Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches

Miloš Kankaraš[*], Jeroen K. Vermunt* and Guy Moors*

Abstract

Three distinctive methods of assessing measurement equivalence of ordinal items, i.e. confirmatory factor analysis, differential item functioning using item response theory and latent class factor analysis, make different modeling assumptions and adopt different procedures. Simulation data are used to compare the performance of these three approaches in detecting the sources of measurement inequivalence. For this purpose, we simulated Likert-type data using two non-linear models, one with categorical and one with continuous latent variables. Inequivalence was set up in the slope parameters (loadings) as well as in the item intercept parameters in a form resembling agreement and extreme response styles. Results indicate that the item response theory and latent class factor models can relatively accurately detect and locate inequivalence in the intercept and slope parameters both at the scale and the item level. Confirmatory factor analysis performs well when inequivalence is located in the slope parameters, but wrongfully indicates inequivalence in the slope parameters when inequivalence is located in the intercept parameters. Influences of sample size, number of inequivalent items in a scale, and model fit criteria on the performance of the three methods are also analysed.

[*] Tilburg University, Department of Methodology and Statistics, Tilburg, Netherlands.

Introduction

There is a growing awareness among social scientists who are involved in empirical comparative research that the issue of measurement equivalence needs to be addressed. Measurement equivalence refers to 'whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute' (Horn & McArdle, 1992). Hence, it questions the comparability of data obtained from different groups, which is, of course, in the centre of any comparative research.

Several approaches for testing measurement equivalence with Likert-type items have been suggested, the more popular of which are multigroup confirmatory factor analysis (CFA) and methods for detecting differential item functioning (DIF) developed in the context of item response theory (IRT) (Steenkamp and Baumgartner 1998; Vandenberg and Lance, 2000; Raju, Laffitte, and Byrne, 2002). A third, less well known but very promising approach which combines multiple group latent class analysis (Clogg and Goodman, 1984; McCutcheon, 2002) with latent class factor analysis (LCFA; Magidson and Vermunt, 2001) was recently proposed by Moors (2004) (see also Kankaraš and Moors, forthcoming).

While the issue of the measurement equivalence (ME) has recently come to the fore in methodological studies, few of these studies focus on the comparison of methods for analysing ME (Meade and Lautenschlager, 2004a; Raju et al., 2002). As a result, applied researchers have little guidance as to which of these methods to use in their own research under which conditions. Whereas CFA, IRT, and LCFA use different terminology, model assumptions, and ME testing procedures, they also share numerous conceptual similarities. One of the two purposes of this article is to illustrate the similarities and differences between the three procedures for studying

ME by formulating them within a generalized latent variable modelling framework (Skrondal and Rabe-Hesketh, 2004). More specifically, it will be shown that each of the three investigated procedures can be viewed as a special case of a more general baseline measurement model. Second, we wish to determine the performance of the three procedures in detecting the different types of sources of measurement inequivalence when dealing with Likert-type ordinal questionnaire items, which is the most commonly used item format in survey research. For this purpose, we employ the three approaches -- CFA, IRT, and LCFA -- under simulated conditions.

The remainder of this paper is organized as follows. First, we present the generalized latent variable modelling framework, as well as describe the three approaches to the analysis of ME that will be compared in this study. After introducing the design of the simulation study, results are presented and discussed.

1. Approaches to investigating measurement equivalence

The CFA, LCFA, and IRT approaches to the analysis of ME stream from different methodological realms and have a somewhat different focus, use different procedures, and label parameters differently. This has, consequently, led to rather isolated practices of ME research which was usually constrained to the specific terminology and methods characteristic for a given methodological framework. However, aside from their apparent differences there are many common elements between the three approaches, from the theoretical assumptions about measurement models to the model parameters and measurement procedures employed. Although these conceptual and procedural similarities may often be overlooked, they can be better understood when approached from the perspective of generalized latent variable

modeling, which contains all three approaches as special cases. In the following we introduce the common framework and terminology that we will use in this study, as well as delineate both the differences and similarities between the three approaches. On the basis of this we formulate the main research questions and define the design factors for the simulation study that was setup to investigate and compare the performance of the three approaches to ME.

1.1 Generalized latent variable models

A common feature of the three relevant approaches to the analysis of measurement equivalence (CFA, IRT, LCFA) is that they are all latent variable models. More specifically, they are all three models in which one or more unobservable variables representing the constructs of interest (such as attitudes, values, traits, abilities) are connected to a set of observed measures, items, or indicators, for instance, to a set of as rating questions in the form of Likert scales.

Let Θ denotes the vector of L latent variables ($l = 1, \ldots, L$), $\mathbf{y}$ the vector of K observed variables ($k = 1, \ldots, K$), and y_k the kth observed variable. A latent variable model is a model for $f(\Theta, \mathbf{y})$, the joint probability density of the latent and observed variables. The causal mechanism shared by the three investigated latent variable models can be represented by their following two main assumptions (Skrondal and Rabe-Hesketh, 2004):

1. The responses on the observed indicators reflect an individual's position on the latent variable(s).
2. Indicators are independent of one another, controlling for latent variables. This is often referred to as the assumption of local independence.

These two assumptions can be expressed mathematically as follows:

$$f(\Theta, \mathbf{y}) = f(\Theta) f(\mathbf{y}|\Theta) = f(\Theta) \prod_{k=1}^{K} f(y_k|\Theta), \quad (1)$$

More specifically, the decomposition of $f(\Theta, \mathbf{y})$ into $f(\Theta) f(\mathbf{y}|\Theta)$ indicates that $\mathbf{y}$ depends on Θ, and the fact that $f(\mathbf{y}|\Theta)$ is replaced by the product $\prod_{k=1}^{K} f(y_k|\Theta)$ expresses that the K item responses are assumed to be independent of one another given Θ. Note that $f(\Theta)$ represents the distribution of the latent variables and $f(y_k|\Theta)$ the distribution of item k conditional on the latent variables scores.

As shown by Bartholomew and Knott (1999), depending on the specification of the distribution of the latent variables - $f(\Theta)$ - and the conditional distributions of the K item -- $f(y_k|\Theta)$ -- four main types of latent variable models can be obtained: factor analysis, item response theory models, latent profile analysis, and latent class analysis (see Table 1).

[INSERT TABLE 1 ABOUT HERE]

This four-fold classification shows that in factor analysis and IRT latent variables are continuous normally distributed, whereas in latent profile and latent class analysis they are discrete and thus have multinomial distribution. Moreover, in IRT and latent class analysis response variables are treated as nominal or ordered categorical variables with multinomial (or binomial) distributions, whereas in factor analysis and latent profile models they are treated as normally distributed continuous variables. It should be noted that the nature of the response variables affects not only

the form of $f(y_k|\Theta)$ but also the type of regression model connecting the item responses to the latent variable(s) (this is sometimes referred to as the link function). In factor analysis and latent profile analysis these are typically linear regression models, whereas IRT and latent class analysis usually make use of logit or probit models, yielding the well-known s-shaped relationship between latent and response variables.

