Jouni Kuha, Irini Moustaki

Non-equivalence of measurement in latent variable modeling of multigroup data: a sensitivity analysis

Article (Accepted version) (Refereed)

Non-equivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis

Jouni Kuha[1]   Irini Moustaki

London School of Economics

[1] Correspondence Author: Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom, email: j.kuha@lse.ac.uk, tel: +44 (0)20 7955 6835.

This research was funded by the Economic and Social Research Council grant RES-239-25-0022 under the Comparative Cross-National Research Methods initiative. We are grateful to the three reviewers and the Associate Editor for their comments which substantially improved the article.

In studies of multiple groups of respondents, such as cross-national surveys and cross-cultural assessments in psychological or educational testing, an important methodological consideration is the comparability or "equivalence" of measurement across the groups. Ideally full equivalence would hold, but very often it does not. If non-equivalence of measurement is ignored when it is present, substantively interesting comparisons between the groups may become distorted. We consider this question in multigroup latent variable modeling of multiple-item scales, specifically latent trait models for categorical items. We use numerical sensitivity analyses to examine the nature and magnitude of the distortions in different circumstances, and the factors which affect them. The results suggest that estimates of multigroup latent variable models can be sensitive to assumptions about measurement, in that non-equivalence of measurement does not need to be extreme before ignoring it may substantially affect cross-group comparisons. We also discuss the implications of such findings on the analysis of large comparative studies.

Keywords: Cross-national surveys; Differential item functioning; Latent trait models; Measurement invariance; Cross-cultural assessment; Item Response Theory.

1. Introduction

The methodological research question considered in this article is the following: When we use a multigroup latent variable model to analyze data from multiple-item scales in a cross-national survey or other multigroup data, when and to what extent are comparative conclusions about the distributions of latent constructs across groups sensitive to an incorrect assumption that measurement equivalence holds for all of the items? We begin by briefly introducing the key terms in this statement.

In a cross-national social survey the same questions, translated into the local languages, are asked of respondents in several countries (for an overview of the field, see Smith 2010 and other chapters in Harkness et al. 2010). Data of this kind arise also from comparative studies in educational and psychological assessment, such as cross-national programmes of educational testing (e.g. the PISA and PIAAC programmes by OECD) where literacy and numeracy tests are administered to subjects in different countries to study levels of competence. A key purpose of such studies is to answer research questions about comparisons between different groups of subjects.

Often several questions (*items*) are used together as a multiple-item scale to measure an underlying (latent) construct, and analyzed using statistical latent variable models. We focus on *latent trait models* for categorical items since many survey questions are of a categorical nature, as are typical items in psychological and educational testing. In psychometric applications, these models are more commonly known as *item response theory* (IRT) models. Furthermore, because we consider data with multiple groups such as countries, we focus on *multigroup* extensions of the models.

As an illustrative example of such analysis, we consider data on a measure of depression from Round 6 of the European Social Survey (ESS) from 2012 (European Social Survey, 2014). The survey included a shortened version of the CES-D depression scale (Radloff 1997; Bracke et al. 2008). This has eight items, but for simplicity of illustration we have shortened it to six by omitting two reverse-coded items. The items ask how much of the time in the past week the respondent felt depressed (labeled item A below), felt that everything they did was an effort (B), had restless sleep (C), felt lonely (D), felt sad (E), and could not get going (F). The items had four response options ('none or almost none', 'some', 'most', or 'all or almost all' of the time) but we have dichotomized these by combining the last three levels, to align the example with the focus on binary items in our sensitivity analysis. For most items and countries, around 30-70% of the respondents gave an answer of at least 'some of the time'. The data are probability samples from general adult populations of 29 countries (as listed in Figure 1), with 752–2968 respondents per country and a total sample size of 54,673.

===== Figure 1 around here. =====

Figure 1 shows results from one multigroup latent trait model for these data. The points connected by the solid line are estimated means of a latent depression trait, measured by the six items, for each of the countries. Here higher values indicate higher levels of depression. The results show clear geographic regularities, for example with the lowest averages being for Nordic countries and the highest for Eastern Europe and the Balkans. The standard deviation of the trait in the UK is fixed at 1, so some of the differences in the country-level means (of up to 1.5 individual-level standard deviations) are clearly substantially large. Many of the between-country differences are statistically significant (standard errors of the estimated means are around 0.05).

One of the crucial issues in the design and analysis of such multigroup studies is *measurement equivalence*. This essentially means comparability: An item is equivalent across groups if it works the same way in all of the groups. If this is not the case, observed differences between groups will reflect not just true differences in the distributions of the constructs of interest but also non-equivalence of measurement.

In a multigroup latent variable model, non-equivalence is operationalized as an association between the group and an item. It is then possible to specify models with and without equivalence, by comparing them to assess the extent of non-equivalence, and, if necessary, to employ a model in which some items are non-equivalent. This is not yet done routinely in cross-national studies, where measurement is typically assumed equivalent by default, but the use of non-equivalence models is increasing. Such models are, however, practically and conceptually problematic to the extent that we would still prefer to avoid them if possible. Lack of measurement equivalence in multigroup comparisons thus presents a dilemma for the data analyst, where neither ignoring non-equivalence nor allowing for it are fully appealing as general approaches.

This dilemma would be fortuitously resolved if comparative conclusions were insensitive to assumptions about the measurement, in particular to wrongly ignoring any non-equivalence in it. But how often is this the case in circumstances that are commonly encountered in comparative research? The implications would be very different if group comparisons were severely distorted even by small variation in measurement, than if they were only affected by the grossest violations of equivalence. This is the question considered in this article. In other words, if we act as if our measures function in the same way in all groups, but in truth they do not, when will this cause substantively interesting comparative conclusions about the groups to be seriously misleading?

Figure 1 examines this question for the ESS data on depression. The points connected by the solid line are estimated country means under the assumption of equivalence, while those labeled with '1' are from the best-fitting model where one of the six items (specifically item F) is allowed to be non-equivalent across the countries. The other sets of estimates come similarly from models with 2, 3 and 4 non-equivalent items. What is of interest here is how the estimated means and rankings of the countries change between these models. It can be seen that there are many noticeable changes, especially with larger numbers of non-equivalent items. For example, the average levels of depression for Germany, UK and France are similar when estimated from the equivalence model, but dramatically different under non-equivalence models.

This illustration is an example of the kind of sensitivity analysis where we focus on a single observed data set. Recently, Oberski (2013) has proposed convenient methods for carrying out such analysis for multigroup structural equation models. Our focus in this article is on the other type of sensitivity analysis, which examines the sensitivity of estimates using numerical analyses of a wide range of scenarios on synthetic data. This has been done previously by Kaplan & George (1995) and De Beuckelaer & Swinnen (2011), using simulation studies for factor analysis models and focusing on significance tests of comparative hypotheses. There appear to be no other published studies of this kind, or any that consider the sensitivity of parameter estimates of the latent variable models for categorical items that we examine here. Furthermore, our analysis is implemented in a way which avoids the need for simulation and requires only one model fit for each scenario that we consider. This makes it possible to examine more cases and in a finer-grained manner than would be possible with a simulation study.

Methods and results of the sensitivity analysis are described in Section 4. Before that, multi-group latent trait models are defined in Section 2. Substantive interpretation of non-equivalence of measurement, statistical approaches to detecting and modeling it, and the motivation of the sensitivity analysis, are discussed in Section 3. The article concludes with a further discussion of the conclusions in Section 5.

2. Multigroup latent variable models

We first define the multigroup latent trait models which will be the focus of the sensitivity analysis. Discussions of different families of latent variable models can be found in, for example, Bollen (1989), Hagenaars & McCutcheon (2002) and de Ayala (2009), and of general frameworks for them in Bartholomew et al. (2011) and Skrondal & Rabe-Hesketh (2004). Literature on multigroup latent variable models goes back to at least Jöreskog (1971), Sörbom (1974) and Byrne et al. (1989) for linear factor analysis and to Muthén & Christoffersson (1981) for latent trait models; current overviews of multigroup models can be found in Kankaraš et al. (2011) and references cited therein. Some of this literature is discussed further in Section 3.2 below.

Let η denote a single latent variable for one respondent (for simplicity of presentation, we focus on a univariate η and omit the respondent subscript). Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be p dichotomous items (with values 0 and 1) that are regarded as measures of η. Both $\mathbf{Y}$ and η are random variables, and an observed value of $\mathbf{Y}$ is

$\mathbf{y} = (y_1, \ldots, y_p)$. Suppose further that each respondent belongs to one of G observed groups. The joint distribution of $\mathbf{Y}$ and η given the group is defined by $p_g(\mathbf{y}|\eta)\, p_g(\eta)$, where $p_g(\cdot)$ denotes a conditional or marginal probability density function and the subscript g that a distribution depends on group $g = 1, \ldots, G$. Here the *measurement model* $p_g(\mathbf{y}|\eta)$ describes how $\mathbf{Y}$ measure η in each group, and the *structural model* $p_g(\eta)$ specifies the distribution of the latent variable given the group.

