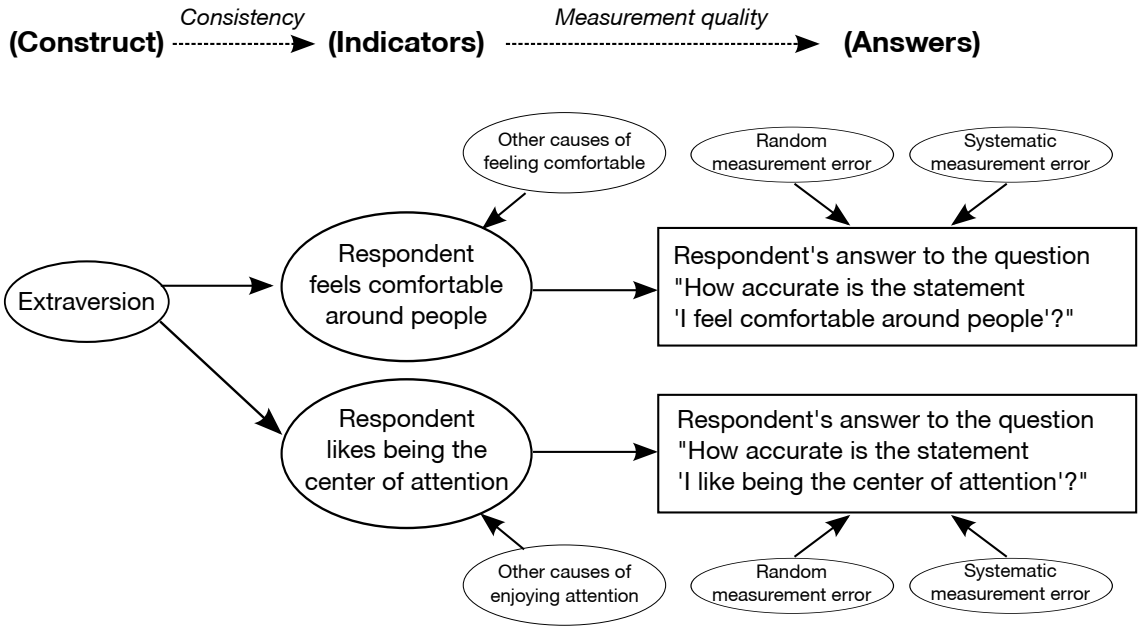


Figure 2: Illustration of the difference between pure measurement quality (the relationship between observed answer and unobserved indicator) and the consistency of indicators (the relationship between the unobserved indicator and the unobserved construct). In the present paper we will only study the connection between indicator and observed answer: pure measurement errors. The indicators are taken from the International Personality Item Pool (http://ipip.ori.org/)

.

Estimating such errors (1) assesses the general quality of a question; (2) allows for the correction of study quantities of interest such as regression coefficients or group differences for the influence of errors; (3) assesses the cross-group comparability of quantities of interest.

The question has been asked for the purpose of measuring a construct. We term the degree to which the indicator, after correction for pure measurement error, measures this construct the 'consistency' of the indicator (Saris and Gallhofer, 2007b). The combination of measurement error and consistency has been called 'construct validity' by Andrews (1984). An illustration of the distinction between measurement errors and consistency is given in figure 2.

Assessing the general quality of items that form a scale and their cross-group comparability is a fairly common activity in psychological research[1]. This quality concerns both the degree to which an indicator is influenced by a construct ('consistency') and the pure measurement errors discussed above. There are, however, advantages to estimating the pure measurement errors separately rather than this combination.

First, there is a scientific interest among survey researchers in the effect on the quality of the questions of various choices to do with survey design. Such choices could refer to the number of response options, use of an agree-disagree scale, linguistic complexity, etc. (Saris and Gallhofer, 2007a). They can also refer to nonresponse (Olson and Kennedy, 2006), or the study of special populations such as immigrants or elderly people (Groves, 2005). In order to separate this effect on quality of survey design from effects on consistency with the construct, it is necessary to estimate measurement error separately.

Second, in studies that compare groups such as cross-national research, the measures must be invariant across groups: only measures with equal consistency across groups allow for comparisons. Having only the combination of measurement and consistency error available results in the stricter requirement that both must be equal across countries. Saris and Gallhofer (2007b) argued that such tests are unnecessarily strict

[1]Dividing the number of matches in Google Scholar (http://scholar.google.com/) to each of the APA's 'core of psychology' four largest impact factor journals (including Psychological Methods) by the number of matches adding the term 'differential item functioning' suggests DIF is mentioned an average of 6%. If the percentages are weighted by the journal's impact factor in 2007, the average is about 4%. Although DIF is not mentioned very often, it is clearly a well-known technique.

and only the higher-order relationship between construct and indicators need to be invariant. Such a test requires separation of measurement error from consistency.

The third reason for estimating pure measurement errors is that in the social sciences, there are few standardized scales. Consequently, questionnaires often contain only one question instead of a number of questions to measure a single construct. A classic example in sociology is the question used to measure social trust: "Would you say that most people can be trusted, or that you can't be too careful in dealing with people?"[2]. Furthermore, it may also happen that different researchers construct different scales *post-hoc* using the same questions. Examples in political science are questions on citizens' trust in various political institutions. In such cases, estimating the extent of measurement errors in the question allows for the correction of the attenuation of relationships with other variables due to errors, and provides an upper bound for construct validity.

For these three reasons it is essential to estimate the quality of single questions. Two general approaches are possible: longitudinal designs using quasi-simplex models (Alwin, 2007) and 'multitrait-multimethod' (MTMM) experiments (Campbell and Fiske, 1959). We will discuss an approach of estimating the quality of single questions based on MTMM experiments that requires fewer assumptions than the approaches used so far for such data. An approach similar to quasi-simplex models for longitudinal designs such as the ones discussed in Alwin (2007) was discussed by Biemer and Bushery (2000)[3].

2 Multitrait-multimethod experiments

Campbell and Fiske (1959) suggested measuring multiple indicators ('traits') by multiple methods (MTMM). The correlations thus obtained were posited to follow a certain pattern. Later, different models were proposed to analyze these patterns, of which confirmatory factor analysis is the most commonly used (for a review, see Schmitt and Stults (1986)).

What the models applied to MTMM data have show is that the MTMM design can be used to separate the relationship between the indicator to be measured and the observed variable from random and systematic measurement errors. Note that here we mean by traits the indicators in the sense of figure 2 rather than the construct.

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 one trait measured with three different methods is given in figure 3.

Collecting data using this MTMM design, data for nine variables are obtained. These variables become the subject of a measurement or MTMM model. There is an ample literature about MTMM models using confirmatory factor analysis and the different choices that can be made for such models. Here we wish to start from a more general model formulation that specifies the relationships between the latent and observed variables without necessarily being a confirmatory factor analysis.

3 The response model

Figure 4 specifies the relationships between the observed scores and their general factors of interest as a graph. This figure shows that each trait (T_i) is measured in three ways. It is assumed that the traits are dependent but that the method factors (M_1, M_2, M_3) are independent.

In figure 4, y_{11} through y_{33} are the observed variables belonging to the experiment. The first digit (i) corresponds to the trait number and the second (j) to the method number. Following the graph, each trait is indicated with T_i and each method with M_j, In total there are $I = 3$ traits and $J = 3$ methods.

The quality of a measure is the strength of the relationship between the trait and the indicator that is supposed to measure it. The amount of systematic error or method effect depends on the strength of the relationship between the method factor and the indicators measured using that method. It should be noted here that a drawback of the MTMM design we use is that one cannot separate method effects from other systematic errors. Thus an assumption is made that all systematic errors are specific to the method used. In

[2]The question was devised by Noelle-Neumann in 1948 in Germany. Later Rosenberg (1956) created a multiple item concept (scale) using this question. But to date, many questionnaires only contain the single question.

[3]It should be noted that the notion of reliability estimated by longitudinal models is different from that employed here: in the longitudinal studies mentioned unique considerations of the moment that form part of the true variance are included as measurement error (van der Veld, 2006).

Method 1

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

	Not at all true	A little true	Quite true	Very true	(Don't know)
There is a lot of variety in my work.	1	2	3	4	8

Method 2

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

Please tick one box.

- ☐ 1 Not at all varied
- ☐ 2 A little varied
- ☐ 3 Quite varied
- ☐ 4 Very varied

Method 3

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

Please tick the box that is closest to your opinion

Not at all varied										**Very varied**
0	1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 3: The trait 'perception of variety of job' measured by three different methods.

an investigation of different explanations for correlated errors in MTMM data, Corten et al. (2002), provided some evidence that this assumption is reasonable.

The most common model applied to this graph is the continuous confirmatory factor analysis (CFA) model. In that case one can define the quality as the amount of variance explained in the indicator by its trait (Saris and Gallhofer, 2007b).

However, the assumption of continuous and interval measurement implicit in the CFA model may be false when responses with only a few categories are obtained. In that case the ordinal CFA (oCFA) model of Muthén (1984) is often applied (Scherpenzeel and Saris, 1997). This model is equivalent to Samejima's graded response model (Samejima, 1969) in item response theory. Such an analysis can be accomplished by applying the CFA model to so-called polychoric correlations, or by special software.