It should be noted that we will use variants of factor analysis and latent classes analysis referred to as confirmatory factor analysis and latent class factor analysis, respectively. We will not use latent profile models.

1.2 Analysis of measurement equivalence

In its most broad term, measurement equivalence has been defined by Mellenbergh (1989) as:

$$f(\mathbf{y}|\,\Theta, g) = f(\mathbf{y}|\,\Theta)\,, \tag{2}$$

where g denotes group membership ($g = 1, \ldots, G$). Thus, measurement equivalence means that the probability distribution of the observed scores $\mathbf{y}$ conditional on the latent variable(s) Θ is the same for all groups. In other words, two individuals with the same Θ but from different groups are equally likely to give any specific set of responses.

Below we describe how the issue of ME is typically dealt within CFA, IRT, and LCFA, as well as introduce an integrating framework and common terminology

based on the generalized latent variable modelling approach presented in the previous section.

1.2.1 CFA

Assuming that the factor structure is the same for all groups, a multi-group CFA model implies the following linear regression model for item *k* for someone belonging to group *g* (Joreskög, 1971):

$$E(y_k \mid \Theta, g) = \tau_k^g + \sum_{l=1}^{L} \lambda_{lk}^g \Theta_l \,. \tag{3}$$

Here, τ_k^g represents the intercept and λ_{lk}^g the factor loading for latent variable *l*. When factor loadings are equal across groups ($\lambda_{lk}^1 = \lambda_{lk}^2 = \ldots = \lambda_{lk}^G$), so called 'metric equivalence' is achieved. However, for the valid comparison of factor means across groups, 'metric equivalence' is not sufficient, but the stricter 'scalar equivalence' condition should be satisfied. This condition requires that both intercepts and loadings are equal across groups ($\lambda_{lk}^1 = \lambda_{lk}^2 = \ldots = \lambda_{lk}^G$ and $\tau_k^1 = \tau_k^2 = \ldots = \tau_k^G$).

1.2.2 IRT

The most commonly used IRT models for polytomous items with ordered categories are the graded response (Samejima, 1969), rating scale (Andrich, 1978), partial credit (Masters, 1982), and generalized partial credit model (Muraki, 1999). The latter three are strongly related IRT models, which use an adjacent category

ordinal logit model to connect the latent variable to the item responses (see, e.g., Heinen, 1996; Vermunt, 2001). In this study, we used a multiple-group version of the generalized partial credit model. The log of odds of selecting category *s* of item *k* instead of category *s*-1 given a persons latent trait and membership of group *g* is assumed to have the following form (Bock and Zimovski, 1997):

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = a_k^g(\Theta - b_{ks}^g), \tag{4}$$

for $2 \leq s \leq S_k$, where *s* denotes one of the S_k categories of variable y_k. Here, a_k^g is the slope or 'discrimination' parameter for group *g* and item *k*, and b_{ks}^g is location or 'difficulty' parameter for group *g*, item *k*, and category *s*. Thus, both difficulty and/or discrimination parameters may vary across groups and cause inequivalence or 'differential item functioning', often referred to as DIF. When DIF is present only in the location parameters b_{ks}^g, it is called uniform DIF. Nonuniform DIF occurs when slope parameters differ across groups.

1.2.3 LCFA

Magidson and Vermunt (2001) proposed a restricted latent class model with multiple ordinal latent variables that they called latent class factor analysis (LCFA). It is a latent variable model with the *L* discrete latent variables with fixed and equidistant category scores. Similar to the multiple group extension of the standard latent class model (Clogg and Goodman, 1985; Hagenaars 1990; McCutcheon, 2002),

it is also possible to define a multiple group variant of the LCFA model (Moors, 2004; Kankaraš and Moors, forthcoming). Among various types of multigroup LCFA models for ordinal indicators we used the model in which, as in the partial credit model, item responses are related to the latent variables by means of an adjacent-category logit model:

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = \alpha_{ks}^{g} + \sum_{l=1}^{L} \beta_{kl}^{g} \Theta_l \qquad (5)$$

Here, α_{ks}^{g} are item- and category-specific intercepts and β_{kl}^{g} item- and factor-specific slopes. As can be seen, each of these can be assumed to differ across groups. The situation in which a set of α_{ks}^{g} parameters differ across groups is sometimes referred to as a 'direct effect' because such a model can also be defined by including the grouping variable as a nominal predictor in the model for item *k*. Such direct effects are present when group differences in item responses can not fully be explained by group differences in the latent factors. Note that this type of inequivalence is conceptually similar to scalar inequivalence in CFA and uniform DIF in IRT. Also β_{kl}^{g} parameters may vary across groups. This is sometimes referred to as 'interaction effects' as such group differences occur when the relationship between item responses and latent factors is modified by the group membership, i.e. by the interaction effect of the grouping variable and the latent factor concerned. Note that this is conceptually similar to 'metric inequivalence' in CFA and nonuniform DIF in IRT.

1.2.4 General model for the analysis of ME

The three presented approaches can be formulated using a unifying notation in following way:

$$E(y_k \mid \Theta, g) = \beta_{0k}^g + \sum_{l=1}^{L} \beta_{1kl}^g \Theta_l \ , \tag{6a}$$

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = \beta_{0ks}^g + \beta_{1k}^g \Theta \ , \tag{6b}$$

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = \beta_{0ks}^g + \sum_{l=1}^{L} \beta_{1kl}^g \Theta_l \ , \tag{6c}$$

As can be seen, we use a common notation for intercept (β_0) and slope (β_1) parameters. The slope parameters are conceptually similar across the three approaches, and indicate the strength of the effect of latent variable l on indicator variable k for group g (McDonald, 1999; Magidson and Vermunt, 2004). These terms were denoted by λ_{kl}^g, a_k^g, and β_{kl}^g in equations (3), (4), and (5), respectively. Whereas the interpretation of the intercept parameters is similar in the IRT and LCFA approaches (equations 6b and 6c), these are not directly comparable with those in the CFA approach (Meade and Lautenschlager, 2004a). Due to the different treatment of the observed variables (continuous vs. ordinal-discrete), CFA models have only one intercept per item, while IRT and LCFA models have S_k-1 free β_0 parameters per item. Note that the intercepts were denoted by τ_k^g and α_{ks}^g in equations (3) and (5). In the

IRT model, $\beta_{0ks}^{g} = -b_{ks}^{g}\, a_{k}^{g}$, that is, minus the item difficulty times the item discrimination.