For categorical $\mathbf{Y}$, the measurement model is given by the probabilities $P_g(\mathbf{Y} = \mathbf{y}|\eta) \equiv p_g(\mathbf{y}|\eta)$. We make the common assumption that the items are conditionally independent given the latent variable, so that $P_g(\mathbf{Y} = \mathbf{y}|\eta) = \prod_{j=1}^{p} P_g(Y_j = y_j|\eta)$ is determined by the univariate measurement models for each of the items Y_j individually. For binary items, $P_g(Y_j = y_j|\eta) = [\pi_j^{(g)}(\eta)]^{y_j}\, [1 - \pi_j^{(g)}(\eta)]^{1-y_j}$ for $y_j = 0, 1$, where $\pi_j^{(g)}(\eta) \equiv P_g(Y_j = 1|\eta)$. Finally, as η is unobserved, any inference will be based on the conditional distribution of $\mathbf{Y}$ given the group only, that is

$$P_g(\mathbf{Y} = \mathbf{y}) = \int \left[\prod_{j=1}^{p} [\pi_j^{(g)}(\eta)]^{y_j}\, [1 - \pi_j^{(g)}(\eta)]^{1-y_j} \right] p_g(\eta)\; d\eta. \tag{1}$$

In the *latent trait models* that we consider, we assume that $\eta \sim N(\kappa^{(g)}, \phi^{(g)})$ is continuous and normally distributed with mean $\kappa^{(g)}$ and variance $\phi^{(g)}$ in each group g. For the measurement models, we use the logistic model (Birnbaum 1968)

$$\text{logit}[\pi_j^{(g)}(\eta)] = \tau_j^{(g)} + \lambda_j^{(g)}\, \eta \tag{2}$$

where $\tau_j^{(g)}$ and $\lambda_j^{(g)}$ are the intercept and loading parameter (also known as the *difficulty* and *discrimination*) of the model for item $j = 1, \ldots, p$ in group g. The curve of $\pi_j^{(g)}(\eta)$ as a function of η is known as the *item characteristic curve* of item Y_j. For identifiability, with one trait η we must have $p \geq 3$ and need to impose some constraints to identify the scale of the latent trait, for example the condition that $\kappa^{(1)} = 0$ and $\phi^{(1)} = 1$ which we will use (in addition, some assumptions on measurement equivalence are needed, as discussed in Section 3.2).

The data on $\mathbf{Y}$ for all the respondents can be summarized as a $G \times 2^p$ group-by-items contingency table of the frequencies n_{gl} of observations with the response patterns $l = 1, \ldots, 2^p$ of the p binary items Y_j in the groups $g = 1, \ldots, G$. The sample size in group g is then $n_g = \sum_l n_{gl}$ and the total sample size $n = \sum_g n_g$. We treat the group totals n_g as fixed, and assume that observations for the n respondents are statistically independent. The log-likelihood function for model (1) is then

$$\ell(\boldsymbol{\theta}) = \sum_{g=1}^{G} \sum_{l=1}^{2^p} n_{gl} \log P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta}) \tag{3}$$

where $\mathbf{y}_l$ denotes the lth response pattern, $P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta})$ is defined by (1), and $\boldsymbol{\theta}$ denotes all the parameters of the model. Maximum likelihood estimates of the parameters are obtained by maximizing $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

Substantively interesting comparative questions typically center on whether the parameters of the structural model (the means and variances of the latent trait) vary between the groups. In contrast, any group differences in the measurement parameters are typically unwelcome and of little substantive interest. They define the question of equivalence or non-equivalence of measurement, which we discuss next.

3. Measurement equivalence and non-equivalence

3.1 Definitions and interpretations

Responses to individual survey or test questions do not only return information about the concepts they aim to measure but are also influenced by the particular question wording and other contextual factors (Schuman & Presser 1981; Tourangeau et al. 2000). When the influence of such factors varies systematically between respondents in different groups, there is non-equivalence of measurement across the groups. In cross-national research reasons for it may include imperfect comparability of translations of the questions, variation in response styles and fieldwork practices, and cross-national differences in the meaning and salience of the questions themselves. Instead of "equivalence", the term "invariance" is also often used (Johnson (1998) gives an excellent review of how these terms have been used in cross-cultural research). In modeling of psychological and educational tests, the terms *differential item functioning (DIF)* or *item bias* are also commonly used for non-equivalence of measurement.

In a multigroup latent trait model, measurement equivalence holds across groups $g = 1, \ldots, G$ for item j if $\pi_j^{(g)}(\eta) = \pi_j(\eta)$ for all groups and at every value of η, which in turn is true if both the intercepts $\tau_j^{(g)} = \tau_j$ and loadings $\lambda_j^{(g)} = \lambda_j$ of the measurement model (2) are the same across the groups. If this is the case for all the items $j = 1, \ldots, p$, measurement of η by this set items is *fully equivalent* across the groups. This definition of equivalence is generally adopted in the literature (see, for example, Mellenbergh 1989, Meredith 1993 and Millsap 2011 for clear discussions). Implicitly it goes back to at least Meredith (1964), who first made a clear distinction between equivalence in the measurement and structural models, and showed that the former may hold even when the latter does not. In particular, it is recognized that non-equivalence in latent trait (IRT) models should be defined as differences in the item characteristic curves between groups (see e.g. Lord 1980 and Mellenbergh 1989, 1992). Millsap (2011) provides an up-to-date discussion of the difficulties related to measurement in psychology and of non-equivalence of measurement in particular.

A measurement model may also possess *partial* rather than full measurement equivalence. This may mean that equivalence holds for some but not all of the individual items (Byrne et al. 1989) and/or of the measurement parameters of a given item (see e.g. Meredith 1993; Steenkamp & Baumgartner 1998 give a summary of the variants of this idea which are commonly distinguished in multigroup factor analysis).

3.2 Using modeling to detect and allow for non-equivalence of measurement

There is a large literature on incorporating non-equivalence within latent variable models. Its main themes are outlined in this section. A much more extensive treatment can be found in the excellent review by Millsap (2011) and references cited therein. Our discussion is fairly brief, because the focus of the rest of the article will then be elsewhere, namely in examining what happens if the analysis does *not* employ the methods reviewed in this section.

The first task is detecting non-equivalence of measurement. Within the latent variable framework, this can be done by comparing the goodness of fit of models with equivalence and (partial) non-equivalence. This approach goes back to the introduction of multigroup factor analysis by Jöreskog (1971); for examples of its use with latent trait (IRT) models, see for instance Thissen et al. (1988; 1993). The model comparison can be done using standard methods of model assessment, including likelihood ratio (LR) tests of nested pairs of models. In large samples this test has high power in detecting even small differences, so less oversensitive alternatives have been proposed. For factor analysis these include fit indices such as the RMSEA (Browne & Cudeck 1993) and the CFI (Bentler 1990), for which various cut-off points have been proposed (e.g. Chen 2007). These indices are not available for latent trait models, but statistics such as the AIC and BIC criteria or residuals in lower-dimensional marginal tables (Bartholomew et al. 2011) may be considered in their stead.

Among the most recent literature on detecting non-equivalence, Raykov et al. (2012) provide a critical discussion of the existing testing procedures, and Kim et al. (2012) study the performance of an adjusted likelihood ratio test when the multiple causes, multiple indicators (MIMIC) model is used for the detection. Jak et al. (2013) develop a test for use with a two-level factor model, and Rutkowski & Svetina (2014) study the appropriateness of the fit indices developed for multi-group confirmatory factor analysis models for testing for non-equivalence in large-scale surveys.

Second, if such methods detect non-equivalence, we can employ partial equivalence models where for some items the parameters of the measurement model do vary across the groups. It is also possible to extend the models further to aim to *explain* the non-equivalence, essentially by modeling with explanatory variables the cross-group variation in measurement parameters (see e.g. Soares et al. 2009; Davidov et al. 2012).

In our example of depression items in the European Social Survey, multigroup latent trait modeling reveals a large amount of non-equivalence in the six items across the 29 countries. First, a model with any one item freed to be non-equivalent fits better than the full equivalence model, according to the LR test and both AIC and BIC. Freeing any 2, 3, or 4 items improves the fit further (the best-fitting models of these types have items F, (B,F), (A,B,F) and (A,B,C,F) non-equivalent; estimated trait means from these models are shown in Figure 1).

The LR test and AIC indicate still better fit if a fifth item is freed, leaving only one cross-nationally equivalent item. Here, however, we are at the limit of identifiability of the model. Having just one equivalent item produces the same fit whichever this item is chosen to be, and also the same fit as the best-fitting model with no equivalent items (and an infinite number of other models; this result is used in the "alignment" method of Asparouhov & Muthén 2014, which seeks the latent scale on which the items are least non-equivalent in aggregate). All of these choices, however, give different estimates for the means and variances of the latent trait. Using such a model thus requires an arbitrary choice of *which* one item is declared to be non-equivalent, or a conclusion that no between-group differences in the structural models are meaningfully identified.

Measurement equivalence across 29 countries is always likely to be difficult to achieve. As a second example, we considered data for just Sweden, Norway and Denmark (this is also discussed in Section 4). Here we select a model with three items (B,C,D) non-equivalent (freeing also E improves the fit further according to LR test and AIC, but only slightly). So even for such a small group of fairly similar countries with closely related languages, half of the items are judged to be formally non-equivalent.

Bracke et al. (2008) also carried out multigroup analysis of the CES-D depression scale in the ESS, using factor analysis of the four-point versions of all eight items from Round 3 of the survey. Considering measurement equivalence by both gender and 25 countries, they settled for a partial equivalence model where 112 of the 400 intercept parameters in the measurement models were non-equivalent and the rest of the intercepts and all the loadings were equivalent. This conclusion was reached using the RMSEA, CFI and TLI statistics, whereas LR tests and the AIC and BIC would again indicate a still larger number of non-equivalent parameters.