The oCFA model takes the discrete and ordinal nature of the responses into account, but at the cost of strong assumptions about the specific form of the relationship between latent and observed variable. In particular, it is assumed that there are continuous latent response variables that have been split up into just a few categories. These latent response variables are assumed to have a normal or logistic distribution, leading to the familiar probit or logit relationship between trait and observed variable (and between method and observed variable). More importantly, the slope parameters of the influence of the trait on the indicator are restricted to be equal for all categories. This implies that the cumulative probabilities of all categories are restricted to be parallel S-shaped curves.

It should be noted that although the normal distribution is a commonly used and computationally convenient choice for the latent response variables, other choices have also been suggested in the literature. Skew-normal (Roscino and Pollice, 2006), copula (Joe, 2005), and mixtures of normal distributions (Uebersax and Grove, 1993) have been suggested. Rost and Walter (2006) applied mixture Rasch models and the LLTM to MTMM data. The alternative approach of optimal scaling should also be mentioned (Takane et al., 1977). These approaches do relax the assumption of normality, but express all relationships only in terms of the latent response variables, which does not allow for a full analysis of the relationship between the traits and observed variables we are interested in.

In this paper we will elaborate on a different approach: the latent class factor model (Vermunt et al., 2004). This model (the LCM) derives from the latent structure model formulated by Lazarsfeld and Henry (1968). Goodman (1974) further developed the latent structure model and gave a method for maximum

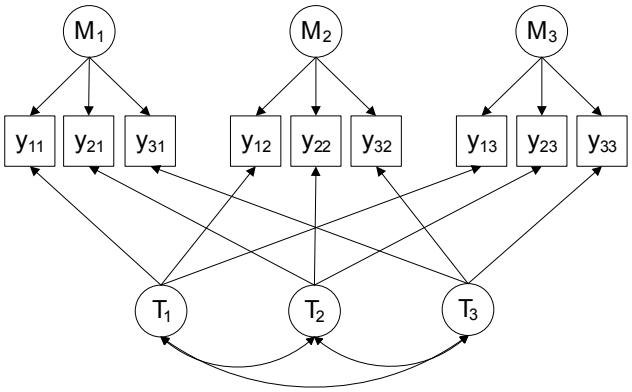


Figure 4: A model graph for multitrait-multimethod data. The method factors (M) represents different answering strategies used by the respondents that may be similar across questions. The trait factors (T) represent the opinion of the respondent after correcting for idiosyncratic response sets and random measurement error. Random error components for each observed variable are not shown here for clarity but can be imagined.

likelihood estimation of the parameters. Haberman (1979) provided the parameterization in terms of log-linear coefficients used here, and was the first to suggest different restrictions on these coefficients yield models for different measurement levels of the observed variables. Different applications of other variants of these models are discussed by Hagenaars and McCutcheon (2002).

The LCM specifies the following relationship between trait, method, and observed variable:

$$p(y_{ij} = k|T_i, M_j) := \frac{\exp(a_{ijk} + b_{ik}^{(t)} T_i + b_{jk}^{(m)} M_j)}{\sum_{l=1}^{K} \exp(a_{ijl} + b_{il}^{(t)} T_i + b_{jl}^{(m)} M_j)}; k, l \in \{1, ..., K\}, \quad (1)$$

where K is the number of categories for the observed variable. The latent variables (traits and methods) are scaled to have equal-distance values lying between 0 and 1. Thus a trait with 5 categories will have scores $\{0, 0.25, 0.50, 0.75, 1\}$ for category numbers 1 through 5. The log-linear parameters a and b are set to sum to zero over all categories of y. This is an arbitrary restriction necessary for identification. By $b_{ik}^{(t)}$ we mean the slope for trait number i and category k, whereas the $b^{(m)}$ are slopes for the method.

The b_k parameters in the model of equation 1 are the associations or log-linear effects and the a_k parameters are intercepts for each category. The effects b_k differ by category. Each effect can also be written as $b_k = bk$, meaning there is only one slope b for the effect of the latent variable on the observed one, and k is the category score. The category scores can be restricted to increase by a certain number for each category, or they can be freely estimated. By different restrictions a different assumption about the measurement level of the observed variable results.

If the observed variable is assumed to be have interval level, the category scores can be assumed to be of equal distance, in general increasing by unity. A common choice is to use the scores 1, 2, 3, 4, 5 for the first, second, third, etc. categories. The effects b_k in equation 1 then become $b, 2b, 3b, 4b, 5b$ for each category k. This model is also known as the uniform association model, since the local odds ratios of adjacent rows and columns in the cross-table of the latent and observed variables have the same value everywhere, namely $\exp(b)$ (Agresti, 2003, pp. 369-370). In item response theory this formulation is equivalent to the partial credit model (Thissen and Steinberg, 1986).

The category scores k can also be estimated by the model. In this case the scores can take on any value, including values that do not increase or decrease monotonically with the category number. In this case the observed variable has nominal measurement level. In practice often the scores do increase monotonically with the category number, yielding an ordinal measurement level. This ordinality is, however, not a restriction imposed by the model but may or may not be found in practice. Since the effects b_k equal bk, one cannot determine whether the differences in b_k stem from different categories or different slope (or both). Thus for this model we will only report their combination b_k. The model where the observed variables have nominal or ordinal measurement level is also known as the row (or column) association model (Agresti, 2003, pp.

	Latent variable	
	Nominal	Interval
Observed variable		
Nominal	(*Classical LCA*)	Row/column association model
Ordinal	(*Classical LCA with constraints*)	(*RC association with constraints*)
Interval	(*Row/column association model*)	Uniform association model

Table 1: Different measurement levels for latent and observed variables can be accomodated within the latent class model (Heinen, 1996). We consider only the models where the latent variable is interval and the observed variable either nominal or interval. Other possible models are indicated in brackets.

373-4). It can also be described as a latent class version of the nominal response model from item response theory (Bock, 1972).

The LCM can thus accommodate different measurement levels of the latent and observed variables (see table 1; see Heinen (1996) for further discussion). Both latent and observed variables are always regarded as discrete, but one can impose the restriction of interval measurement on latent and/or observed variables. We use this possibility to examine whether the responses can be taken to have been measured at interval level or not, and whether the assumption of ordinal categories is warranted. In this table 'classical LCA' indicates the model where no restrictions are placed on the pairwise loglinear parameters.

Again the quality and amount of systematic error of the observed variable can be defined in terms of the relationship between the trait and method variables and the observed variable. Where in the CFA model the quality is the amount of explained variance, in the LCM the relationship is more complex and depends on the value of the latent trait. It is determined by the log-linear a and b parameters of equation 1, which can be used to express the effect of the trait or method on the observed variable in odds ratios. We can say that the quality of the measure is zero for all values of the trait if the relative odds of choosing any category do not increase or decrease with the trait. This is the case only when all b coefficients are zero. In contrast, a good measure has a high quality for values of the trait that cover as much of its distribution in the studied population as possible. Note that typically very high and very low (and unlikely) values of the trait will still be inaccurately measured, even by a high-quality measure.

The odds ratios aid in understanding the model, but make it more difficult to interpret in terms of probabilities. We will therefore also examine the probability of each category given the trait (item category characteristic curves). We will further evaluate the quality of the questions by plotting the amount of information that each item provides on its own about its latent trait. The so-called 'item information' is a generalization of the concept of reliability and used in test construction in IRT (Hambleton et al., 1995). In sum, although the relationship between latent variable and indicator is more complex than in the linear CFA model, this relationship can still be examined, and in great detail. In general the LCM provides much more detailed information about the use of the scale than the other models mentioned above.

The LCM approach also has disadvantages. First, in our previous conceptualization, the latent variables were continuous and measurement errors arise partly because answers are obtained only in categories. To put it another way, only an unknown range of values on an underlying continuous variable is observed, and the latent response variables are discretized into the observed variables. In the latent class model this aspect of the errors is not modeled: the latent variables are still discrete. Thus, if the real variable of interest is continuous the latent classes still contain measurement error. Whether the classes of the LCM provide an accurate enough approximation to this continuous distribution is a topic for discussion and research.

A second arguable disadvantage of the LCM approach is that the models have many parameters. In many cases a simpler model might provide a sufficiently good description of the measurement process without the need for 116 parameters. At the same time, this is also the strength of the approach, since it can be employed to investigate with precision the quality of the measures and the need for a specific set of assumptions. We will to some extent try to avoid the tendency toward overparameterized models by employing the Bayesian Information Criterion (BIC), which penalizes extra parameters.

4 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 (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[4], and for the different translations we refer to the ESS archive[5].

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).

We applied the LCM model specified above to the experiment on the role of men and women in society. Three traits were measured in this experiment, namely:

1. "A woman should be prepared to cut down on her paid work for the sake of her family;"
2. "Men should take as much responsibility as women for the home and children;"
3. "When jobs are scarce, men should have more right to a job than women."

These traits were measured using a five category agree-disagree scale in different phrasings of the question for the first two methods, and using an item-specific scale as the third method (see appendix for question formulations). Barplots and descriptive statistics for the variables we will study are given in figure 5.