[INSERT TABLE 2 ABOUT HERE]

Table 2 summarizes the relevant characteristics of the three approaches to ME. It presents the assumptions related to the latent and the response variables, along with the conceptual similarities between the model parameters – intercepts and slopes – as well as between the two most important forms of inequivalence in these parameters.

1.2.5 Procedures of the three approaches for analyzing ME

In all three approaches, the study of ME is based on the comparison of models that differ in the degree of inequivalence - in the number of item parameters that is allowed to vary across groups – with the aim to find the best fitting model with the lowest level of inequivalence possible. The most commonly used model comparison test in CFA is the chi-square difference test, which is in fact a likelihood-ratio (LR) test between nested models. Other popular fit indexes are measures such as Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Akaike Information Criterion (AIC). Note that a multiple group CFA typically starts from the baseline, unrestricted model in which all parameters are group specific, and subsequently moves to more restricted models (Vandenberg and Lance, 2000; Steenkamp and Baumgartner, 1998). Models are compared on a scale level with models in which they are nested, starting with the model with equal loadings and

followed by the model with equal loadings and intercepts. When inequivalence is found on a scale level, a researcher can proceed with item-level analysis in search of partially equivalent models. Measurement inequivalence is present to the degree that inclusion of equality restrictions of parameters across groups significantly deteriorates the model fit.

Similar model comparison procedures based on the LR chi-square tests are used in IRT based ME analysis. Differently form CFA, multiple group IRT starts from the most restricted equivalent measurement model, which is then compared with models in which the parameters in a single item are allowed to vary freely across groups (Thissen, Steinberg, and Wainer, 1988; Meade and Lautenschlager, 2004a). As for CFA, for IRT models guidelines are provided for the required level of invariance in order to be able to compare the latent scores across groups to be comparable; i.e., the minimal requirement is that parameters of at least one item should be invariant across groups (Mead and Lautenschlager, 2004; Steenkamp and Baumgartner, 1998).

In LCFA, the study of ME is based on the comparison of measurement models that differ in the number of direct and interaction effects included. LCFA typically relies on information criteria such as AIC, BIC and AIC3 that evaluate models both in terms of their fit and their parsimony, as well as on LR test (Moors, 2004; Kankaraš and Moors, forthcoming).

2. Data and method

2.1 Study overview

We performed a simulation study to determine the ability of the three latent modelling approaches to detect measurement inequivalence in rating scale questions

under a variety of conditions. The investigated conditions were related to (1) type of inequivalence, (2) nature of the latent variable, (3) number of inequivalent items in a scale, (4) sample size, (5) model fitting strategy, and (6) statistics used for model selection. More specifically, our research questions were:

1) How well do CFA, IRT and LCFA perform in detecting inequivalences in the slope and intercept parameters?
2) What is the influence of the assumed distribution of the latent variable (categorical or continuous) on the performance of the three approaches?
3) Does the performance of the three approaches depend on the number of inequivalent items in a scale? We compared conditions with one and three inequivalent items in a scale.
4) Does sample size have an effect on the ability of the three approaches to detect inequivalences? There were two different sample sizes: with 200 and with 1000 respondents per group.
5) Is the performance affected by whether one performs a scale-level analysis (the typical CFA approach) or an item-level analysis (the typical IRT approach)?
6) Are conclusions different when using LR (or chi-square difference) tests compared to using AIC?

We generated data sets containing three types of inequivalences: two types concerned the intercepts (β_0) and one the slope (β_1). Differences in β_0 parameters across groups reflected two well-documented response styles occurring with rating scales, namely, acquiescence and extreme response. Acquiescence is defined as a respondent's tendency to agree (or disagree) with given statements, irrespective to

their (positive or negative) content (Paulhus, 1991). To the extent that agreement tendency is associated with cultural background it is prone to be one of the factors that can cause measurement inequivalence in the cross-cultural comparisons (Moors, 2003; Billiet and McClendon, 2000). Extreme response style is defined as a respondent's tendency to choose the extreme categories of a response scale (i.e., completely agreeing or disagreeing) independently of the specific item content (Greenleaf, 1992). It may bias cross-cultural comparisons as it is a characteristic that has been shown to differ across cultures (Hui and Triandis, 1989).

Given that acquiescence and extreme response style might occur, what can we expect from the three approaches? When dealing with rating questions, probably the most important difference between the three approaches is the fact that the CFA, unlike the IRT and LCFA, does not contain a separate β_0 parameter for each item category (has only one intercept per item). Hence, it is expected that the CFA will experience more difficulties in detecting inequivalences in β_0 parameters (Meade and Lautenschlager, 2004a). On the other hand, since the β_1 parameters have a similar interpretation in each of the three models, they should have similar success rates in detecting differences between groups in these parameters.

Another important distinction between the approaches is in the nature of the latent variable(s). We were interested in what consequences misspecification of the latent variable distribution has on the validity of results of these approaches. For this purpose, we generated data sets based on two measurement models: one with a continuous, normally distributed latent variable, and one with an ordinal, uniformly distributed latent variable with three categories. The uniform distribution was chosen because it is rather different from a normal distribution and still relatively common in latent class analysis. From a theoretical point of view, one would expect that CFA and

IRT perform better with the continuous latent variable and LCFA with the discrete one.

One of the factors found to have substantial influence on the performance of the CFA and IRT methods is the sample size, i.e., because of lack of power a smaller sample size reduces the ability of the two approaches to detect inequivalence in data (Meade and Lautenschlager, 2004a; Meade and Bauer, 2007; French and Finch, 2006; Meade and Lautenschlager, 2004b). Aside from sample size we also varied the number of inequivalent items in a scale. This factor is found to influence the results of the CFA and IRT analyses in a similar manner (but to a lesser extent) as sample size does (Meade and Lautenschlager, 2004a).

Finally, we wanted to see whether the choice of model fitting strategy (scale-level versus item-level analysis) and fits measures (the chi-square difference test and the AIC) affects the encountered results. These latter two factors do not determine how the data are simulated, but how they are analyzed.

2.2 Properties of the simulated data sets

Data set were generated from two measurement models, one with a continuous and another with a discrete latent variable. Both types of models contained five items with five ordinal response categories, where the relationship between the latent variable and the items was based on the logit link function. The number of groups was set to two. The specific choice of the group-specific population parameters is discussed below. It should, however, be noted that we fixed the latent variable mean and variance (to 0 and 1) and assumed these to be equal across the two groups. The two different sample sizes were 200 and 1000 observation per group: a sample size of

200 can be considered to be the minimal recommended size in both IRT and CFA (Meade and Lautenschlager, 2004a), while a sample size of (at least) 1000 is common in cross-national research.