3.3 What to do about non-equivalence? — Motivation of the sensitivity analysis

When such modeling is applied to data from large cross-national surveys or testing programmes, it is very often concluded that many of the measurement items are not equivalent across all the countries. This was the case in our example, and also for instance in Davidov et al. (2008), Kankaraš & Moors (2009), Meuleman & Billiet (2012), Nagengast & Marsh (2014), and several contributions to Davidov et al. (2011).

What then can be done about non-equivalence of measurement in multigroup analysis? Broadly, the main possibilities may be labeled *allow* and *ignore*. The "allow" approach means searching for a well-fitting partial equivalence model and using it as the basis of the final conclusions from the analysis. This may seem like a natural and straightforward way to address and contain the challenge of non-equivalence. However, we would argue that it is not, but that it may bring up severe difficulties of its own. Some of these difficulties are practical, with increased computing times and numbers of models to be considered. But the more profound problems are conceptual. For example, a well-fitting partial equivalence model with an unambiguously defined latent scale may not exist at all, i.e. there may not be even two items which are judged to be equivalent (this was the case in our example). These conditions on identifiability can be relaxed in Bayesian modeling with informative prior distributions (e.g. Soares et al. 2009, Fox 2010, and Muthén & Asparouhov 2012, 2013), but this rather hides rather than resolves the inherent limit of identifiability in partial equivalence models.

Other conceptual complications arise as soon as *any* items are treated as non-equivalent, i.e. when they work differently in different groups. This has several implications for the interpretation of the results. First, the non-equivalent items contribute relatively little to the estimated structural model; for example, the country means in Figure 1 are mostly driven by the data on the items which are equivalent in these models. Second, any latent variable scores assigned to individuals from such models will depend also on the group and not just the responses to the items. The dilemma is whether such measurement still means the same thing in all of the groups, and whether we are comfortable with using a model which incorporates it to draw comparative conclusions about the constructs. The dilemma does not admit an easy solution, as both positive and negative answers to it can be reasonably defended.

In contrast, the "ignore" approach means simply using the full equivalence model, irrespective of any evidence of non-equivalence. This treats each item in the same way in all groups, so it avoids the complications of non-equivalence models. Its own conceptual disadvantage is also clear: if we believe in the latent variable model with non-equivalence, the equivalence model is misspecified and will in general give biased estimates of the parameters. Ignoring this, however, can also be defended. Doing so means de-emphasizing the latent variable interpretation and treating the model mainly as a pragmatic aid in deriving a single measurement scale applied to all the groups.

3.4 Sensitivity analyses — Observed vs. synthetic data

Both ignoring non-equivalence of measurement and allowing for it thus have their disadvantages, and the choice between them is not obvious. This would not matter much if these approaches generally gave similar conclusions about the parameters of interest — in which case we could with confidence use the simpler equivalence model. But when is this likely to be case in practice? This is the question of the sensitivity of the equivalence model to misspecification: if there is true non-equivalence but it is ignored, how does this affect comparative conclusions about the groups?

There are two distinct but complementary approaches to such analysis. The first is to assess sensitivity in individual observed data sets, by fitting both equivalence and non-equivalence models and comparing estimates for the parameters of interest from them. Our analysis of the depression items, as discussed above and displayed in Figure 1, is an example of this approach. In this case we fitted all 58 distinct models where different subsets of the six items were non-equivalent. A method which reduces such modeling burden has recently been proposed by Oberski (2013). His "expected parameter change for parameters of interest" or *EPC-interest* provides a good approximation of what *would* happen to estimates of the structural parameters if specified measurement parameters were allowed to be non-equivalent. This in turn builds on older measures of Modification Indices and Expected Parameter Change (Jöreskog & Sörbom 1986; Saris et al. 1987; Saris et al. 2009) where the focus is on testing or estimating what would happen to some parameters if model constraints on those *same* parameters were relaxed.

In the second type of sensitivity analysis, changes in estimates are examined not for specific real data but in different hypothetical circumstances, varying any aspects of the true models that may affect the results. This provides a complementary approach which can be used for what analysis of observed data sets cannot do, namely to build the big picture of the effects of ignoring non-equivalence of measurement, both in outline and in detail. It can also provide information on choices in study design and on potential sources of sensitivity in individual studies. We will carry out sensitivity analysis of this type, as explained in the next section.

4. The sensitivity analysis

4.1 Research questions

Suppose that, in groups g, the observed frequencies n_{gl} of response patterns l for a set of binary items are generated by a latent trait model as defined in Section 2. The parameters of this model are the means $\kappa^{(g)}$ and variances $\phi^{(g)}$ of the latent trait in the groups, and the intercepts $\tau_j^{(g)}$ and loadings $\lambda_j^{(g)}$ of the measurement model (2) for each item j. The true model will in general contain some non-equivalence of measurement, so that for some items the true values of $\tau_j^{(g)}$ and $\lambda_j^{(g)}$ will be different across the groups.

Another model is then fitted to the data. This is of the same form as the true model, except that it will assume full measurement equivalence. In other words, the fitted model will allow the parameters of the latent trait to vary across groups, but it will constrain the measurement intercepts $\tau_j^{(g)} = \tau_j$ and loadings $\lambda_j^{(g)} = \lambda_j$ of each item to be the same in all groups. This model is thus in general misspecified.

Parameter estimates from the fitted model will then converge to some values.

Our goal is to examine what these will be in different circumstances, and how they relate to the parameter values of the true model. In particular, we want to compare estimates of $\kappa^{(g)}$ and $\phi^{(g)}$ that are obtained from the fitted model to the true values of these parameters, i.e. to assess how sensitive the conclusions about the parameters of the structural model are to misspecifications of the measurement model.

There are many degrees and dimensions in which the true model may deviate from full equivalence. We will focus on three of them: (i) the number or proportion of items which are non-equivalent; (ii) the degree of non-equivalence in these items; (iii) the relative strength of these items as measures of the latent trait. The aim is to gain understanding on how variation in these dimensions affects the bias due to model misspecification. Overarching these specific questions is a looser and semi-qualitative research question, which is whether the misspecification matters in the first place — that is, whether the bias in the estimated structural parameters may be large enough to be of concern, given levels of non-equivalence which may be expected in real data.

We will throughout consider examples where the fitted model has the same number of latent traits as the true model (one, in all of our examples). Another potential type of sensitivity is that the apparent number of traits, as chosen by empirical criteria, may also be affected by misspecification of the model. In other words, when measurement equivalence is imposed on the fitted model, true non-equivalence may instead manifest itself as a apparent extra traits in the structural model. The sensitivity of this aspect of the modeling will depend on which model selection criteria are used and how they behave under model misspecification. Examining these questions is beyond the scope of this paper but is an important topic for future research.

4.2 Methodology

The most obvious way to implement the sensitivity analysis would be a simulation study where many data sets were generated from each true model, the misspecified equivalence model fitted to each of them, and the results averaged over the simulations. This was done by Kaplan & George (1995) and De Beuckelaer & Swinnen (2011) in their sensitivity analyses for factor analysis models. This, however, requires a large number of data sets and model fits for every true model, so it will be time-consuming if we want to consider many different situations — as we wish to do, in order to examine the effect of several different factors and of varying the magnitude of measurement equivalence by small steps. Here we achieve this by replacing simulation studies with an approach which requires only one data set and model fit for each true model.

Our approach draws on general results on maximum likelihood (ML) estimates of misspecified models, as derived by Berk (1966), Huber (1967), Akaike (1973) and White (1982). These show that if data are generated from a model with true parameters $\boldsymbol{\theta}_0^*$ and a model with parameters $\boldsymbol{\theta}$ is then fitted to the data, ML estimates of $\boldsymbol{\theta}$ will converge (as $n \to \infty$) to the value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ which maximizes the expected value of the log-likelihood (3) of the misspecified model evaluated over the true model, i.e.

$$\mathrm{E}_{\boldsymbol{\theta}_0^*}[\ell(\boldsymbol{\theta})] = \sum_{g=1}^{G} \sum_{l=1}^{2^p} n_{gl}^* \log P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta}) \tag{4}$$

where $\mathrm{E}_{\boldsymbol{\theta}_0^*}$ denotes expectation with respect to the true model,

$$n_{gl}^* = \mathrm{E}_{\boldsymbol{\theta}_0^*}(n_{gl}) = n_g \, P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta}_0^*), \tag{5}$$

$P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta}_0^*)$ are the probabilities of the response patterns given the true parameter values, and $P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta})$ the probabilities under the fitted model. Here $\boldsymbol{\theta}_0$ is thus what ML estimates of the parameters of the misspecified equivalence model are estimating when the true model is a non-equivalence model with parameters $\boldsymbol{\theta}_0^*$. As the result is asymptotic, $\boldsymbol{\theta}_0$ will be a good representation of the expected value of the misspecified estimate in real data sets with sufficiently large samples, in roughly the same way as is required for the adequacy of ML estimates for correctly specified models.

Crucially, (4) is of the same form as the log-likelihood (3) but with n_{gl}^* in the role of the data. This suggests a simple two-step procedure for a sensitivity analysis:

(1) For a given true model with parameter values $\boldsymbol{\theta}_0^*$, calculate the true probabilities $P_g(\mathbf{Y} = \mathbf{y}_l; \boldsymbol{\theta}_0^*)$ of the different response patterns, and thus generate a contingency table with frequencies n_{gl}^* from (5).