In order to be able to compare countries on the quality of measurement, we selected the country with the highest and the country with the lowest quality for the questions, as estimated in the confirmatory factor analysis model. In this experiment Greece (n=2406) had the highest qualities and Slovenia (n=1442) the lowest. More information can be found together with the precise quality estimates for all countries in Oberski et al. (2007).

5 Methods

We used the program Latent Gold 4.5[6] (Vermunt and Magidson, 2005) to estimate the following models which result from different assumptions about the relationships in figure 4:

[4] http://www.europeansocialsurvey.org
[5] http://ess.nsd.uib.no
[6] http://www.statisticalinnovations.com/products/latentgold_v4.html

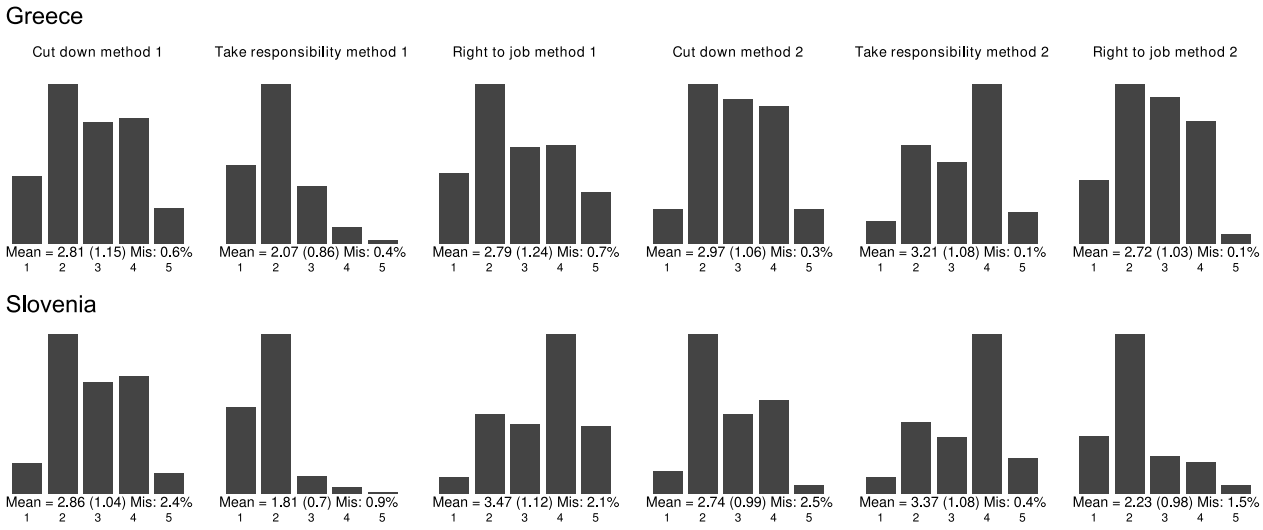


Figure 5: Barplots of the the observed variables for the first two methods of the 'role of women' experiment in two countries. Below each barplot the mean and standard deviation (in brackets) are given with the percentage of item missing data.

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	×	×	×	2	×	×	×
3	×	×	×	3	×	×	×

In all models we take the latent variables to be of interval level measurement, while the observed variables may be interval or nominal. We also investigate models with different numbers of classes, to the extent allowed by the amount of information in the data and the estimation procedure[7]. In order to limit the number of possible models, we vary the number of classes for all traits at the same time and for all methods at the same time. We do not consider models with 5 classes for one trait and 3 for another trait, for instance.

No restrictions are imposed on the associations between the latent traits, except that there are no third-order interactions. This implies that the associations between any two latent traits may be of any form but do not vary across levels of the third trait.

Although not shown here, we also estimated models with no method factors and differing numbers of classes for the traits. In all cases the fit indices indicated a strong need to introduce method factors.

In the analysis, aside from dealing with the planned missing data design that can be considered missing completely at random (MCAR), we also take into account the design weights provided by the ESS, interviewer clustering effects on estimates and standard errors, and data missing at random (MAR). The solutions are obtained by the EM algorithm with at least 10 random starting values in order to find the global optimum, switching to Newton-Raphson at the end of optimization.

6 Results and discussion

6.0.1 Model selection

We estimated the latent class MTMM model described above with different numbers of classes and different assumptions about the measurement level of the observed variables. That is, a so-called linear-by-linear or uniform association model and a row association model. Table 2 shows the resulting BIC model selection criteria and selected models. Lower numbers indicate a better fit to the data. Note that models with differing

[7]For example, estimating the nominal model with 5 trait classes and 3 method classes for Greece took 2.5 days on our computer.

Greece

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	31438	31209	31017	2	30922	30595	***30478***
3	31083	30852	30880	3	.	30502	30498

Slovenia

Observed variable measurement level							
Interval				Nominal			
	Traits (no. classes)				Traits (no. classes)		
Methods	3	4	5	Methods	3	4	5
2	17417	17342	17335	2	.	**16149**	*16160*
3	17427	17332	.	3	.	16158	.

Table 2: BIC for the different models estimated on the 'role of women' experiment. The model selected by the BIC is shown in **bold face**. The model selected by the AIC (not shown) is shown in *italics*. For Greece BIC and AIC select the same model.

number of classes are not nested and cannot be compared using a likelihood ratio test. For such comparisons the BIC can be used (Raftery, 1995). The selected model according to the AIC criterion is also indicated.

In both countries, the BIC and AIC indicate that models including method factors fit the data much better than models without method factors. Therefore the criteria indicate that method factors must be introduced. Also for both countries, the observed variables cannot be taken to be measured at interval level: a model with nominal (or ordinal) level observed variables fits the data much better. This brings into question the assumption of interval level measurements made by the confirmatory factor analysis model. The degree of the difference between the equal and unequal interval models can be deduced from the parameter estimates discussed later.

In Greece the AIC and BIC select the same model, which has 5 classes for the traits and 2 classes for the method factors. The observed variables are measured at nominal level in this model. The model has 2257 degrees of freedom. In Slovenia the AIC selects this same model (1246 degrees of freedom), while the BIC selects the more parsimonious 4-class solution for the traits (1249 df). In the interest of being able to compare the two countries, we will select the same solution for both countries, choosing the model with 5 classes for the traits and 2 for the methods for both Greece and Slovenia.

6.0.2 Quality of the questions

Parameter estimates The results for the selected model for Greece and Slovenia are shown in table 3. This table only shows the parameter estimates for the questions measured using the first method (i.e. the questions asked in the main questionnaire of the ESS). The estimates for the other six questions can be found in the appendix.

The model selected has a separate parameter for each category of the observed variable. This parameter can be seen as a varying the effect of the trait on the observed variable (see equation 1). The parameters can be interpreted in terms of odds ratios: if the latent trait increases by one category, the odds of choosing category 2 over category 1 of the first item in Greece, for instance, increase by 20. This is so because a one category increase of the latent trait is scored as 0.25 and $0.25(e^{-17.4}/e^{-21.8}) \approx 20$. So for each one-category increase in the trait the odds of choosing category two rather than one increase 20-fold.

The model does not restrict the items to be of ordinal measurement level. Ordinality *may* hold, however. An item is ordinal if the estimated log-linear effects (b) of the trait on the observed variable are all increasing (or all decreasing) numbers. The table therefore shows that ordinality holds for all observed variables shown here except for the 'Take responsibility' item in Slovenia, although the difference between the offending coefficients is not statistically significant. This item has an exceptionally low measurement quality in all analyses we have performed.

The same effects are also very unevenly spaced. In some cases, such as categories three and four ('neither

		Greece				Slovenia			
		Traits b_{kt}	s.e.	Method b_{km}	s.e.	Traits b_{kt}	s.e.	Method b_{km}	s.e.
Cut down									
Agree strongly	1	-21.82	(4.02)	2.76	(1.51)	-6.77	(2.30)	3.33	(0.89)
Agree	2	-17.40	(3.49)	-1.22	(1.39)	-5.85	(1.32)	-1.92	(0.67)
Neither agree nor disagree	3	-5.14	(4.13)	-1.58	(1.14)	0.22	(1.03)	-2.73	(0.66)
Disagree	4	17.78	(2.58)	-2.37	(0.44)	4.53	(1.15)	-2.04	(0.53)
Disagree strongly	5	26.59	(10.36)	2.40	(3.78)	7.87	(3.18)	3.36	(1.34)
Take responsibility									
Agree strongly	1	-29.17	(4.36)	1.37	(0.70)	5.34	(1.87)	0.88	(0.54)
Agree	2	-13.05	(1.99)	-2.24	(0.57)	0.68	(2.08)	-3.94	(0.40)
Neither agree nor disagree	3	12.76	(1.78)	-1.11	(0.43)	-2.64	(1.74)	-3.49	(0.87)
Disagree	4	12.65	(1.94)	-1.06	(0.61)	-6.25	(5.69)	-1.97	(0.59)
Disagree strongly	5	16.80	(4.21)	3.03	(1.18)	2.86	(3.18)	8.52	(0.93)
Right to job									
Agree strongly	1	-25.33	(9.97)	0.06	(2.00)	-18.65	(4.68)	3.40	(0.93)
Agree	2	-20.17	(2.77)	-2.73	(0.46)	-14.11	(2.64)	-0.71	(0.57)
Neither agree nor disagree	3	-7.91	(3.10)	-3.45	(1.13)	1.77	(2.30)	-1.42	(0.33)
Disagree	4	11.05	(4.16)	-2.69	(0.97)	10.94	(2.50)	-2.72	(0.62)
Disagree strongly	5	42.37	(6.33)	8.81	(1.21)	20.05	(2.55)	1.45	(0.82)

Table 3: Estimates of the log-linear effects of the traits on their respective observed variables (column 2 for Greece and 6 for Slovenia, with robust standard errors in columns 3 and 7), and of the relationships between the method factors and the observed variables (columns 4 and 8). For the sake of brevity only the three questions from the main questionnaire (method 1) are shown.

agree nor disagree' and 'disagree') for the 'Take responsibility' item in Greece, they are almost equal for two different categories. This suggests that these categories represent much the same opinion and that therefore these items can not be taken to be of interval measurement level.