Out of five items in the scale, either one item (item 3) or three items (items 3, 4, and 5) were set up to be inequivalent across the two groups. There were three forms of inequivalence:

- Inequivalence in β_1 parameters,
- Inequivalence in β_0 parameters in form of agreement bias, and
- Inequivalence in β_0 parameters in form of extreme response bias;

The baseline value for the β_1 parameters was 1.00 (in logit units). The inequivalent β_1 condition was created by setting its value to 2.00 in Group 2 for the inequivalent item(s), that is, by assuming a stronger relationship between the latent variable and the item(s) concerned. This form of inequivalence may occur when other factors influence the relationship between latent and response variables and groups differ on these factors.

The baseline values for the category-specific β_0 parameters were -1.0, 0.5, 1.0, 0.5, and -1.0, which roughly approximates a normal distribution for the answers on a Likert scale.[1] In the agreement bias condition, inequivalence was defined by adding 1 to the highest rating parameter and subtracting 1 from lowest rating parameter for Group 2. This resulted in following β_0 parameters for Group 2: -2.0, 0.5, 1.0, 0.5, and 0.0. Presuming that the Likert scale represent answer options ranging from 'completely disagree' (category 1) to 'completely agree' (category 5), then this

[1] The implied item distributions for each level of the ordinal latent variable and for each group are provided in Appendix 1.

pattern of inequivalence would represent 'agreement bias' in Group 2 since it is more likely to select the higher categories than Group 1, controlling for the latent variable.

The extreme response bias condition was defined by adding 0.5 to the β_0's for the two extreme categories (categories 1 and 5), and subtracting 0.17 from categories 2 and 4 and 0.67 from category 3. Consequently, for inequivalent item(s) in Group 2 the β_0's were -0.5, 0.17, 0.67, 0.17, and -0.5. In this way, we have 'flattened' the answer distribution for Group 2 by making the β_0 parameters more similar to one another. This pattern of β_0 parameters resulted in higher probability of answering with the extreme options (1 and 5) compared to the respondents from Group 1, controlling for latent variable.

By combining the four design factors – distribution of latent variable (2 conditions), form of inequivalence (3), sample size (2), and number of inequivalent items (2) – we obtained 24 different conditions. For each of these conditions, we performed 100 replications. So, in total 2400 data sets where generated and analyzed.

2.3 Analyses of the simulated data sets

The three studied latent variable modeling approaches are accompanied with somewhat different model fitting strategies for studying ME. CFA is typically conducted at the scale level, i.e., by changing the parameter settings for all scale items simultaneously and subsequently comparing this model with a baseline model. A researcher would typically proceed to the item-level analysis after inequivalence is found at the scale level. In contrast, the IRT approach starts with separate item-level tests, without prior testing for scale-level inequivalence, which involves changing the parameter settings for a single item at a time. The LCFA procedure combines scale-

and item-level analysis. In order to foster comparison of the results between the three approaches, which was our primary research interest, we used both scale- and item-level procedures. Moreover, we conducted item-level analyses for all items, irrespective of whether there was evidence for scale-level inequivalence, which allowed us to determine how well the three procedures can detect inequivalence in a given item or set of items, irrespective of their ability to detect inequivalence at the scale level.

For the scale-level analyses, we first estimated a model in which all item parameters were allowed to vary across the two groups (Model A). This unrestricted model served as our baseline model. Then, we tested the equivalence of β_1 parameters by comparing the former unrestricted model with a model in which all β_1 parameters are fixed to be equal across the two groups (Model B). When the β_1 parameters were found to be equivalent, we conducted the final step in which we tested equivalence of β_0 parameters by contrasting the previous restricted model with equal loadings with a model in which both β_0 and β_1 parameters are restricted to be equal across groups (Model C).

For the item-level analysis, we compared models in which the β_1 or β_0 for one items are equated across the two groups with the unrestricted model. More specifically, inequivalence in the β_1 for item k is assessed by comparing the unrestricted model (Model A) with a model in which this parameter is equated across the two groups for item k (Model Bk). In order to test for inequivalence in β_0 at the item level we need to assume equivalence in β_1 parameters. Therefore, testing scalar equivalence of item k was based on the comparison of the model with equal loadings for all items (Model B discussed above) with the model which in addition assumes that the intercept is equal for the item concerned (Model Ck).

As model selection measures we used the chi-square difference (or LR) test, which is common in CFA- and IRT-based procedures (Vandenberg and Lance, 2000; Meade & Lautenschlager, 2004a), and the AIC that is typically used in LCFA as well as in CFA (Kankaraš and Moors, forthcoming). An alpha level of 0.05 was used in all analyses. In CFA and IRT, for identification purposes, it is required that one item is specified to be invariant. We used the first item for this purpose, except for the models in which the invariance of this item was tested, in which case the second item was chosen as the reference item.

Data simulations and analyses were conducted using version 4.5 of the Latent GOLD program (Vermunt and Magidson, 2008). It includes a syntax module which proves to be very flexible in modelling options necessary for our simulation study. Examples of Latent GOLD syntax used are presented in Appendix 2.

3. Results

The results of our study will be summarized using three outcome measures. The first one is the number of *true positives*; that is, the number of replicate samples (out of the 100 per design cell) in which the applied procedure correctly identified inequivalence in the item parameters. The second is the number of *false positives* or type I errors, referring to the number of replicates in which inequivalence is identified in the wrong parameters. One such example is the case in which the actual inequivalence was in the β_1 parameters but the model concerned finds inequivalence in the β_0 parameters (or vice versa). The third outcome variable involves type II errors or *false negatives*, which counts the number of replication samples in which

inequivalence in neither β_1 and β_0 parameters was detected where it should have. This occurs, for instance, as a result of lack of power of the performed test.

[INSERT TABLE 3 ABOUT HERE]

Table 3 reports the number of replicate samples in which inequivalence is detected for the conditions with three inequivalent items in a scale. Information is given for the scale as well as the item-level analyses. These results are obtained when using the chi-square difference test as the statistic for model selection. Three forms of inequivalence are listed, i.e., inequivalence in the form of agreement bias, extreme response bias, and inequivalence in the slope parameters. Separate results are presented for the two types of latent variable distributions (ordinal and continuous) and two different sample sizes (200 and 1000). Below, we first discuss the results for agreement bias, then for extreme response bias, and subsequently for inequivalence in the slope parameters. Later on we present more detailed findings regarding the effects of the form of inequivalence, the number of inequivalent items, and the used fit statistic.