(2) Fit the misspecified model of interest (here the equivalence model) with n_{gl}^* as the data. The resulting parameter "estimates" are $\boldsymbol{\theta}_0$.

For the pseudo sample sizes n_g used in step (1), only their *ratios* between groups g matter while the absolute values of n_g make no difference to $\boldsymbol{\theta}_0$. In all of our examples we have set n_g to be equal across the groups.

An analogous approach has been used for other purposes, by Rotnitzky & Wypij (1994) to examine biases due to missing data, Heagerty & Kurland (2001) to study the effects of misspecification of generalized linear mixed models, and Biemer (2011) to examine properties of latent class models (Biemer coins the term "expeculation" for it), but the method has not before been applied in the context where we use it here.

The models were fitted using Mplus 6.12 (Muthén & Muthén, 2010), and calculation of the expected frequencies n^*_{gl} and other data management were in R (R Core Team, 2012). Because likelihood functions for latent trait models can have several local maxima, multiple starting values were used for the estimation algorithm. Since in a one-trait latent trait model the direction of the trait is arbitrary, the fitted trait was rotated (i.e. reversed) where necessary so that the trait always had the same direction. Further information about the computer implementation and annotated code are available as supplementary material on the *Psychological Methods* website.

4.3 Design of the studies

Table 1 shows the settings used for the sensitivity analysis in Section 4.4. It gives the true parameter values for a one-trait latent trait model for three groups. In all of the cases the notional sample sizes n_g are taken to be equal across the groups.

===== Table 1 around here. =====

Part (a) of the table shows the values of the structural parameters, i.e. the mean $\kappa^{(g)}$ and variance $\phi^{(g)}$ of the single normally distributed latent trait in the groups. These are fixed at $\kappa^{(1)} = 0$ and $\phi^{(1)} = 1$ in group 1 for identification, but freely estimated in the other groups. Their true values are in fact 0 and 1 in all the groups, so the true situation is such that the structural model does not depend on the group. This equality is not imposed on the fitted model, where the trait mean and variance are estimated as separate parameters for groups 2 and 3. Their estimates will then converge to the true values, i.e. be equal across the groups, when the whole model is correctly specified, but not necessarily when the measurement model is misspecified. We use a true model of this kind, with a common structural model across the groups, in all of the analyses for convenience of interpretation: It is then easy to remember that any difference between the groups in the fitted structural model must be bias due to model misspecification.

The true measurement models are specified by fixing values of the response probabilities $\pi_j^{(g)}(\eta)$ for each item j and group g at two values of the trait η (at -1.5 and $+1.5$); these then determine the parameters ($\tau_j^{(g)}$ and $\lambda_j^{(g)}$) of the measurement model (2), and thus the response probabilities at other values of η. Part (b) of Table 1 shows how these fixed values are chosen for those items which are specified to have equivalence of measurement. We use items with two distinct measurement models, referred to as Lower- and Higher-discrimination items ('L' and 'H' for short). Their item response curves $\pi_j^{(g)}(\eta)$ are shown by the solid curves in plots (a) and (c) of Figure 2. They differ in that the response probability depends on η more strongly for an H-type than for an L-type item, i.e. that an H-type item is a more discriminating measure of η.

To introduce non-equivalence into the measurement model, we set the response probabilities to be different across the groups for some items. The probabilities shown in part (b) are 0.8 when η takes on the value -1.5. If an item is non-equivalent, this 0.8 is changed across the groups to the values shown in part (c) of Table 1. This shows 20 different cases, each with a different true measurement model. Across the 20 cases, the item probabilities for one of the groups (group 1 throughout) remain the same, while those for the other groups take on different values, corresponding to different magnitudes of nonequivalence. In case 16 the probabilities become equal across the groups, so this is the case of full equivalence. In each of these cases the response probability at $\eta = -1.5$ remains at the same fixed value (as in part (b) of the table) in all groups, and the probabilities at other values of η are determined accordingly.

===== Figure 2 around here. =====

A non-equivalent measurement model of this kind is illustrated by plot (a) of Figure 2 for an L-type item, and plot (c) for an H-type item. These show how the true probabilities $\pi_j^{(g)}(\eta)$ for a non-equivalent item depend on the trait η in the three groups, in the non-equivalence case 9 in part (c) of Table 1 (the line of circles indicates the fitted equivalence model in one of our scenarios; this will be discussed in the next section). This is a case where the level of non-equivalence might be regarded as moderate but not extreme. We will often discuss results for it specifically below, in order to provide additional focus for the discussion and a link for results across different plots.

Part (d) of Table 1 summarizes the other settings. They are defined by the eight combinations of three binary choices. First, the total number of items is either 4 or 8, and in each case half are of type L and half of type H. Second, the proportion of non-equivalent items is 25% (1 out of 4, or 2 out 8) or 50% (2/4 or 4/8). Third, the non-equivalent items are either all of type L or all of type H. For brevity, the different scenarios defined by these choices are sometimes referred to as “1H/4” (1 H-type item out of a total of 4 items is non-equivalent) and so on. For each scenario, we examine each of 20 cases of degree of non-equivalence that are defined in part (c) of Table 1.

When reporting the results, we consider also two further measures in order to summarize the degree of non-equivalence in the different cases. The first is an overall measure of *Differential Test Functioning* (DTF) which is often used to summarize non-equivalence (DIF) in IRT applications (see Millsap 2011 for a discussion). We define it as follows. First, let the “score” T be the average of a single respondent’s responses to the p binary items Y_j. Second, the expected score is $\mathrm{E}_g(T|\eta) = \sum_j \pi_j^{(g)}(\eta)/p$ for an individual who belongs to group g and has the value η of the latent trait. Next, consider two individuals who have the same value of the trait but who belong to our groups 1 and 3 (for which the measurement models are furthest apart). The expected difference in scores between these individuals is $D(\eta) = \mathrm{E}_3(T|\eta) - \mathrm{E}_1(T|\eta)$. The DFT measure we use is the square root of the expected value of $D(\eta)^2$ calculated over the distribution of η in group 1, i.e. the square root of $DFT = \int D(\eta)^2\, p_1(\eta)\, d\eta$ where $p_1(\eta)$ is the standard normal distribution. Somewhat loosely, this can be thought of as the difference in the score that a randomly selected member of group 1 would be expected to receive if the measurement model of group 3 rather than group 1 applied to that individual. This is 0 under full equivalence and positive otherwise.

The second summary measure arises from the work of Satorra & Saris (1985), who derived an approximation of the power of a likelihood ratio test of nested models in general covariance structure analysis. This approximation is very convenient here because it makes use of the quantities which are already produced as part of our sensitivity analysis. Generalized to our case, the result shows that if we were to carry out a likelihood ratio test of the null hypothesis that the full equivalence model holds, against the alternative hypothesis defined by the true model, the power of this test is calculated using a non-central χ^2 distribution (and degrees of freedom equal to those of the test). The crucial part of the result is that the correct noncentrality parameter of this distribution can be approximated by the likelihood ratio test statistic for this hypothesis, calculated with the true probabilities $P_g(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}_0^*)$ and the estimated probabilities $P_g(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}_0)$ in the roles of the observed and fitted proportions respectively. Calculating this power measure also requires that a notional sample size is specified for the test. Below we use a sample size of 1000 in each group.

4.4 Results of the sensitivity analysis for latent trait models

We begin with an example of the sensitivity of the estimated measurement models, before focusing on the structural models in the rest of the discussion. In Figure 2, the circled line in plot (a) shows the values to which the estimated response probabilities $\pi_j(\eta)$ for the one non-equivalent item of type L (i.e. with Low discrimination) would converge when an equivalence model was fitted to data where the true model followed case 9 of the 1L/4 scenario of the examples in Table 1 (i.e. one where 1 L-type item is non-equivalent, out of a total of 4 items). These values have been determined using the method explained in Section 4.2. We will refer to them (somewhat imprecisely) as the "estimated values" of the parameters under different degrees of misspecification. For the measurement model the question of interest is then what this estimate will be, given that the true values of the measurement probabilities vary across the groups.

===== Figure 2 around here. =====

In this first example the answer is that the estimated common measurement model is effectively an average of the true group-specific models in the three groups. This can also be seen in plot (b), which shows for this scenario the estimated intercepts τ_j and loadings λ_j for this item across all 20 cases of non-equivalence. The horizontal axis of the plot spans these 20 cases, and is labeled by the values of the measurement probability for group 3 shown (in bold) in part (c) of Table 1. Case 16 where the true model has measurement equivalence is indicated by the dashed vertical line, and the non-equivalence case 9 which we often use for illustration by the dotted vertical line.

The estimates in the correctly specified full equivalence case 16 are, as they should be, the true parameter values. In case 9, in contrast, the estimated common intercept is further from, and the loading closer to 0 than in case 16, thus defining the flatter curve which is seen in plot (a). This happens across all the 20 cases, in all of which the estimated measurement model is roughly an average of the group-specific ones. On the other hand, the estimates for the three equivalent items (one of which is also shown in plot (b)) are not affected by the misspecification of the non-equivalent item, but retain essentially their true values across all the cases of this scenario.