Turning to the effects of the method factors, it can be seen that these represent an 'extreme versus middle response' factor. Take, for example, the first item in Greece (column three in the table). If a person were to go from class 1 to class 2 on the method factor, their odds of choosing 'agree completely' rather than 'agree' increases about 50-fold, *keeping the trait score constant*. At the same time, their odds of choosing 'disagree completely' rather than just 'disagree' increase about 100-fold. Considering that disagreeing with the statement is the obviously socially desirable answer, higher scores of the method factor are associated with answers that are extreme, but more so on the socially desirable side. This finding holds for all items and methods, including the ones not shown here.

The table shows also that the parameters have been estimated with considerable uncertainty. This uncertainty includes sampling design and interviewer effects, since we have included these in the model estimation procedure. In spite of the large standard errors, most coefficients have been estimated with sufficient precision to distinguish between the parameter values. The highest uncertainty is associated with categories one and five, which were chosen much less often than the other three categories. The lack of data points for these categories, which, as we shall see, is a consequence of the poor quality of some of these items, aggravates the uncertainty inherent in the analysis.

Item characteristic curves The parameters shown in table 3 can be used to compare the countries on the relationships mentioned. But they do not provide a complete picture of the quality of the indicators. We are primarily interested in the conditional probabilities of belonging to each category of the observed variables given the latent class, and these probabilities are also determined by the intercepts in equation 1. As a clarification, one can consider that in an analysis with two classes and two observed categories the conditional probabilities would be the true positives and false negatives rate. To shed more light on the precise relationships the traits have with the indicators, therefore, we also provide plots of the so-called item characteristic curves[8]. These are sometimes also called 'item-category response functions'.

[8]Note that here we show the conditional probabilities rather than the cumulative probability often graphed.

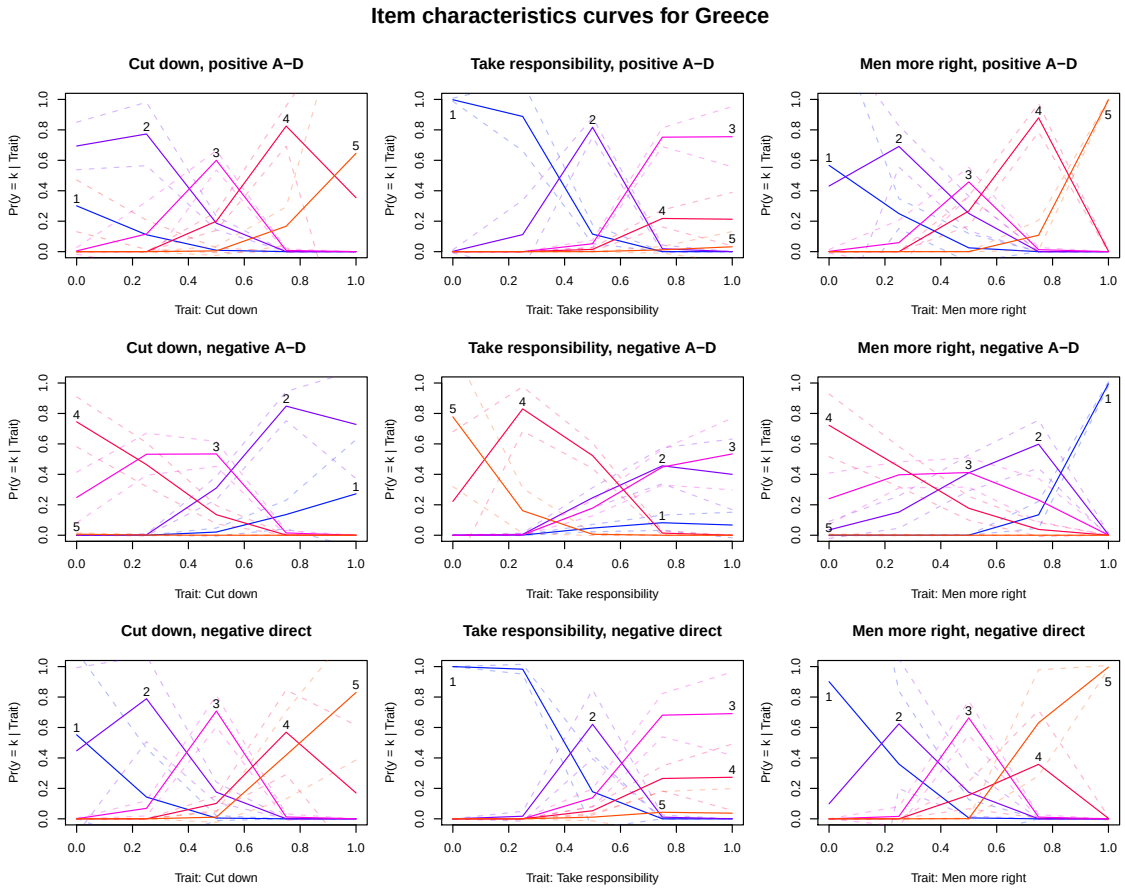


Figure 6: Item characteristic curves for Greece. The lines indicate the probability of choosing a category, given that value for the trait to be measured. Each line is marked with its category number at the point where this probability is highest (its peak). The dotted lines are approximate 95% confidence intervals around the probabilities. The columns show measurements of the three different traits, while the rows show measurements using the three different methods.

Figure 6 provides the curves describing the conditional probability of belonging to each category, given the score on the latent trait the variable is supposed to measure.

In the figures, the three methods of asking the question correspond to the rows, so that the first row contains the three graphs of the ICC's for the first method (main questionnaire), and the second and third rows the graphs for the supplementary questionnaires. The columns and graph titles correspond to the three traits described above.

The solid lines in the graph correspond to the probability of choosing a category, given the trait score. The lines have been marked with a color and the number of the category. This category number is moreover plotted at the point where the conditional probability of choosing that category is highest, i.e. at the peak of the item characteristic curve. If an item is ordinal then the ICC's peak in succession, and one will read either '1 2 3 4' or the reverse from left to right. It is also of interest whether the peak is high (close to one) or not, as this is an indication of the specificity of the category. Last, the peak should ideally not be underneath another curve.

The dotted lines provide approximate 95% confidence intervals around the ICC's. This is an example of the richness of the output that can be obtained from Latent Gold. The uncertainty noted in table 3, which is considerable for the extreme categories, is again reflected here.

The top left graph for Greece shows that item 'cut down' from the main questionnaire has very good measurement properties in this country. The category curves peak in succession, meaning that all categories provide information about the score on the latent trait. Moreover these peaks, which can be likened to the