3.1 Agreement bias

The results of the scale-level analyses indicate that CFA, IRT and LCFA perform well in detecting inequivalence in the form of agreement bias. For all conditions, the 'false positive' rate was smaller than 8% for each of the three approaches. Furthermore, invariance in β_0 parameters ('true positive' rate) is found in most of the remaining cases, although some differences in performance across

conditions can be observed. One differences between the three approaches that attracts attention is that the CFA test shows somewhat higher 'false negative' rates with the smaller sample sizes of 200 (23 and 32 for the ordinal and continuous models respectively) suggesting that the power of the model is somewhat to low. Other than this problem, there are no major differences in results of the three approaches across the given conditions.

Results from the item-level analysis show a rather similar pattern. Here, again all three approaches detect inequivalent items rather well, but with reoccurring power problems with the smaller sample size. 'False positive' rates for items 1 and 2 are rather low and inequivalent items 3, 4, and 5 are identified with perfect accuracy in the large sample size conditions. However, with the smaller sample size of 200, detection rates drop significantly, especially for CFA and IRT.

3.2 Extreme response bias

When inequivalence appears in the form of extreme response bias, scale-level results shows huge differences in performance between approaches. IRT and LCFA approaches turn out to have low 'false positive' rates (<10) and high rates of 'true positives' (>80) in all conditions. In contrast, CFA either wrongly indicates that inequivalence is present in β_1 rather than in β_0 parameters ('false positive') or indicates that there is no inequivalence ('false negative'). Again, power issues are noticeable, as detection rates are systematically lower for the smaller sample size condition.

Similar results are obtained with the item-level analysis. CFA fails in detecting items with inequivalence in β_0 parameters. The IRT and LCFA approaches, on the

other hand, perform much better with high ‘true positive’ rates and small ‘false positive’ rates. ‘True positive’ rates are somewhat smaller in the smaller sample size simulations, but even then, they are rather high (>64). The only significant difference between these IRT and LCFA is noticeable when a continuous latent variable is assumed for which LCFA has somewhat higher ‘false positive’ rates (10) than IRT.

3.3 Inequivalence in slope parameters

All three approaches prove to be good in detecting inequivalence regarding the β_1 parameters (the loadings). When samples with 1000 respondents are simulated ‘true positive’ rates are almost perfect. However, for the smaller sample size the ‘true positive’ rates are substantially lower, particularly in the IRT approach. Once again, the LCFA approach has somewhat better performance when the latent variable is ordinal compared with the continuous latent variable condition (100 versus 75 true positives, respectively). An unexpected finding is that in the small sample case also CFA has somewhat higher ‘true positive’ rates in the ordinal compared to the continuous latent variable condition (100 and 84, respectively). It is also worth noting that the CFA more frequently yields 'false negatives' than 'false positives', whereas for IRT and LCFA both types of mistakes show up more equally.

The analyses at the item level show similar patterns as those at the scale-level. While all three approaches demonstrate good ability to detect inequivalent items, their efficacy is affected by the sample size. Power issues are, however, less pronounced for the CFA and LCFA when the latent variable is ordinal, whereas in the IRT

approach similar rates are observed irrespective of the nature of latent variable. 'False positive' rates are generally small (<10).

3.4 Other results

[INSERT TABLES 4, 5 and 6 ABOUT HERE]

In this section, we present more details on the simulation results for three different design factors: Table 4 focuses on the type of inequivalence, Table 5 compares the two conditions which differ with respect to the number of inequivalent items, and Table 6 compares the results obtained with two different fit statistics. We report the average and the standard deviation of the number of replications in which inequivalence was detected across levels of the other conditions.

Comparing results for different types of inequivalence (Table 4) we find that in the case of agreement bias, the LCFA is somewhat better than the other two approaches in detecting inequivalence at the item level, while all three approaches have similar success at the scale level. CFA tests are clearly not able to correctly identify extreme response bias on both scale and item level (low 'false positive' rates). In this situation IRT and LCFA perform much better, with similar, relatively high 'true positive' rates across the board. When β_1 parameters are different across groups, the CFA approach has somewhat better 'true positive' rates compared to the other two approaches, while IRT and LCFA tests have somewhat higher rates of 'false positives'. Rates of 'false negatives' are generally low (M<26) for all approaches indicating relatively satisfactory levels of tests' power.

Varying the number of inequivalent items in a scale (Table 5) has similar effects on the results as varying the sample size which was discussed in the previous section. In particular, the presence of more inequivalent items increases the 'true positive' and decreases the 'false positive' rates. Nonetheless, there are some differences. First, contrary to the sample size effect, the number of inequivalent items does not affect detection rates in an item-level analysis. Secondly, while sample size affects the power of all three approaches, the number of inequivalence items in the scale does not influence the performance of CFA.

The two fit criteria that we used in our analyses yielded, generally speaking, similar results across approaches and conditions. However, as can be seen from Table 6, in the IRT and LCFA scale-level analyses AIC performed less well than the various chi-square difference tests. The AIC measure yielded somewhat lower 'true positive' and higher 'false negative' rates for these conditions. In the item-level analyses, on the contrary, AIC has slightly better 'true positive' rates with all three approaches, but also higher rates of 'false positives', especially with LCFA.

4. Discussion

The main finding of our simulation is that all three investigated approaches --- CFA, IRT and LCFA -- are generally able to detect inequivalences in rating scale items at both the scale and item level. There is one clear exception to this general finding, i.e., when the item intercepts differ as a results of differential extreme response styles, the CFA test is less adequate and wrongfully points at inequivalent slopes instead of intercepts. Although this might come as a surprise at first glance, it

becomes more understandable if one realizes that an extreme response style has a similar effect on the item distribution as larger slope parameters have. The subtle difference between these two forms of inequivalence can only be detected with a model with separate β_0 parameters for each response category, like IRT and LCFA, but unlike CFA.

The fact that CFA may have difficulty to detect inequivalences in β_0 parameters has been indicated before (Meade and Lautenschlager, 2004a). However, contrary to Meade and Lautenschlager's conclusions, we showed that this is not always the case, but that it depends on the source of inequivalence. With inequivalences associated with differential acquiescence, CFA was equally successful as were IRT and LCFA. Hence, we need to conclude that the performance of the CFA approach in identifying inequivalence in β_0 parameters depends upon the specific form in which β_0 parameters differ across groups. Inequivalence in β_1 parameters is detected in a rather high number of cases by all three approaches, although with slightly more precision by the CFA. These results for the CFA are expected given that the CFA is well designed for analysis of β_1 parameters and are in accordance with results in previous studies (Meade and Lautenschlager, 2004a; French and Finch, 2006). However, we like to underscore that the analyses also confirmed that IRT and LCFA, aside for being well suited for inspecting differences in β_0 parameters, are also successful in detecting difference in β_1 parameters.