However, these broad conclusions do not hold in all situations. Plots (c) and (d) show analogous results for the 2H/4 examples. Here the common estimated measurement model is not an average of the three within-group ones but close to just one of them, the most discriminating model which applies in group 1. Estimates for the equivalent items are now also slightly more strongly affected than they were in the first example. Similar results are obtained for the 1H/4 scenario, suggesting that the pattern seen in plots (c) and (d) is due to the fact that the non-equivalent items are highly discriminating rather than that two of them are non-equivalent. These results illustrate that the common measurement model of an item estimated under equivalence need not be simply an average of the group-specific models for that item, but that there may instead be more complex interplay of models between groups and between different items — and with the structural models, as will be seen below.

Plot (d) of Figure 2 shows also, for case 9, approximations of the same estimated measurement parameters obtained from a standard simulation study. This is included to provide one illustration of this alternative to our non-simulation method of sensitivity analysis, and to show that the two approaches do produce similar results. In the simulation study, 1000 simulated data sets were generated with $n_g = 1000$ observations in each group (in a simulation, unlike in our approach, specific sample sizes do need to be specified) and with the true parameters of case 9. The stars shown in the plot are the average values, across the simulations, of the estimates of the four measurement parameters considered here. It can be seen that, as expected, they agree well with the asymptotic estimates obtained from our calculations. This is also true for the estimated structural parameters, which are shown in Figures 3 and 4 below.

In the rest of this section we focus on the estimated distributions of the latent trait in the groups. Figure 3 shows the estimated means $\kappa^{(3)}$ in group 3 for all the scenarios and cases defined in Table 1, and Figure 4 the standard deviations $\sqrt{\phi^{(3)}}$ similarly (in group 1 these parameters are always fixed at 0 and 1 respectively, and their estimated values for group 2 are roughly half-way between the ones for groups 1 and 3). Recall that in each case the mean in group 3 is really 0 and the standard deviation is 1, a result which is correctly recovered when the true measurement model agrees with the fitted full equivalence model. In the other cases, where the equivalence model is misspecified, the estimated parameters of the structural model incorrectly indicate some differences between the groups. We are interested in assessing how the nature and degree of this bias depends on the three factors considered here, i.e. the number and type of non-equivalent items, and the degree of non-equivalence in them.

===== Figures 3 and 4 around here. =====

For the degree of non-equivalence the conclusion is, unsurprisingly, that for each given scenario the bias increases with increasing non-equivalence, that is when moving further away from the full equivalence case. This increase is mostly smooth, but with one conspicuous exception, which occurs in the example where two low-discrimination items out of four items are non-equivalent (scenario 2L/4). There the bias in both the means and the standard deviations increases abruptly when we move into the most extreme cases of non-equivalence. Behind this sudden change is a dramatic example of the interplay between estimated measurement and structural models. What happens in these extreme cases is that the estimated measurement model is abruptly reconfigured, so that the (non-equivalent) low-discrimination items obtain larger discrimination parameters (loadings) than do the true high-discrimination items. The latent trait implied by the estimated measurement model is thus effectively re-defined, from a trait which is measured more strongly by the true H-items than the L-items into a trait for which the opposite is true. The estimated parameters of the trait are then similarly reconfigured, to reconcile this measurement model with the observed data.

We can examine the extent of non-equivalence in terms of both the proportion and total number of non-equivalent items. It appears that the proportion matters more, in that for a given type of item the bias is roughly similar for the same proportion of non-equivalent items irrespective of the total number (e.g. for the 1/4 and 2/8 cases), although with larger proportions there is also evidence that smaller totals are associated with slightly more bias. For the proportion, the results are clear and unsurprising: the higher the proportion of items which are really non-equivalent, the larger is the bias in the structural parameters estimated under the assumption of equivalence.

In terms of which types of items are non-equivalent, we focus on the contrast between more or less discriminating ("H" vs. "L") items. Here Figures 3 and 4 suggest somewhat different conclusions for estimated means vs. standard deviations of the latent trait. For the means there is evidence that when the proportion of non-equivalent items is large, the bias is larger when these items are more strongly discriminating, but these differences are relatively small. For the standard deviations, the differences are in the same direction but much clearer: biases in the estimated standard deviations are in all cases low when only L-type items are non-equivalent, but much larger when H-type items are non-equivalent. These differences are related to those for measurement models in Figure 2. There it was observed that for non-equivalent L-items the measurement model estimated under equivalence was an average of the group-specific models, while for H-items it was instead close to the most discriminating group-specific model. It appears that in the former case much of the impact of the model misspecification is absorbed by the measurement model, leaving the standard deviations of the trait largely unaffected — whereas the reverse is true when H-items are non-equivalent, in which case the estimated standard deviations are badly biased.

Motivated by these results on the standard deviations, we considered a variant of these analyses (detailed results are not shown). This was identical to the ones discussed above, except that in the estimation the variance of the latent trait was constrained to be equal (to 1) across all the groups. Recall that in these cases the true variances are in fact equal, and in practice we might often be willing to make this assumption. We might then suspect that specifying this part of the structural model correctly could reduce bias in the other part, i.e. the trait means. Instead, however, the opposite was true: the bias in the estimated means was worse when the variances are constrained than when they are not. This happens, in essence, because the estimated misspecified model involves a compromise of biases in different parameters in order to yield the best overall fit to the observed data. Since the measurement parameters are not allowed to assume their true, nonequivalent values, all the estimated parameters — both measurement and structural — shift to achieve this compromise. If the structural variances are fixed at particular values — even when these are the true ones — all of the shifting in the structural model is done by the means, which thus end up having still more bias.

Returning to the main results in Figures 3 and 4, we may also consider the absolute magnitudes of the biases. Considering first our illustrative, moderate non-equivalence case 9, we observe that the bias in the estimated mean is here around 0.1-0.15 when only a quarter of the items are non-equivalent, but rises to around 0.3 when half are non-equivalent. The former is arguably not large for a variable with a standard deviation of 1, but the latter could be considered non-negligible. These biases increase when the magnitude of the non-equivalence increases, although only slowly when the proportion of non-equivalent items is low. In the estimated standard deviations the biases are small when the non-equivalent items are low-discriminating but can easily be substantial when high-discriminating items are non-equivalent.

Figure 5 shows the estimated trait means (i.e. their bias, since the true value is 0) in all the 160 cases in Figure 3, against the two summary measures of the degree of non-equivalence defined in Section 4.3. In the upper plot, the means are plotted against the square root of the Differential Test Functioning (DTF) measure. Here the bias in the estimated means generally increases with increasing non-equivalence as measured by DTF. However, this relationship is not simple, in that similar levels of DTF lead to different degrees of bias in different situations. This finding, which is likely to hold for any single-number measure of non-equivalence, suggests that no such overall measure is likely to be a precise predictor of the bias arising from ignoring non-equivalence.

The second measure shown in Figure 5 is the approximate power of the test of non-equivalence. This power is high in most cases. Typically the only cases where it is not very near 1 are the ones with very small degrees of non-equivalence. In these cases the bias is also inevitably small. However, the bias is also small in very many cases where the power is effectively 1, for example in our reference case 9 (marked with triangles in the plot) in several scenarios. This suggests that in practice methods of model assessment like the likelihood ratio test may often detect non-equivalence in cases where ignoring it would make very little difference on the estimates of interest.

===== Figure 5 around here. =====

To relate one real example to these scenarios, consider data on the depression items in the ESS for Sweden, Norway and Denmark. As discussed in Section 3.2, model selection suggests that three of the six items are non-equivalent between these three countries, so we treat this as the true measurement model. In terms of the scenarios, the nonequivalent items in this model are between the H and L types in strength of measurement, and their nonequivalence across the countries is roughly comparable to the cases at around 0.55–0.7 on the horizontal axes of Figures 3 and 4. For such cases, the sensitivity analysis suggested a bias of around 0.2 trait standard deviations in the estimated means, and a downward bias in the standard deviations. This is also what we see in this example. Estimated trait means from the true nonequivalence model are 0 (fixed), -0.36 (standard error 0.07) and -0.19 (0.06) and standard deviations 1, 1.22 and 1.03 for Sweden, Norway and Denmark respectively. From the full equivalence model, the means are 0, -0.17 (0.05) and 0.02 (0.04) and standard deviations 1, 0.96, 0.80 similarly. The mean difference between Sweden and Denmark becomes nonsignificant when estimated under the equivalence model. Square root of the estimated DFT measure (between Sweden and Denmark) is 0.045 in the example.

This example is thus a "3/6" scenario with nonequivalence roughly comparable to our cases 10–14 (with a mixture of items and cases). That this is so for Sweden, Norway and Denmark, which form one of the most homogenous subsets of three countries in the ESS, suggests that for most other countries the non-equivalence of measurement will be greater and thus closer to more extreme ones of our scenarios — and the estimates from the equivalence and non-equivalence models will be correspondingly more different. These differences will also be large for any analysis of all or most of the 29 countries together, comparison of which would often be our main goal. Examples of the sensitivity of the estimated trait means for them can be seen in Figure 1.

In summary, our sensitivity analysis indicates that the magnitude of non-equivalence and the proportion of non-equivalent items are the strongest factors which affect the bias that arises from incorrectly employing a full equivalence model, while other factors such as which kinds of items are non-equivalent have a smaller but non-negligible impact. Constraining the variances of a latent variable to be equal across groups increases the bias in their means. Ignoring a limited amount of non-equivalence in a small proportion of items may not have a serious impact on estimates of the latent distribution across groups (even in many cases where standard model selection criteria would strongly reject the equivalence model). However, this will change when the amount of non-equivalence increases, in which case the bias can be substantial.