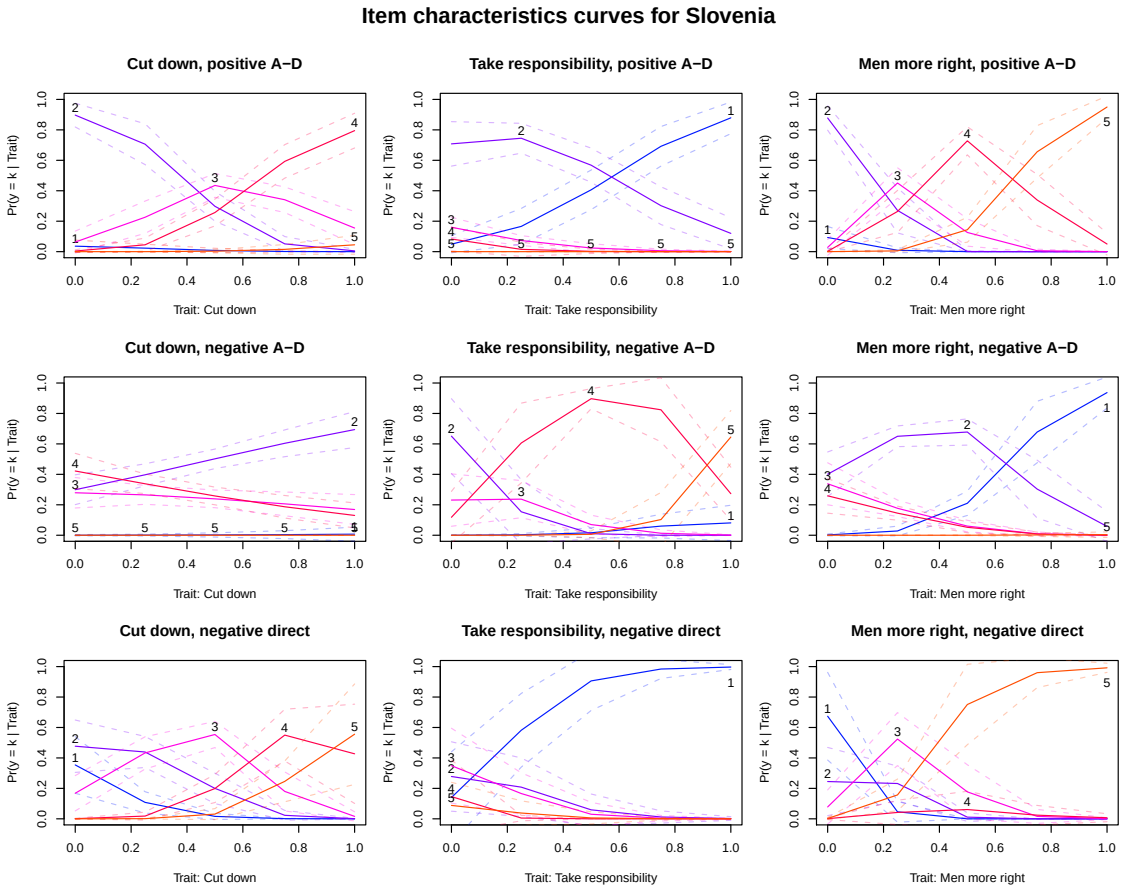


Figure 7: Item characteristic curves for Slovenia. The lines indicate the probability of choosing a category, given that value for the trait to be measured. Each line is marked with its category number at the point where this probability is highest (its peak). The dotted lines are approximate 95% confidence intervals around the probabilities. The columns show measurements of the three different traits, while the rows show measurements using the three different methods.

probability of true positives or sensitivity in two category models, are quite high, in the $Pr(y = k|T_1 = k, M = E(M)) = 0.8$ range, except for the first category. Since all the curves are steep, the probability of choosing any other category than the modal one–false negatives or (un)specificity–decreases sharply. This graph is highly similar to the same graph as calculated from the probit IRT model by the program Mplus (Muthén and Muthén, 1998) based on our previous research. Thus for this indicator the probit IRT or categorical factor analysis model may describe the relationship between trait and indicator adequately.

The same graph for Slovenia (figure 7) is quite different. Here it is clear information is only being obtained from the three middle categories. The extreme categories are hardly used at all. For these middle categories, however, the peaks are successive and relatively high for categories 2 and 4. Thus, although not all categories are used, resulting in a loss of information, the discriminating power of the three middle categories is quite good.

This is not the case for the same item measured by the second method in Slovenia. Here the quality is extremely low, as almost no discriminating power exists except for choosing the second category versus all the others. In general the measurement quality for the second method is much worse in both countries. The third method fares better in Greece then it does in Slovenia, where the measurement quality is disastrous; in the second item only choosing the first versus all the other categories provides any information.

It can also be seen that in general the measurement properties in Slovenia are worse than in Greece. This is in line with the findings from CFA and ordinal probit models; in fact, it was the reason these two countries were selected. As an example one can compare item 2 'take responsibility' in the main questionnaire across

the countries. In Greece again only the three middle categories provide good measurement properties. Thus this item has intermediate quality. But in Slovenia the middle category is equally likely to be chosen for all values of the trait, and in fact just as likely as categories 4 and 5. The only differentiation one can make between people on the latent trait comes from a distinction between categories one and two[9]. This item has an extremely low quality in all models we have examined for these data so far; in the continuous CFA model the percentage of variance explained in the item by the trait was estimated at 25%. An explanation is now found in the extremely limited use of the scale. In the estimated marginal distribution (prevalence) of this opinion in Slovenia the proportion of people in categories of the latent trait associated with disagreement is below 0.10.

When the comparisons described above are made across the methods it can be seen in figures 6 and 7 that the second row consistently has worse measurement properties than the other two rows. The first and third methods have comparable measurement quality. The same conclusion was also drawn in the categorical CFA analyses that were conducted earlier. The linear CFA analysis suggested that the first method was slightly better than the other two.

Item information A more direct measure of the quality than has been used so far is the item 'information'. It is the inverse of the error variance of the maximum likelihood estimate of the trait that one can get from each item, and can be seen as a generalized reliability. The information function $I(T)$ is a measure of precision in the estimation of the trait T: $\sigma(\hat{T}) = 1/\sqrt{I(T)}$. Thus, as the curve approaches zero, less and less can be said about the person's trait score. The item information functions are shown for all items in figure 8. For more details about the information function and how it was computed we refer to the appendix.

Because of the non-linear specification of the model, and contrary to CFA, the information varies across levels of the trait[10]. Instead of a single number, a plot is obtained across the range of the trait. One can also obtain the marginal or average information in a particular country by averaging over categories of the trait, weighting the information at each category by the prevalence of that category (e.g. Donoghue, 1994). This average information is a single number that provides the expected information for that country. It is important to note, however, that it depends on the marginal distribution of the trait: items in two countries with the same information curve but different marginal distributions will in general provide different average information.

In the figure the information has been plotted on a log scale to allow for comparison of the different items, which vary widely in information provided. Therefore any visible differences in height of the curves are usually substantially large. One can appreciate the absolute values of the curves by considering for example that an information value of 74 (the average for the direct version of item 1 in Greece) implies that the best estimate of the latent trait for a particular person that one can obtain with this item will have a standard deviation of 0.12, on a scale of 0 to 1. One can also compute the relative efficiency of two items as the ratio of their information (Hambleton et al., 1995). The average information in the country has been indicated at the top of the graphs.

The agree-disagree versions of the 'cut down' item in Greece are clearly asymmetrical. This implies that opinions against the 'feminist direction' are measured much less accurately than 'pro-feminist' opinions. The item-specific scale is much better overall and also provides better coverage of the entire range of opinions. It is for the population studied slightly (1.1 times) more efficient than the first method and 1.7 times as efficient as the second method.

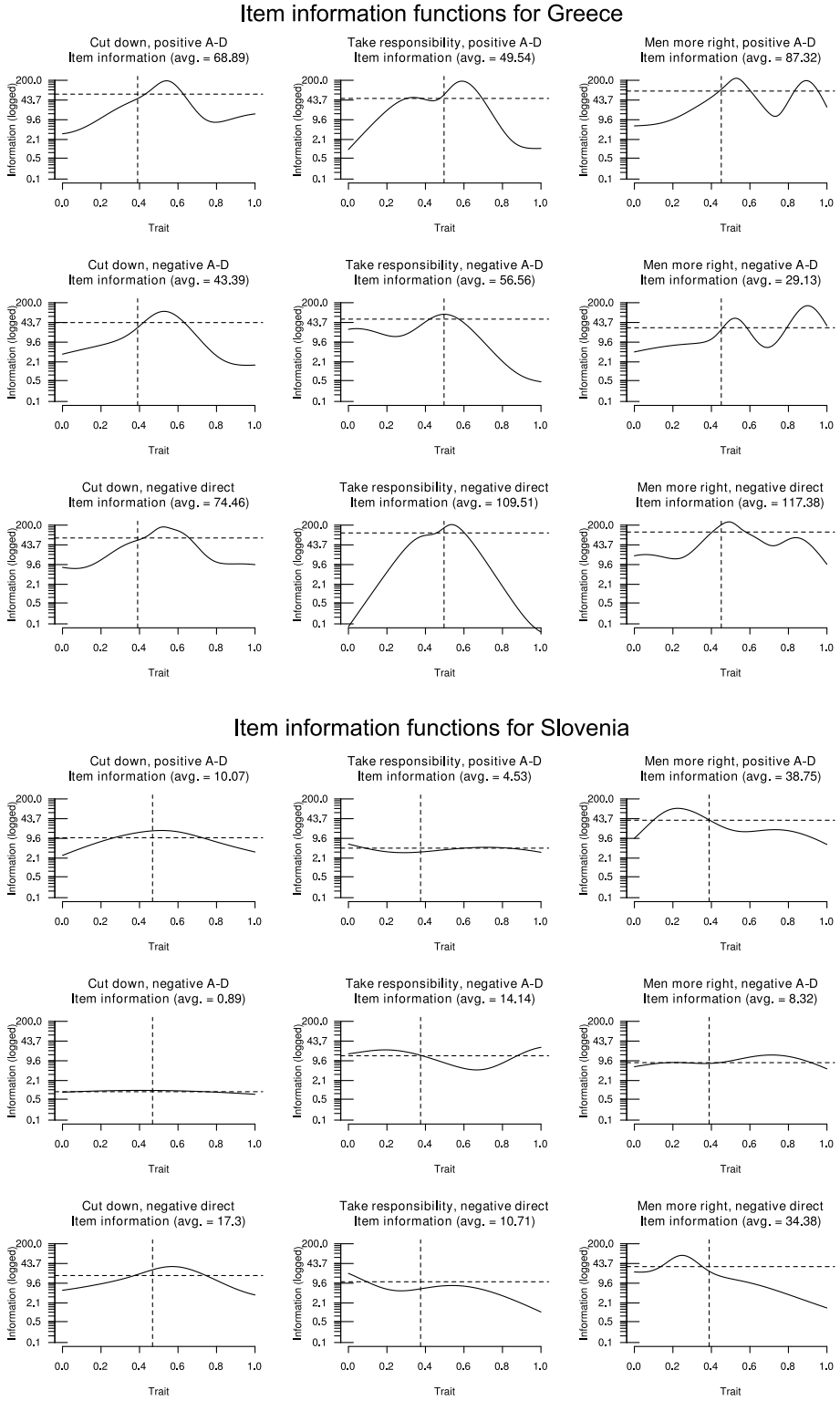
The direct version of the 'take responsibility' item has a very high peak and provides much more information about opinions close to the average Greek opinion than the agree-disagree scales. However, away from the average the information provided is much higher for the first two methods. In principle these are therefore better adjusted to measure relatively 'feminist' or 'anti-feminist' opinions than the item-specific scale, where 'feminist' opinions are again better measured than 'anti-feminist' ones. On average the item-specific scale is still about twice as efficient as the agree-disagree scales for the population studied.