Another important finding, both from a theoretical and practical point of view, is that the nature of the latent variable had little influence on the performances of different approaches. CFA and IRT that assume continuous normally distributed latent variables, have very similar performances with ordinal and continuous latent variables. Likewise, the LCFA also proves its omnipotence with generally good

performance in models with continuous latent variable, notwithstanding that it still performs slightly better when used in its own playground, with discrete-ordinal latent variables. Admittedly, we only compared a continuous normal with a three-category uniform distribution and in principle other latent variable distribution could give different results. Nevertheless we feel that this robustness of the three methods for possible misspecification of the latent variable distribution is, from a researcher's practice point of view, surely encouraging.

The number of respondents per group (sample size) proved to be one of the most important factors affecting the performance of the three approaches. CFA, IRT, and, to somewhat lesser degree, LCFA are all vulnerable to lack of power caused by small number of subjects, which is found to be the case in many previous Monte Carlo studies (Meade and Bauer, 2007; French and Finch, 2006; Meade and Lautenschlager, 2004a). Thus, when using the CFA and IRT tests a researcher should use as large a sample as possible in order to accurately detect inequivalences at the scale and item levels. In the situation with smaller sample sizes, it is necessary to place additional emphasis on adopting the correct procedure and, if feasible, use alternative methods.

In comparing two fit statistics, the chi-square difference test and the AIC, we found that the former has higher power in the scale level analyses. The AIC-test reveals somewhat higher 'true positive' rates in the item-level analyses accompanied, however, with higher 'false positive' rates. Nevertheless, generally speaking two fit statistics test performed very similarly and are therefore best to be used together since they complement and verify each other.

One of the important aspects of this study is the use of the same model fitting procedure for analysing ME in all three approaches. By using one standard procedure we have inevitably made some adjustments to the specific procedures of the three

approaches. In particular, both IRT and LCFA tests, when analysing individual items, usually adopt procedures based on the comparison of the 'restricted' model with all items set to be equal across groups with subsequent models in which parameters are set free one-by-one, yielding a forward-inclusion procedure. This procedure is not that different from the backward-elimination procedure that we used in this research, since both procedures are based on the comparison of subsequent models that differ only in the status of the parameters of one item. However, we favour and recommend the 'backward-exclusion' procedure used in this study, as it insures that the more restricted model is compared with a model that fits data, which is not always the case in the 'forward including' procedure.

One of the novelties of this research was the comparison of the LCFA approach, which is less known, with two other more standard CFA and IRT approaches to ME. An important finding from this comparison was that LCFA proved to be a valid and reliable alternative for dealing with inequivalence both at the scale and the item level. When latent variable(s) are of categorical nature, the LCFA should be the first choice for the analysis of ME. What is more, although it has shown slightly better performance when used with ordinal latent variables, results from this study indicate that the LCFA is a viable option in case of continuous latent variable(s) too, at least for the given characteristics of the simulation data used in this study.

Meade and Lautenschlager (2004a) and Raju et al. (2002) called for further studies on the CFA and IRT-based tests of ME and the relationships between them. In this study, we widened the range of compared approaches to ME by including a rather new and promising procedure – multi group LCFA. Although this study shredded some light on the distinctive characteristics and inherent similarities of the three approaches as well as on the conditions under which they are most suitable to be

applied, we feel that further work is still needed. The two measurement models in this study, one with a discrete and one with a continuous latent variable, are both modelled with non-linear relationship between latent and indicator variables, which is – to some extent – favouring the IRT and LCFA methods. Of course, this choice is made because indicator variables in attitude research are rarely interval or ratio level variables. Nevertheless, there is also need for simulation studies modeling linear relationships in the measurement scales and then assess the performance of the CFA, IRT and LCFA procedures. Furthermore, the number of design factors included in the analysis was limited. Future studies should investigate the influences of differences in latent variable means and variance across groups, different forms and magnitudes of imbedded inequivalence, number of groups compared, items per scale, and response categories per item, in order to determine and compare performances of the three approaches to ME.

References

Andrich, D. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43: 561–573.

Bartholomew, D. J. and Knott, M. 1999. *Latent Variable Models and Factor Analysis*, London: Arnold.

Billiet, J.B. and McClendon, M.J. 2000. "Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items." *Structural Equation Modeling* 7: 608-628.

Bock, R.D. and Zimovski, M.F. 1997. "Multiple Group IRT." Pp. 433-448 in *Handbook of Modern Item Response Theory*, edited by W.J. van der Linden & R.K. Hambleton. Springer: New-York.

Clogg, C.C. and Goodman, L.A. 1985. "Simultaneous Latent Structure Analysis in Several Groups." *Sociological Methodology* 1985: 81–110.

Greenleaf, E. A. 1992. "Measuring Extreme Response Style." *Public Opinion Quarterly* 56: 323-351.

French, B. F. and Finch, W. H. 2006. "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance." *Structural Equation Modeling* 13: 378-402.

Hagenaars, J. A. 1990. *Categorical Longitudinal Data–Loglinear Analysis of Panel, Trend and Cohort Data.* Newbury Park, CA: Sage.

Heinen, T. 1996. *Latent Class and Discrete Latent Trait Models: Similarities and Differences.* Thousand Oaks, CA: Sage.

Horn, J. L. and McArdle, J. J. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18: 117-144.

Hui, C.H. and Triandis, H.C. 1989. "Effects of Culture and Response Format on Extreme Response Style." *Journal of Cross-Cultural Psychology* 20(3): 296–309.

Jöreskog, K.G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 6: 409–426.

Kankaraš, M. and Moors, G. Forthcoming. "Measurement Equivalence in Solidarity Attitudes in Europe. Insights from a Multiple Group Latent Class Factor Approach." *International Sociology*.

Magidson, J. and Vermunt, J.K. 2001. "Latent Class Factor and Cluster Models, Bi-Plots and Related Graphical Displays." *Sociological Methodology* 31: 223-264.

------. 2003. "Comparing Latent Class Factor Analysis with the Traditional Approach in Data Mining. Pp. 373-383 in *Statistical Data Mining and Knowledge Discovery,* edited by H. Bozdogan. Boca Raton: Chapman & Hall/CRC.

Masters, G. N. 1982. "A Rasch Model for Partial Credit Scoring." *Psychometrik,* 47: 149–174.

McCutcheon, A. 2002. "Basic Concepts and Procedures in Singe- and Multiple-Group Latent Class Analysis." Pp. 56-88 in *Applied latent class analysis*, edited by J. Hagenaars and A. McCutcheon. Cambridge University Press.

Meade, A.W. and Bauer, D.J. 2007. "Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance." *Structural Equation Modeling* 14(4): 611-635.

Meade, A.W. and Lautenschlager, G.J. 2004a. "A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance." *Organizational Research Methods* 7 (10): 361–88.

------. 2004b. “A Monte-Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance.” *Structural Equation Modeling* 11(1): 60-72.