5. Conclusions and discussion

We have studied the sensitivity of conclusions from multigroup latent variable modeling when non-equivalence of measurement is incorrectly ignored. The simplest version of our research question was whether this can severely affect between-group comparisons of the distributions of the latent variables, given realistic levels of non-equivalence. The answer is yes in general, although there is also a fairly wide range of circumstances where the conclusions are relatively unaffected. These findings broadly agree with those of earlier studies by Kaplan & George (1995) and De Beuckelaer & Swinnen (2011) who studied analogous questions on significance tests.

We have focused on latent trait (IRT) models where the latent variable is continuous but the observed items are treated as categorical. We might expect fairly similar results for factor analysis models where both types of variables are continuous. Latent class models, where all the variables are categorical, could plausibly give rise to effects which were not observed in our examples with a continuous latent trait. Sensitivity of latent class models is thus a promising subject for future research. More work of this kind remains to be done also for models with continuous latent variables, especially in situations which were not considered here, These include, for example, cases with larger numbers of items or with more heterogeneity in the discrimination and non-equivalence of the items, and sensitivity of the selection of the number of latent variables. With the method of sensitivity analysis we have proposed, such explorations can be carried out much more efficiently than with conventional simulation studies.

Such findings are of interest because levels of non-equivalence which can lead to serious biases can and do occur frequently in large cross-national and other comparative studies. Our results thus do not support the convenient and reassuring conclusion that non-equivalence could in general be simply ignored. It is also not likely to be commonly limited to a few items only, which could be omitted at little cost. Instead, the analyst needs to make an explicit choice between two imperfect ways forward, as we discussed in Section 3.4. One of them is to use models which allow for non-equivalence in some of the observed items, an approach which comes with practical and conceptual costs of its own. The other is to use an equivalence model after all; this is always possible, but it effectively means treating the observed items as comparable at face value, and de-emphasizing their interpretation as measures of latent constructs. The results of our sensitivity analysis would then be interpreted as indicating how conclusions of interest might vary depending on how we treated the observed survey or test items, without arguing that treating them equally across groups is necessarily wrong. Either way, the choice faced by the comparative researcher is not a simple one.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kaidó. (Reprinted in Samuel Kotz and Norman L. Johnson (eds.), *Breakthroughs in Statistics, Volume I.* Springer-Verlag, New York, 1992, pp. 599–624, with an introduction by J. deLeeuw)

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*, 495–508.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Third ed.). Chichester: Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, *37*, 51–58.

Biemer, P. P. (2011). *Latent class analysis of survey error.* Hoboken, NJ: Wiley.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 425–435). Reading, MA: Addison-Wesley. (Chapter 18)

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bracke, P., Levecque, K., & Velde, S. Van de. (2008). The psychometric properties of the CES-D 8 depression inventory and the estimation of cross-national differences in the true prevalence of depression. In *Proceedings of the International Conference on Survey*

Methods in Multinational, Multiregional, and Multicultural Contexts (3MC) 2008. Berlin: Berlin-Brandenburg Academy of Sciences and Humanities.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

Chen, F.-F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558–575.

Davidov, E., Schmidt, P., & Billiet, J. (Eds.). (2011). *Cross-cultural analysis: Methods and applications.* New York: Routledge.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in: the adequacy of the european social survey to measure values in 20 countries. *Public Opinion Quarterly*, *72*, 420–445.

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 117–147). New York: Routledge.

European Social Survey. (2014). *Ess-6 2012 documentation report. edition 2.1.* European Social Survey Data Archive, Norwegian Social Science Data Services, Bergen.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York: Springer.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis.* Cambridge: Cambridge University Press.

Harkness, J. A., et al. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts.* Hoboken, NJ: Wiley.

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*, 973–985.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium in mathematical statistics and probability.* Berkeley: University of California Press.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 265–282.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. A. Harkness (Ed.), *Zuma-nachrichten spezial no. 3: Cross-cultural survey equivalence.* Mannheim: ZUMA.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Jöreskog, K. G., & Sörbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares methods [Computer software manual]. Mooresville, IN.

Kankaraš, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology*, *24*, 557–579.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and Research*, 279–310.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling*, *2*, 101–118.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*, 469–492.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mellenbergh, G. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. (1992). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Meuleman, B., & Billiet, J. (2012). Measuring attitudes toward immigration in europe: The cross-cultural validity of the ESS immigration scales. *Ask: Research & Methods*, *21*, 5–29.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York: Routledge.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335.

Muthén, B., & Asparouhov, T. (2013). *BSEM measurement invariance analysis* (Mplus Web Notes: No. 17). http://www.statmodel.com.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

Muthén, L. K., & Muthén, B. O. (2010). Mplus user's guide (sixth edition) [Computer software manual]. Los Angeles, CA.

Nagengast, B., & Marsh, H. W. (2014). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In L. L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues and methods of data analysis* (pp. 317–344). Boca Raton, FL: CRC Press.

Oberski, D. L. (2013). *Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models.* (To appear in *Political Analysis*)

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (`http://www.R-project.org`)

Radloff, L. S. (1997). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401.

Raykov, T., Marcoulides, G. A., & C.-H., L. (2012). Measurement invariance for latent constructs in multiple populations: A critical review and refocus. *Educational and Psychological Measurement*, *72*, 954–974.

Rotnitzky, A., & Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, *50*, 1163–1170.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57.

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

Saris, W. E., Satorra, A., & Veld, W. M. Van der. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561–582.

Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrica*, *50*(1), 83–90.

Schuman, H., & Presser, S. (1981). *Questions & answers in attitude surveys.* New York: Academic Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal,and structural equation models.* Boca Raton, FL: Chapman & Hall / CRC.

Smith, T. W. (2010). The globalization of survey research. In J. A. Harkness et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 477–484). Hoboken, NJ: Wiley.

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348–377.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge University Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

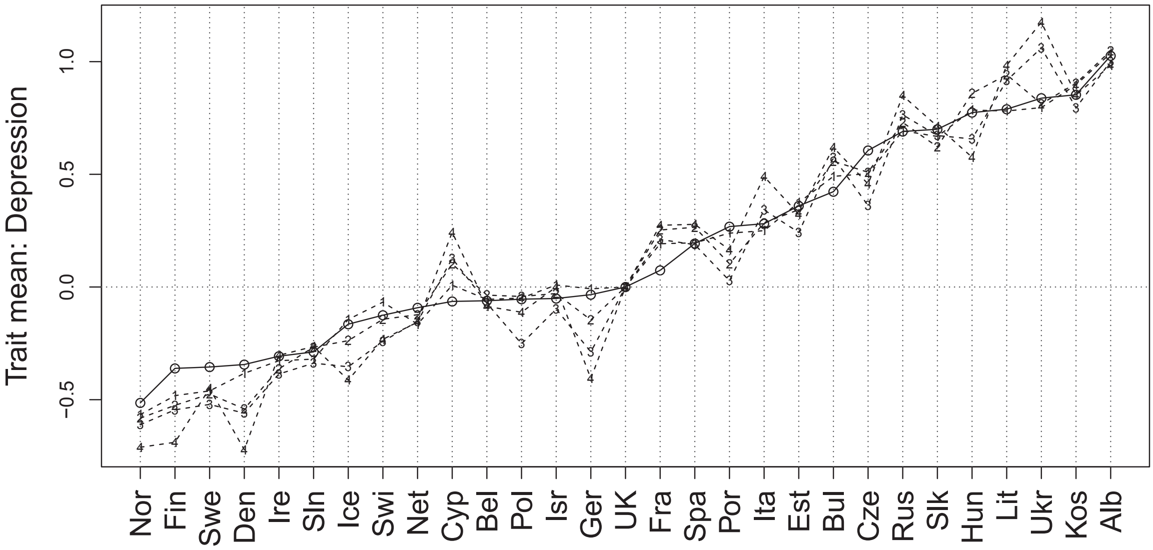

Figure 1. Estimated means by country for a latent trait indicating depression, from multi-group latent trait models for six binary items in Round 6 of the European Social Survey (2012). Higher values indicate higher levels of depression, and the average for UK is fixed at 0 for identification. The estimates connected by the solid line are from a model which assumes equivalence of measurement for all six items, and the other estimates from best-fitting models where 1, 2, 3 or 4 items are allowed to be non-equivalent across the countries.