The 'men more right' item has high quality overall, and covers the whole range of opinions quite well. In this respect the item-specific scale again does much better than the agree-disagree scales, whose information curves are skewed towards the measurement of 'feminist' opinions. On average the item-specific scale is 1.3

[9]Note that the scale of this latent trait has been reversed relative to Greece. This ordering of the classes is arbitrary and does not affect the results.

[10]A complication omitted here is that it also varies across levels of the method. The curves shown provide the marginal information collapsed over categories of the method. See the appendix for an explanation.

Figure 8: Model-based item information functions for both Greece and Slovenia. Note the log scales.



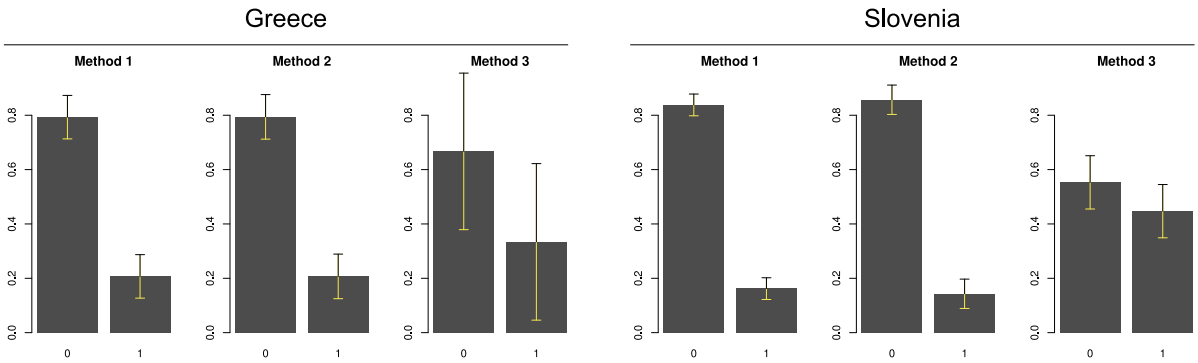


Figure 9: Estimated histograms of the latent method factors with approximate 95% error bars.

times more efficient than the positive agree-disagree version, and much (4 times) more efficient than the negative agree-disagree version.

The graphs for Slovenia immediately reveal the much lower overall measurement quality of the items in that country. The median item in Greece is 4.3 more efficient on average than in Slovenia. This is more than the largest information ratio in Greece. Thus the differences between the countries are much larger than the differences between the items within each country.

'Men more right' is also the better item in Slovenia. There the measurement is skewed towards measurement of 'pro-feminist' opinions for both the direct and positive agree-disagree versions. Contrary to the pattern found in Greece that the item-specific scale provided more equal measurement across the whole range of the scale, in Slovenia the item-specific scale's information curve is more skewed than the other two method's curves. The average information for the Slovenian population is not very different, however, reflecting the highly skewed marginal distribution of the trait in that country.

The 'take responsibility' trait is not well measured in Slovenia. The negative agree-disagree is slightly (1.3 times) better than the item-specific scale, and seems to be able to pick up also negative opinions somewhat. This is the only item where the negative agree-disagree scale is better than the other two methods.

'Cut down' negative agree-disagree is the worst item of all. It has a variance which is higher than the entire range of the trait scale. This means one knows about as much about a Slovenian's opinion after asking this question as before asking it. The other measures are better, the direct version being the best of the three.

Overall we found that the item-specific scales were better than the agree-disagree versions (3.5 times more efficient on average), and that the positively formulated items were better than the negatively formulated ones (1.8 times on average). This finding is in line with Saris et al. (frth). The difference in quality between the countries that motivated the choice of countries in the first place was also clearly found.

6.0.3 Method effects

So far we have only discussed the relationship between the traits to be measured and their observed variables, that is, the quality of the questions as indicators of the trait they are supposed to measure. As discussed previously, another important part of answers to survey questions can be described as 'method effects'. These have been modeled in our case as latent variables that affect the answers in the same way for questions asked in the same way, but are unrelated to the traits to be measured.

The coefficient estimates for these method factors are shown in table 3 and were already discussed. The method factors found in both countries represent a contrast between only using the middle categories or giving extreme answers, more so on the socially desirable side. The fact that class 1 on the method factor is more associated with disagreement is not in line with the hypothesis that respondents tend to 'acquiesce', that is, to tend to agree with any statement.

The estimated proportion of people (with approximate 95% confidence intervals) in each of the two categories of the method factors is shown in figure 9. The majority of people are in the class which uses only the middle categories. But a substantial proportion of people also are extreme responders. For the third method there are more extreme responders, with the difference in proportions of extreme and middle responders not statistically significant. This may be a consequence of the fact that the first two methods

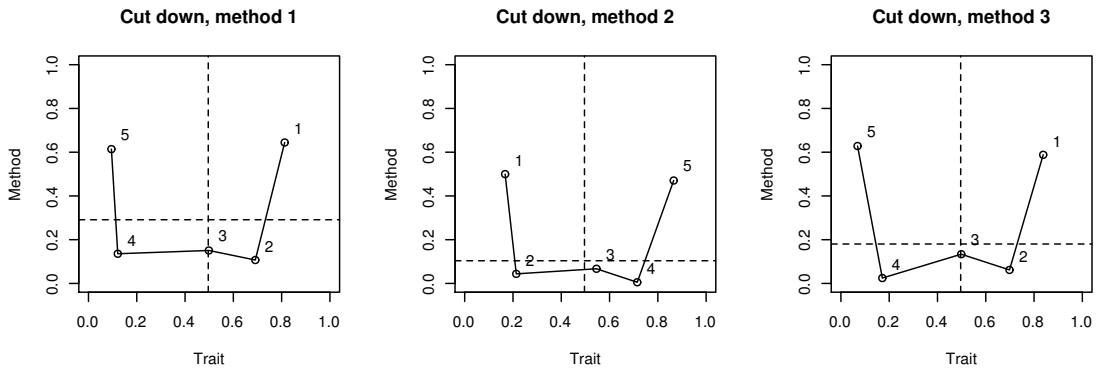


Figure 10: Bi-plots for the first item, 'Women should be prepared to cut down on their paid work', for all three methods in Greece. Plotted is the conditional mean of the trait (horizontal axis) and method factors (vertical axis) influencing the item given that a particular category (the points labeled with a category number) was chosen.

were fully labeled scales while the third method has only the two extremes labeled, perhaps attracting more responses.

An instructive way of examining the relationship an item's categories have with its trait and method factors is through a bi-plot (Magidson and Vermunt, 2001). Bi-plots for the first item ('women should be prepared to cut down on paid work') for Greece are shown in figure 10. The plots shown in this figure plot the conditional mean of the trait and method factors given a choice for each of the five categories of the items. The plot again makes the meaning of the method factors readily apparent: categories 1 and 5 versus the rest.

When one projects the points (categories) onto the trait axis, it can be seen that the categories are quite unevenly spaced as was already remarked. For the first method categories 5 and 4 are much closer together than categories 1 and 2, suggesting the scale is not symmetric. The same happens in the opposite direction for the second method. For this method choosing the middle category represents an above-average opinion, suggesting the phrase 'neither agree nor disagree' does not have its intended neutral meaning in this case. In all three methods two of the categories are much closer to the middle category than the other two. It can also be seen that the method and trait factors represent very different things as they are unrelated.

External validation of the method factors The model as estimated so far appears to give valid inferences about the items. However, the meaning of the method factors has been assumed rather than checked by using external data. We now do this, demonstrating how one can use the factor score estimates of the latent class MTMM model to perform additional analyses.