Mellenbergh, G.J. 1989. “Item Bias and Item Response Theory.” *International Journal of Educational Research* 13: 127 – 143.

Moors, G. 2003. “Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Re-examined.” *Quality & Quantity* 37: 227-302.

------. 2004. “Facts and Artefacts in the Comparison of Attitudes among Ethnic Minorities. A Multi-Group Latent Class Structure Model with Adjustment for Response Style Behaviour.” *European Sociological Review* 20: 303-320.

Muraki, E. 1999. “Stepwise Analysis of Differential Item Functioning Based on Multiple-Group Partial Credit Model.” *Journal of Educational Measurement*, 36: 217–232.

Paulhus, D. L. 1991. “Measures of Personality and Social Psychological Attitudes.” Pp. 17-59 in *Measures of Social Psychological Attitudes Series*, vol. 1, edited by J.P. Robinson and R.P. Shaver. San Diego: Academic.

Raju, N.S., Laffitte, L.J., and Byrne, B.M. 2002. “Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory.” *Journal of Applied Psychology* 87(3): 517–29.

Samejima, F. 1969. “Estimation of Latent Ability Using a Response Pattern of Graded Scores.” *Psychometric Monograph Supplement* No.17.

Skrondal, A. and Rabe-Hesketh, S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Steenkamp, J.E.M., and Baumgartner, H. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25: 78-90.

Thissen, D., Steinberg, L., and Wainer, H. 1988. "Use of Item Response Theory in the Study of Group Differences in Trace Lines." Pp. 147-169 in *Test Validity*, edited by H. Wainer and H. Braun. Hillsdale, NJ: Erlbaum.

Vandenberg, R.J. and Lance, C.E. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 2: 4-69.

Vermunt, J.K. 2001. "The Use of Restricted Latent Class Models for Defining and Testing Nonparametric and Parametric IRT Models." *Applied Psychological Measurement* 25: 283-294.

Vermunt, J. K. and Magidson, J. 2008. *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*, Belmont, MA: Statistical Innovations Inc.

Table 1: Classification of latent variable models

		LATENT VARIABLES	
		Continuous (Normal)	Categorical (Multinomial)
RESPONSE VARIABLES	Continuous (Normal)	Factor Analysis	Latent Profile Analysis
	Categorical (Multinomial)	Item Response Theory	Latent Class Analysis

Table 2 Characteristics of the CFA, IRT and LCFA models for ME

	CFA	IRT	LCFA
A. Model assumptions			
Distribution of latent variable - $f(\Theta)$	Continuous Normal	Continuous Normal	Discrete Multinomial
Distribution of response variables - $f(y_k\|\Theta)$	Continuous Normal	Discrete Multinomial	Discrete Multinomial
Regression model for response variables	Linear	Logit	Logit
B. Model parameters			
Intercept parameter β_0	Item intercept	Function of difficulty parameter	Intercept
Slope parameter β_1	Factor loading	Discrimination parameter	Beta loading
Inequivalent β_0	Scalar inequivalence	Uniform DIF	Direct effect
Inequivalent β_1	Metric inequivalence	Non-uniform DIF	Interaction effect

Table 3. Number of samples (out of 100) in which measurement inequivalence was detected in the condition with 3 inequivalent items in a scale

Form of inequivalence	Latent variable distribution	Sample size	Scale-level analysis									Item- level analysis													
			Loadings			Intercepts			Equivalent			Item 1			Item 2			Item 3			Item 4			Item 5		
			CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA	CFA	IRT	LCFA
Agreement bias	Ordinal	200	0	6	8	**77**	**92**	**91**	23	2	1	2	2	8	3	3	4	**42**	**53**	**79**	**42**	**57**	**91**	**42**	**66**	**9 3**
		1000	4	5	5	**96**	**95**	**95**	0	0	0	1	3	5	5	4	7	**99**	**100**	**100**	**100**	**100**	**100**	**99**	**100**	**100**
	Continuous	200	1	2	6	**67**	**92**	**92**	32	6	2	1	2	11	1	5	10	**40**	**51**	**80**	**41**	**49**	**76**	**37**	**43**	**7 2**
		1000	0	5	4	**100**	**95**	**96**	0	0	0	2	3	8	3	2	5	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Extreme bias	Ordinal	200	53	9	4	**2**	**88**	**95**	45	3	1	1	4	4	0	1	3	**0**	**84**	**86**	**1**	**71**	**78**	**0**	**78**	**8 6**
		1000	100	2	5	**0**	**98**	**95**	0	0	0	1	3	2	2	4	5	**2**	**100**	**100**	**2**	**100**	**100**	**2**	**100**	**100**
	Continuous	200	39	10	6	**0**	**80**	**90**	61	10	2	1	2	11	3	1	10	**1**	**65**	**80**	**2**	**64**	**76**	**1**	**66**	**7 2**
		1000	100	6	8	**0**	**94**	**92**	0	0	0	1	0	10	0	2	10	**0**	**100**	**100**	**1**	**100**	**100**	**0**	**100**	**100**
Bias in slope parameters	Ordinal	200	**100**	**79**	**100**	0	16	0	0	5	0	2	3	4	2	3	3	**82**	**64**	**91**	**81**	**63**	**84**	**86**	**60**	**8 3**
		1000	**100**	**99**	**100**	0	1	0	0	0	0	2	4	2	2	4	6	**100**	**99**	**100**	**100**	**100**	**100**	**100**	**99**	**100**
	Continuous	200	**84**	**70**	**75**	0	22	25	16	8	0	4	3	10	4	3	8	**68**	**48**	**52**	**68**	**46**	**44**	**67**	**47**	**5 4**
		1000	**100**	**100**	**98**	0	0	2	0	0	0	1	7	6	1	7	8	**100**	**100**	**96**	**100**	**100**	**100**	**100**	**100**	**9 8**

*Numbers in bold represent 'true positive' rates, columns 'Equivalent' give 'false negative' rates, while rest of the numbers shows 'false positive' rates

Table 4. Mean and standard deviation of the number of samples in which inequivalence was detected by form of inequivalence