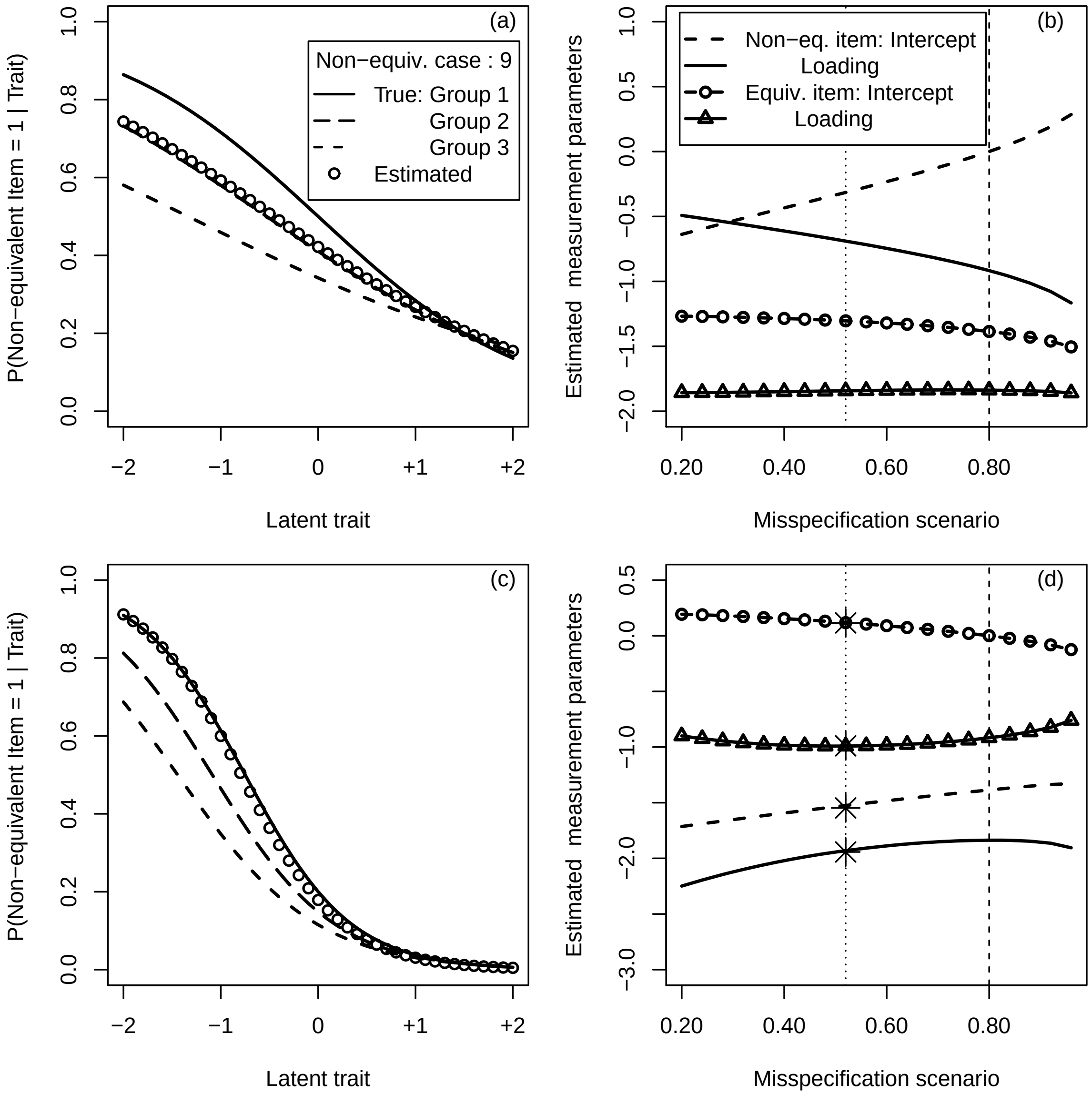

Figure 2. Results for measurement models estimated under measurement equivalence when the true model is specified as in Table 1. In each case there are 4 items, 2 of type H and 2 of type L (as specified by part (b) of Table 1). In the example shown in plots (a) and (b), one item of type L is non-equivalent. Plot (a) shows the true item response probabilities for this item as a function of the latent trait in the three groups, in non-equivalence case 9 in part (c) of Table 1. It also shows the common response probabilities that would be estimated under equivalence in this case. Plot (b) shows such estimated values of the parameters of the measurement model across all 20 scenarios of non-equivalence considered here, for the non-equivalent item and one equivalent item. Plots (c) and (d) are similar to (a) and (b) respectively, but for an example where the there are two non-equivalent items of type H. The horizontal axis of plots (b) and (d) is labeled by the values of the probabilities shown in bold in (c) of Table 1, and the vertical dashed and dotted lines indicate cases 16 (full equivalence) and 9 respectively. The stars in plot (d) indicate corresponding estimates from a standard simulation study discussed in the text.

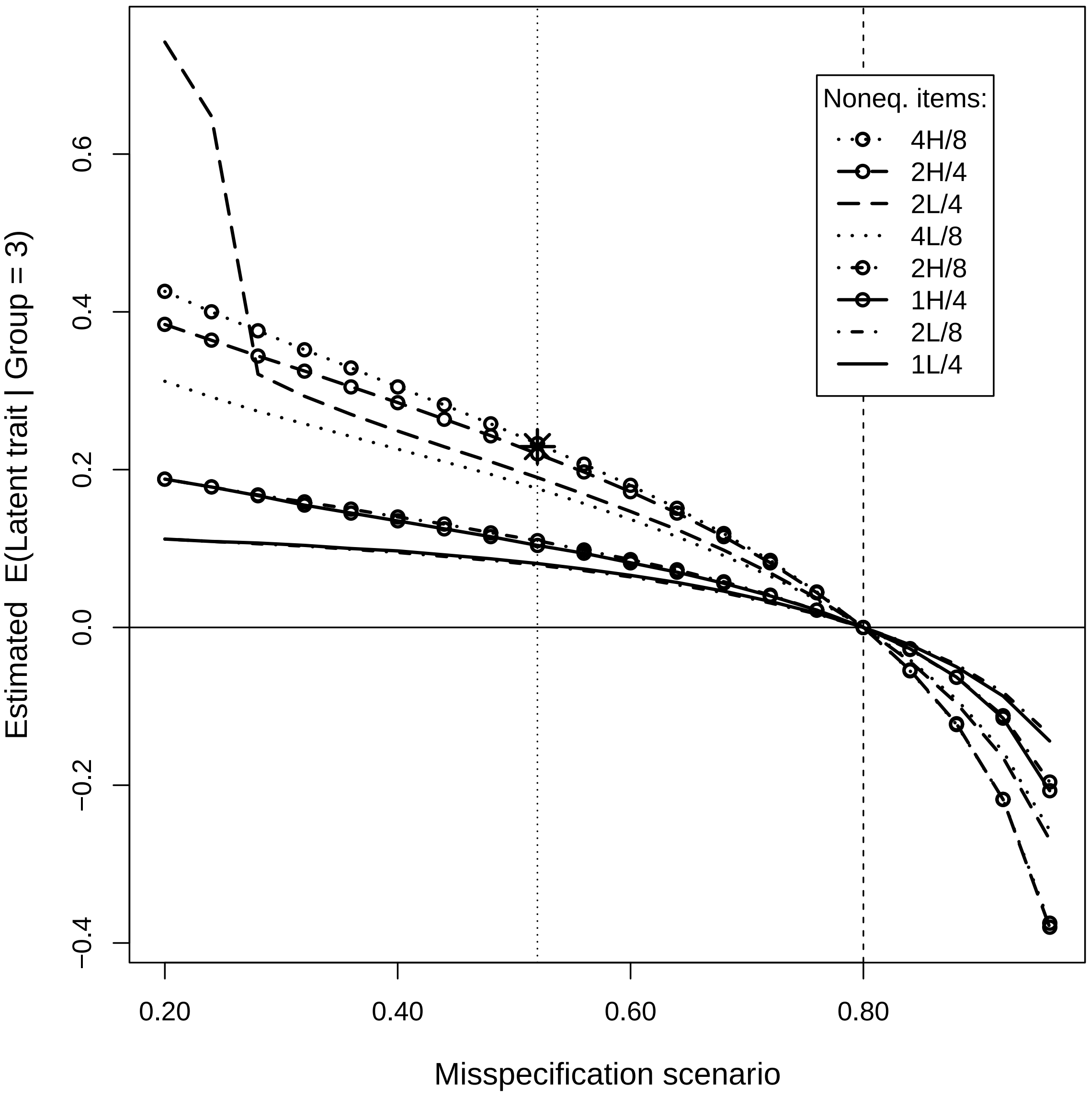

Figure 3. Results of sensitivity analysis of all the scenarios of latent-trait models defined in Table 1. For each of them, the figure shows the value to which the estimate of the mean $\kappa^{(3)}$ of the latent trait in group 3 would converge if a full equivalence model was fitted to data from the true model defined by the scenario. The true value of this mean is 0 throughout. The horizontal axis is labeled by the values of the probabilities shown in bold in (c) of Table 1, and the vertical and dashed and dotted lines indicate cases 16 (full equivalence) and 9 respectively. The star shows an analogous value of $\kappa^{(3)}$ obtained from a standard simulation study (as discussed in the text) for scenario "2H/4" (i.e. one where 2 high-discrimination items out of 4 total items are non-equivalent).