It was suggested that the method factors represent an 'extreme response style' (ERS), and, to a much smaller extent, social desirability. This conclusion was based on the coefficient estimates. We now test the same conclusion with data not used in the model. Similar studies have been done using CFA by Billiet and McClendon (2000) and using a latent class factor model by Kieruj and Moors (frth).

One important reason for this is that the literature on ERS suggests that it is a stable personality trait that is different for different people but the same across all questions for each person (Billiet and McClendon, 2000). In the MTMM model developed above, however, the method factors are independent, suggesting that ERS on one set of items does not imply ERS on another. Therefore the correlation with external measures can be seen as a validation of the MTMM model.

We selected 39 variables from the ESS main questionnaire that had an answer scale on which extreme response was possible. In order to prevent confounding of variables, the items on position of women studied here were excluded. A measure of extreme response style was constructed by counting the number of times each respondent chose the most extreme possible categories on the answer scale (the minimum or the maximum). This variable, called 'stylesum', had mean 7.1, median 6 and interquartile range 6.

The method factor scores of the three methods was estimated for each person and added to the data set with the variable 'stylesum'. The modal (most likely) method scores (0 or 1) were also added.

	Method 1–ERS	Method 2–ERS	Method 3–ERS
Pearson correlation with factor scores	-0.222*	-0.027	-0.076*
Polyserial correlation with modal category	-0.261*	-0.069	-0.086*

$*$Significantly different from zero ($p < 0.01$)

Table 4: Correlations between the method factor scores and the external extreme response style (ERS) measure. This measure is constructed as the number of extreme responses on 39 other questions.

We then computed the correlation between the method factor scores and the extreme response style (ERS) measure that was computed completely independently of the 'role of women' variables. We also computed the polyserial correlation between the modal category of the method and the ERS measure. The results are shown in table 4.

It can be seen that the first and third method factor scores correlate significantly with the independent ERS measure. Method 1 correlates much higher with this measure than the other two methods. This shows that extreme response style works differently for different items; a person who answers one type of item in an extreme manner does not necessarily do the same for another. The differences in correlation can in part also be explained by the amount of time between the questions; most questions used to measure ERS were asked in the main questionnaire, closer to the questions used to estimate the factor scores of method 1. The factor scores for methods 2 and 3 were estimated from questions asked in the supplementary questionnaire, approximately one hour after the start of the main interview. This suggests that respondents may also change their reponse style during the interview.

Another method of testing the suggestion that extreme response style is a stable personality trait is to correlate it with other stable personality traits. To this end we correlated the method factor scores and modal categories, as well as the ERS 'stylesum' measure, with 26 questions from the Schwartz 'human values scale' asked in the supplementary questionnaire of the ESS (Schwartz, 1992). These correlations were very small; we found none above 0.1.

From the significant correlation of -0.26 above we can conclude that to some extent the method factors do indeed measure a response style independent of the content of the questions. However, the low correlations with other methods, and of the methods with a person's values, it would appear that the this response style is not a stable personality trait but can vary across methods and even during the interview. Thus the model with separate method factors used here appears more warranted than a model with one style factor that represents a personality trait of extreme response tendency. It should also be noted that it would be very difficult to model extreme response using a traditional or ordinal factor analysis model.

It was clearly shown that the method factors represent a middle versus extreme response. It was also suggested that, to a much smaller extent, they represent susceptibility to socially desirable answers to some extent. So far this claim does not rest on much more than the fact that one of the positive log-linear coefficients for each method factor is larger than the other one, and this happens on the socially desirable side. However, it can also be validated directly.

In the European Social Survey besides the data from the main and supplementary questionnaires data was also gathered in interviewer questionnaires. These included a question for the interviewer on whether 'anybody [was] present, who interfered with the interview'. If the method factor truly represents social desirability in part, then the probability of belonging to the classes should be influenced by the presence or absence of another person during the interview. It is not completely that simple, however; given the content of the questions men and women should show opposite behavior depending on whether their partner is present. Also, presumably, religion plays a role in what is considered desirable.

When each of the items are regressed on explanatory variables and the presence of another person during the interview, it is clear that this variable has a statistically significant influence. For women the expected mean of the 5-point scale increases by 0.5 when their partner is not present compared to when they are. For men this effect is in the opposite direction but much smaller[11]. Thus it is clear that social desirability effects are present in the items. It remains to be investigated, however, whether the method factors estimated in our analysis account for the effect of social desirability.

For this reason we created a data set that combines the original variables from the ESS main, supplemen-

[11]The model controls for age, gender, education, religion, living with a partner, and marital status. The analysis is not shown here but can be obtained upon request from the first author.

	Method 1			Method 2			Method 3		
	Estimate	S.E.	t	Estimate	S.E.	t	Estimate	S.E.	t
(Intercept)	1.58	0.24	6.65	3.15	0.47	6.76	2.37	0.29	8.19
Other person present	-0.04	0.51	-0.08	-0.07	1.03	-0.07	1.13	0.81	1.39
Female	-0.12	0.13	-0.93	-0.63	0.30	-2.13	0.24	0.17	1.41
Religion	-0.14	0.22	-0.61	0.21	0.42	0.49	0.37	0.26	1.41
Church attendance	0.10	0.05	1.94	0.26	0.11	2.35	-0.12	0.06	-1.96
Married	0.01	0.13	0.07	-0.21	0.28	-0.76	-0.26	0.17	-1.53
Present×Female	0.00	0.28	0.00	-0.60	0.73	-0.83	-0.86	0.34	-2.53
Present×Religion	-0.17	0.48	-0.36	1.49	0.80	1.87	-1.53	0.78	-1.96
Present×Married	0.29	0.29	0.99	-0.57	0.85	-0.67	0.69	0.35	1.99
N	2401			2401			2401		
$Deviance$	2118			608			1460		
$-2LLR(Model\chi^2)$	6.52			18.03*			24.52*		

Table 5: Logistic regression of the probability of belonging to the first class on each method as influenced by the presence or absence of a third person during the interview, mediated by different variables that influence what is socially desirable: gender, religion, and marital status.

tary, and interviewer questionnaires with the factor scores (modal classes and probabilities) obtained from our LCM analysis. We then regressed the logit of the probability of belonging to the first class of each method factor on presence of another person, as well as gender, religion, marital status, and their interactions with the presence or absence of another person during the interview. The results of this analysis for the three method factors for Greece are shown in table 5.

It can be seen in the table that the effects for the first method are all non-significant. For the second method there are main effects of church attendance and gender, and the interaction effects, though not statistically significant, are in the expected direction. For the third method the model is clearest. All of the interaction effects as well as the main effect of church attendance are statistically significant.

To give an example of the meaning of the above analysis, consider the estimates for the third method. For a woman who is religious, the probability of moving into class 1 of the method factor increases from 0.80 to 0.93 if somebody is present at the interview. This in turn increases her chances of saying that 'a woman should be prepared to cut down on her paid work', for instance. Incidentally it also decreases the chance that she will use an extreme category considerably. All these effects happen, in our model, while keeping her trait score constant. Thus any difference in the answers provided by respondents differing on the characteristics in table 5 has nothing to do with a change in their underlying opinion.

It should be noted that the social desirability effects found on the method factors are small relative to the effects found on the items themselves. This suggests that there is an element of social desirability, different across respondents, that still remains to be explained. The model could be expanded to include an acquiescence style factor, for instance (Billiet and McClendon, 2000). However, it is questionable whether a model with such an extra latent variable can be estimated with the experimental design used here. This remains a topic for further investigation.

7 Conclusion

The goal of this study was to show how more general measurement models can be formulated, and in particular to demonstrate the use of latent class models for analysis of the quality of single questions.

We have formulated latent class factor models from our general graphical model and applied these models to a multitrait-multimethod experiment on the role of women in society. Furthermore we compared the results for two countries, one of which was previously estimated to have low question qualities (Slovenia) and the other (Greece) high ones.

We investigated the quality of the questions using the item characteristic curves and information functions. To our knowledge this paper is the first to provide formulas for the item information function of latent class factor models (see appendix).

The investigation of question quality using the LCM yielded a wealth of information about the functioning

of the questions. It was established that for the agree-disagree scales only the middle three categories (out of five) provide information about the traits to be measured, and that for the items with exceptionally low quality the number of categories providing information is reduced even more.

The quality in Slovenia was again found to be much lower than in Greece, in line with previous findings. It was also clear that the positively worded items were better than the negatively worded ones, and that the item-specific scales provided much more information than the agree-disagree format. With two exceptions, they also provided more equal information across the whole range of the trait. The agree-disgree versions provide more accurate measurement of 'pro-feminist' than of 'anti-feminist' opinions.

This finding is important: items with an approximately equal amount of information across the range of the trait are desirable, especially in cross-national research.

An item with much skew in its information function is less likely to be useful for cross-national comparisons. This is so because even if the information functions were the same in all countries, countries with higher average opinion would have a higher measurement quality[12]. Since measurement errors affect the analysis of means and regression (Fuller, 1987), differential measurement errors across countries will invalidate comparisons of means and relationships.

In the present analysis we found that the information functions for Slovenia and Greece were very different. Thus it is not clear that the lower quality in Slovenia is due to a difference in the average opinion. The analysis shows that the difficulty of categories indicating 'anti-feminist' opinions was far higher in Slovenia. Thus most Slovenians are left with only two choices, 'agree' versus 'neither agree nor disagree', which are used differently by different people as evidenced by the method effects. Clearly this item does not measure opinions in a way equivalent to the way they are measured in Greece. If answers are to be compared or the items in Slovenia to be analyzed, therefore, improvements should be made. One suggestion would be to rephrase the question so that it is less extreme.

We examined the method effects. Bi-plots showed clearly what the method factors represent: a distinction between extreme versus middle responses. Most people (about 80%) were found to use only the middle categories. This is a strong indication of satisficing; the question might not be clear enough or too cognitively difficult to answer. The parameter estimates suggested that the method factors also represent a susceptibility to answering in a socially desirable way. This was investigated by regressing the estimated method factor scores onto the presence or absence of a third person during the interview. Effects were found for the third and second methods, but not the first. Considering that the effects of this variable on the items are much larger precisely for the first method, there may still be room in the model for a social desirability or style factor. Such an study would also shed light on the plausibility of the assumption we have had to make that all systematic errors are specific to the method of asking the question. This is, however, is outside the scope of the present study.

The latent class analysis elaborated in this paper provides much information about the precise workings of the items, as well as suggestions for their improvement. Furthermore this was achieved without any assumption of normality or of parallel probability curves. Indeed these assumptions, made in (ordinal) confirmatory factor analysis models usually applied to MTMM data, were found not to hold. Therefore we hope to have shown the utility of this approach for the evaluation of categorical items using multitrait-multimethod designs.

[12] Skew in the amount of information will also bias regression analyses with interactions. As an example, consider the 'take responsibility' item's information function in Greece. A regression of a dependent variable on 'take responsibility', a third variable, and their interaction is formulated. Figure 8 shows that agreement is measured accurately, while disagreement is not. Thus two groups of Greeks which have a different average opinion on the trait will have different amounts of measurement error in this item and therefore different correlations with other variables. Therefore a regression analysis which includes both a main effect and an interaction with the opinion on this item will give biased estimates.

References

Agresti, A. (2003). *Categorical data analysis*. Wiley-Interscience.

Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley-Interscience.

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

Biemer, P. P. and Bushery, J. M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26(2):139–152.

Billiet, J. B. and McClendon, M. K. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4):608–628.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

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

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.

Crowne, D. P. and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349–354.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, pages 295–311.

Fuller, W. A. (1987). *Measurement error models*. Wiley New York.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215.

Groves, R. M. (2005). *Survey Errors and Survey Costs*. Wiley-Interscience.

Haberman, S. (1979). Analysis of qualitative data.

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. and McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge University Press Cambridge United Kingdom:.

Hambleton, R. K., Rogers, H. J., and Swaminathan, H. (1995). *Fundamentals of item response theory*. Sage Publ.

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

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage Thousand Oaks.

Hui, H. C. and Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20(3):296–309.

Jackson, D. N. and Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55(4):243–252.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Kieruj, N. D. and Moors, G. B. (frth). Response scales' vulnerability to acquiescence and extreme response style behavior.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):pp213–236.

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

Lord, F. M. and Novick, M. R. (1968). Statistical theories of mental scores. *Reading, Addison–Wesley*.

Magidson, J. and Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, pages 223–264.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17(4):351.

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

Muthén, L. K. and Muthén, B. O. (1998). Mplus user's guide. *Los Angeles: Muthén & Muthén*, 2004.

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.

Olson, K. and Kennedy, C. (2006). Examination of the relationship between nonresponse and measurement error in a validation study of alumni. *Proceedings of the Survey Research Methods Section*.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Roscino, A. and Pollice, A. (2006). A generalization of the polychoric correlation coefficient. In Sergio Zani, A. C. M. R. and Vichi, M., editors, *Data Analysis, Classification and the Forward Search*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin Heidelberg.

Rost, J. and Walter, O. (2006). Multimethod item response theory. In Eid, M. and Diener, E., editors, *Handbook of Multimethod Measurement in Psychology*. American Psychological Association, Washington, DC.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

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. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S., editors, *Measurement errors in surveys*, pages 575–599. John Wiley & Sons, New York.

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

Saris, W. E. and Gallhofer, I. N. (2007b). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience.

Saris, W. E., Révilla, M., Krosnick, J. A., and Schaeffer, E. M. (frth). Comparing questions with agree/disagree response options to questions with item-specific response options.

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.

Scherpenzeel, A. C. and Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, 25:341.

Schmitt, N. and Stults, D. M. (1986). Methodology review: Analysis of Multitrait-Multimethod matrices. *Applied Psychological Measurement*, 10(1):1–22.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology, vol*, 25.

Takane, Y., Young, F., and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7–67.

Thissen, D. and Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4):567–577.

Uebersax, J. S. and Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49:823–835.

van der Veld, W. (2006). *The survey response dissected: A new theory about the survey response proces*. University of Amsterdam, Amsterdam.

Vermunt, J. K. and Magidson, J. (2005). Latent gold 4.5 user's guide. *Belmont, MA: Statistical Innovations Inc*.

Vermunt, J. K., Magidson, J., and Inc, S. I. (2004). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. *New developments in categorical data analysis for the social and behavioral sciences*, pages 41–63.

A Phrasing of the questions used in the 'role of women' experiment

Main questionnaire (method 1)

CARD 59 I am now going to read out some statements about men and women and their place[81] in the family. Using this card, please tell me how much you agree or disagree with the following statements.

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

Suppplementary questionnaire first (method 2)

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

iS8[26] "A women should not have to cut down on her paid work for the sake of her family."
Please tick one box.

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

iS9[27] "Women should take more responsibility for the home and children than men."
Please tick one box.

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

iS10[28] "When jobs are scarce, women should have the same right to a job as men."
Please tick one box.

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

Suppplementary questionnaire second (method 3)

iS22[57] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
A woman should be prepared to cut down on her paid work for the sake of her family	☐	☐	☐	☐	☐	A woman should not have to cut down on her paid work for the sake of her family

iS23[58] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
Men should take as much responsibility as women for the home and children	☐	☐	☐	☐	☐	Women should take more responsibility for the home and children than men

iS24[59] If you had to choose between the following options which would you prefer? Please show how close your opinion is to the statements below by choosing a number between 1 and 5.
Please tick one box.

	1	2	3	4	5	
When jobs are scarce, men should have more right to a job than women	☐	☐	☐	☐	☐	When jobs are scarce, women should have the same right to a job as men

B Item information function for the polytomous Latent Class Factor Model

In this section we explain how to obtain the variance of the best estimate of the trait T that one can obtain from an observed item y. This variance is also called the item information, as it equals the Fisher information in the likelihood of the item given only the trait (Hambleton et al., 1995).

The observed variable has a multinomial likelihood:

$$L = \prod_{k=1}^{K} P_k^{U_k},$$

where U_k is an indicator function that equals 1 if $y = k$ and 0 otherwise.

The item-category response function for the model used is

$$p(y = k|T, M) := P_k(T, M) = \frac{\exp(a_k + b_k T + m_k M)}{\sum_{c=1}^{K} \exp(a_c + b_c T + m_c M)}. \tag{2}$$

For succinctness we will refer to $P_k(T, M)$ simply as P_k. This model is very similar to a generalized partial credit model (PCM, see Muraki (1993)). Indeed, if for a given relationship one replaces the category scores for the observed variable in the PCM by the slope for that category and sets the discrimination parameter to unity, identical first and second derivatives result.

The equation above gives the conditional probability of choosing category k, given both the trait and the method. In total there are K categories and item-category response functions. These functions are also called the item characteristic curves.

The item information function (IIF) is now equal to the Fisher information in the item, with respect to the trait:

$$I(T) = -E(\frac{\partial^2 \ln L(T)}{\partial T^2})$$

For any given value of M we can derive the second partial derivative of the item likelihood with respect to T as

$$\frac{\partial^2 \ln L}{\partial T^2} = \sum_{k}^{K} [U_k(\lambda^2 - \nu)],$$

where

$$\lambda = \sum_{k}^{K} [b_k P_k];\ \nu = \sum_{k}^{K} [b_k^2 P_k].$$

A proof can be obtained from the first author upon request. Alternatively, we refer to the derivation in the appendix of Donoghue (1994), replacing in that paper the quantities D and a by 1 and all category scores (k, c) by the slope b_k for that category.

Noting that $E(U_k|T, M) = P_k(T, M)$ and $E(x|T) = \sum_l^L E(x|T, M)p(M = l)$ for any random variable x, we can conclude that the information in the item about T, conditional on M is

$$I(T|M) = \sum_{k}^{K} \beta_k^2 P_k(T, M) - [\sum_{k}^{K} \beta_k P_k(T, M)]^2. \tag{3}$$

and the marginal information then equals

$$I(T) = \sum_{l}^{L} [I(T|M)\ p(M = l)], \tag{4}$$

where the index l runs over all scores of the method factor. In the model selected in this paper $l \in \{0, 1\}$.

One can also calculate a trait score estimate for each person based on the parameters. The standard error of the estimation of this score then equals $1/\sqrt{I(T)}$ (Hambleton et al., 1995).