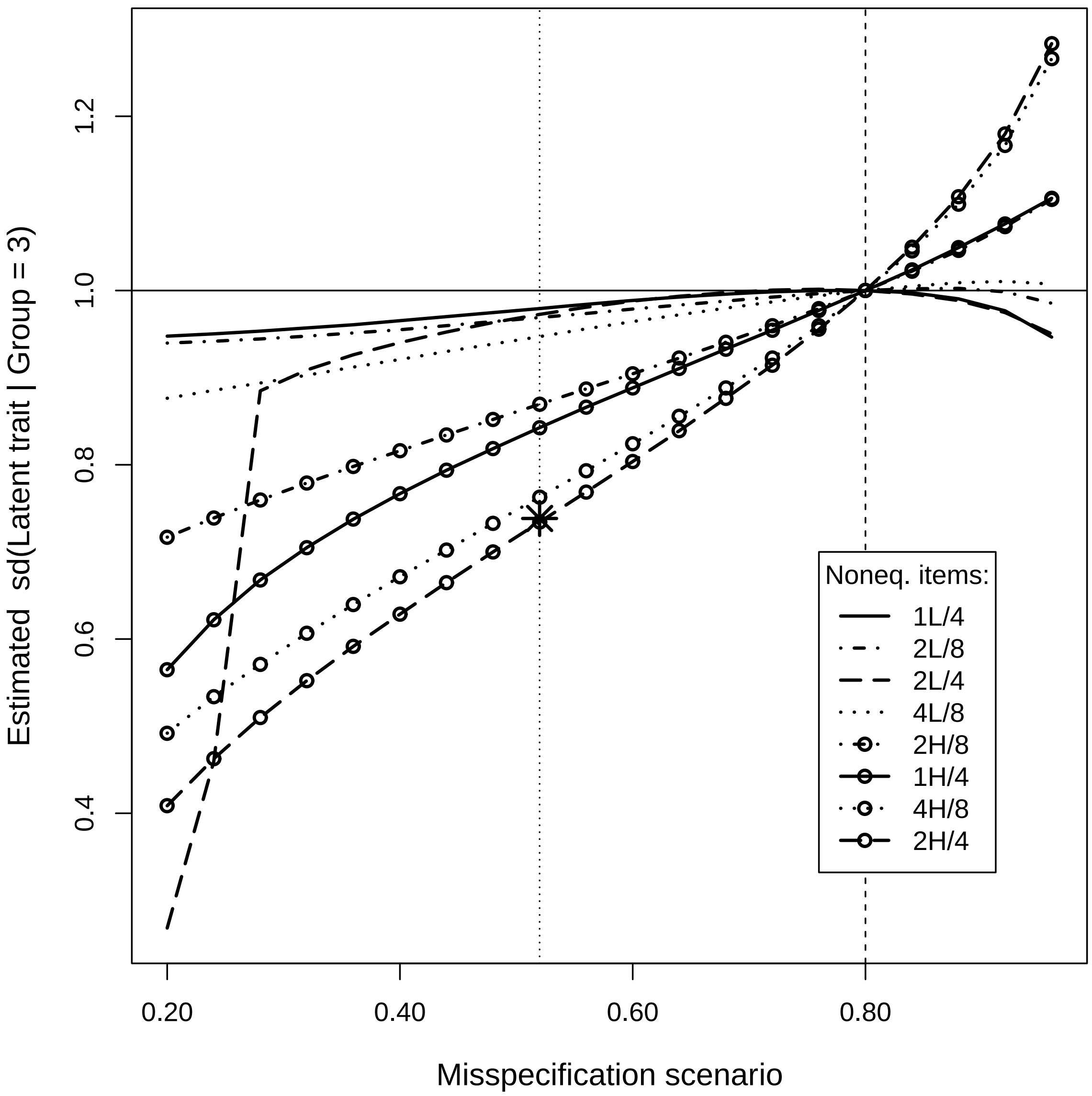

Figure 4. This figure is analogous to Figure 3, and shows the values to which estimates of the standard deviation $\sqrt{\phi^{(3)}}$ of the latent trait in group 3 would converge in the same scenarios. The true value of this standard deviation is 1 throughout.

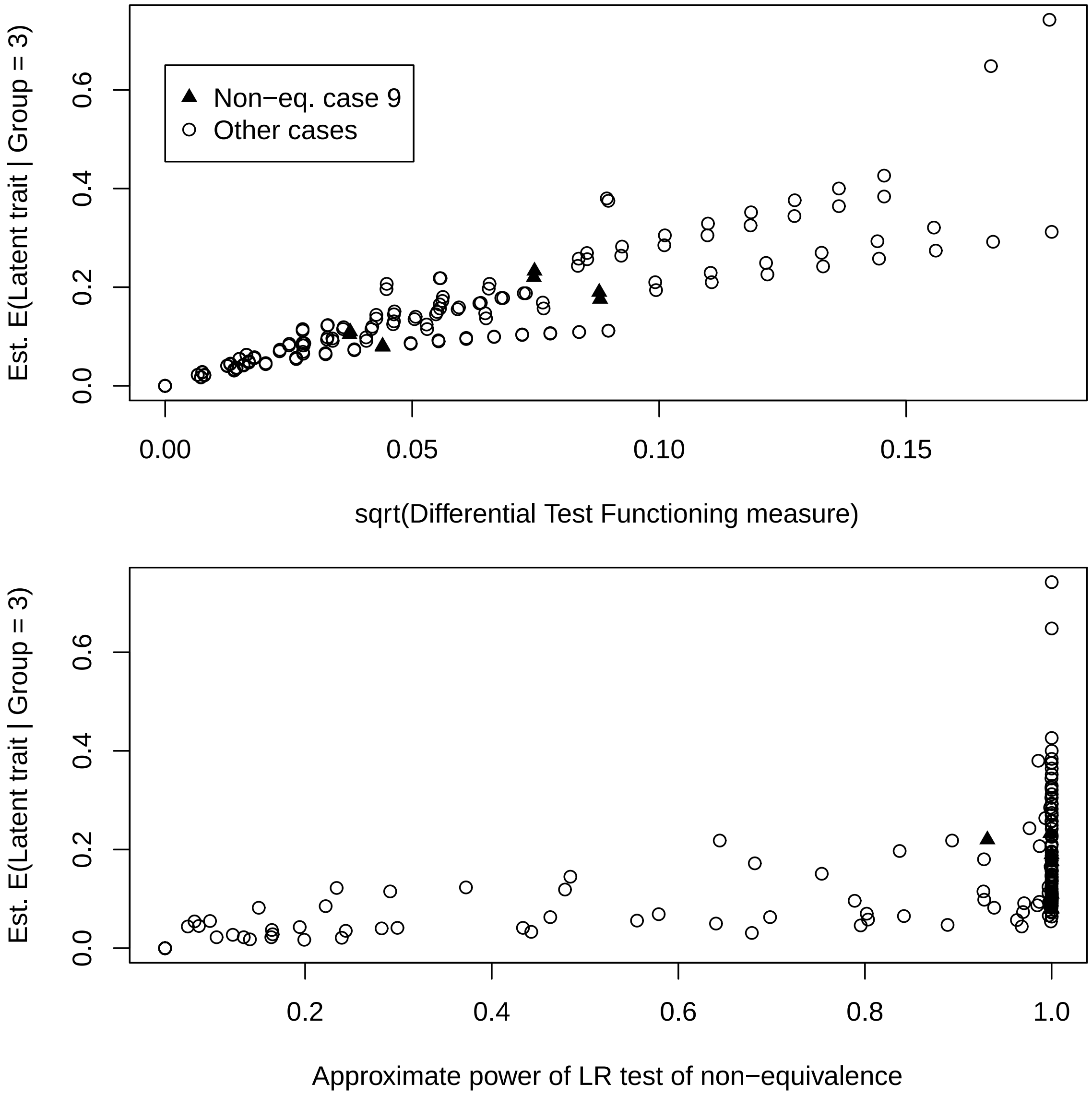

Figure 5. Plots of the bias in the estimated means of the latent trait in the cases shown in Figure 3 (on the vertical axes) against the square root of the measure of Differential Test Functioning (upper plot) and approximate power of a test of non-equivalence (lower plot) which are defined in Section 4.3.

(a) Structural model: One normally distributed latent trait η, with the following parameters in three groups:

Group (g)	Mean ($\kappa^{(g)}$)	s.d. ($\sqrt{\phi}^{(g)}$)
1	0 (fixed)	1 (fixed)
2	0	1
3	0	1

Sample sizes n_g are the same in all groups.

(b) Measurement model for those binary response items Y_j which are equivalent across the groups: Response probabilities $\pi_j(\eta) = P(Y_j = 1|\eta)$, and the intercepts τ_j and loadings λ_j of the measurement models:

	$\pi_j(\eta)$ given η at:				
	-1.5	0	$+1.5$	τ_j	λ_j
Lower-discrimination items (‘L’):	$0.8^{‡}$	0.5	$0.2^{†}$	0.000	-0.924
Higher-discrimination items (‘H’):	$0.8^{‡}$	0.2	$0.015^{†}$	-1.386	-1.848

The probabilities marked with ‡ and † are fixed to determine τ_j and λ_j and thus all other response probabilities.

(c) Measurement model for those items which are non-equivalent across the groups: 20 different scenarios, with the following values for the probabilities marked with ‡ in table (b):

	Sensitivity case (1–20)											
Group	1	2	3	...	8	9	...	15	16	17	...	20
1	.80	.80	.80	...	.80	.80	...	.80	.80	.80	...	.80
2	.50	.52	.54	...	.64	.66	...	.78	.80	.82	...	.88
3	**.20**	**.24**	**.28**	...	**.48**	**.52**	...	**.76**	**.80**	**.84**	...	**.96**

In each case the probabilities marked with † remain the same as in table (b), and the parameters of the measurement model are determined by these two fixed probabilities.

(d) Number and configuration of the items:
* 4 or 8 items, half of type L and half of type H.
* If 4 items, 1 or 2 non-equivalent. If 8 items, 2 or 4 non-equivalent.
* The non-equivalent items are all of type L, or all of type H.

Table 1: Parameter settings for the true model in the sensitivity analysis of latent trait models. Unless otherwise mentioned, the horizontal axis of each figure in Section 4.4 spans the 20 cases in table (c) and is labeled by the values of the probability in Group 3 (shown here in bold). The dashed vertical line in each of these figures corresponds to case 16 (full equivalence) and the dotted vertical line to case 9.