Type of inequivalence	Analysis	SCALE LEVEL						ITEM LEVEL			
		True positives		False positives		False negatives		True positives		False positives	
		M	SD	M	SD	M	SD	M	SD	M	SD
Agreement bias	CFA	80.6	17.5	3.3	3.0	16.1	18.1	74.6	26.6	3.1	2.1
	IRT	78.2	22.6	7.1	2.7	14.7	22.1	78.7	22.6	3.8	1.9
	LCFA	79.8	21.3	8.6	2.8	11.7	19.2	91.3	11.1	14.1	11.0
Extreme response bias	CFA	0.6	0.8	72.8	29.0	26.7	28.5	2.6	2.4	2.8	2.1
	IRT	73.0	28.0	8.3	3.1	18.7	26.5	88.4	13.0	3.6	2.1
	LCFA	79.0	24.1	7.3	2.5	13.7	22.7	90.8	11.1	13.8	11.5
Inequivalence in β_1 parameters	CFA	94.6	9.1	0.0	0.0	5.4	9.1	91.0	11.4	6.3	4.8
	IRT	80.7	23.0	12.4	16.1	6.9	9.9	79.8	21.9	10.9	7.0
	LCFA	83.4	21.9	12.5	14.7	4.1	9.7	87.0	17.7	11.5	6.8

Table 5. Mean and standard deviation of the number of samples in which inequivalence was detected by number of inequivalent items per scale

Number of inequivalent items	Analysis	SCALE LEVEL						ITEM LEVEL			
		True positives		False positives		False negatives		True positives		False positives	
		M	SD	M	SD	M	SD	M	SD	M	SD
1	CFA	56.3	43.0	24.4	37.2	19.3	24.4	56.5	42.6	4.0	3.6
	IRT	67.4	29.4	11.4	12.3	21.2	26.3	80.5	22.4	6.2	5.7
	LCFA	70.3	26.5	12.1	10.5	17.6	23.0	88.1	16.0	13.6	10.4
3	CFA	60.8	44.4	26.3	39.0	12.9	18.4	55.9	42.1	4.3	3.6
	IRT	87.2	11.9	7.2	5.3	5.6	8.4	82.9	19.2	5.9	5.1
	LCFA	91.2	8.2	6.8	5.9	2.0	4.1	90.2	12.9	12.1	9.1

Table 6. Mean and standard deviation of the number of samples in which inequivalence was detected by fit statistic

Fit criteria	Analysis	SCALE LEVEL						ITEM LEVEL			
		True positives		False positives		False negatives		True positives		False positives	
		M	SD	M	SD	M	SD	M	SD	M	SD
Chi-square difference	CFA	58.0	43.9	23.7	37.6	18.3	24.2	52.9	42.4	1.8	1.5
	IRT	81.1	20.6	9.5	12.0	9.4	15.3	79.1	22.7	3.3	2.2
	LCFA	84.2	18.1	9.5	11.0	6.3	12.0	87.5	16.0	7.7	6.5
AIC	CFA	59.2	43.7	27.0	38.5	13.9	19.0	59.2	41.9	6.3	3.6
	IRT	73.5	27.5	9.1	6.7	17.4	24.9	85.5	16.3	8.8	6.4
	LCFA	77.3	25.4	9.4	6.4	13.4	22.4	91.9	10.6	18.4	9.8

Appendix 1 – Implied class-specific item distribution under agreement bias, extreme response bias, and bias in slope parameters

Agreement bias:

		Item category				
		1	**2**	**3**	**4**	**5**
Group 1	Class 1	0.3251	0.4281	0.2074	0.0370	0.0024
	Class 2	**0.0545**	**0.2442**	**0.4026**	**0.2442**	**0.0545**
	Class 3	0.0024	0.0370	0.2074	0.4281	0.3251
Group 2	Class 1	0.1497	0.5360	0.2597	0.0463	0.0082
	Class 2	**0.0189**	**0.2306**	**0.3801**	**0.2306**	**0.1398**
	Class 3	0.0006	0.0237	0.1332	0.2749	0.5675

Extreme bias:

		Item category				
		1	**2**	**3**	**4**	**5**
Group 1	Class 1	0.3251	0.4281	0.2074	0.0370	0.0024
	Class 2	**0.0545**	**0.2442**	**0.4026**	**0.2442**	**0.0545**
	Class 3	0.0024	0.0370	0.2074	0.4281	0.3251
Group 2	Class 1	0.5245	0.3002	0.1454	0.0259	0.0039
	Class 2	**0.1098**	**0.2139**	**0.3526**	**0.2139**	**0.1098**
	Class 3	0.0039	0.0259	0.1454	0.3002	0.5245

Bias in slope parameters:

		Item category				
		1	**2**	**3**	**4**	**5**
Group 1	Class 1	0.3251	0.4281	0.2074	0.0370	0.0024
	Class 2	**0.0545**	**0.2442**	**0.4026**	**0.2442**	**0.0545**
	Class 3	0.0024	0.0370	0.2074	0.4281	0.3251
Group 2	Class 1	0.6921	0.2678	0.0381	0.0020	0.0000
	Class 2	**0.0545**	**0.2442**	**0.4026**	**0.2442**	**0.0545**
	Class 3	0.0000	0.0020	0.0381	0.2678	0.6921

APPENDIX 2 – Latent Gold syntax files used in the simulation study

This appendix explains the Latent GOLD 4.5 syntax files we used for our simulation study. The variables and equations sections of the syntax file for a fully heterogeneous CFA model in which the first item serves as reference item is as follows:

```
variables
   dependent y1 continuous, y2 continuous, y3 continuous,
             y4 continuous, y5 continuous;
   independent group nominal;
   latent theta continuous;
equations
   theta | group;
   theta <- group;
   y1 <- 1 + (1) theta;
   y2 <- 1 | group + theta | group;
   y3 <- 1 | group + theta | group;
   y4 <- 1 | group + theta | group;
   y5 <- 1 | group + theta | group;
```

In the variables section we provide the relevant information on the dependent, independent, and latent variables to be used in the analysis. In a factor analysis, the dependent and latent variables are defined to be continuous. The first two equations concern the variance and the regression model for the latent variable: the variance is assumed to depend on group (indicated with "| group") and the mean is regressed on group (the intercept is omitted for identification purposes). The other five equations concern the regression models for items y1 to y5. These contain the term "1" referring to the intercept and the term "theta" referring to the slope. Except for the reference item y1, these are indicated to differ across groups with "| group". The term "(1)"

preceding “theta” in the model for y1 indicates that the loading for the first item is fixed to 1, which is required for identification purposes.

The only necessary modification to obtain an IRT model for ordinal items is that the definition of the dependent variables should be

```
dependent y1 ordinal, y2 ordinal, y3 ordinal,
          y4 ordinal, y5 ordinal;
```

that is, by indicating that the items are ordinal instead of continuous. A latent factor class model is obtained by indicating that also the latent variable is ordinal :

```
latent theta ordinal 3;
```

where “3” indicates that there are 3 latent classes. Moreover the first three equations should be replaced by these two:

```
theta <- 1 + group;
y1 <- 1 | group + theta | group;
```

i.e., there is no latent variable variance, the latent variable intercept is identified can thus be included in the model, and no identifying constraints need to be imposed on the item parameters. The other more restricted models assuming equivalent item intercept and/or slopes across groups are obtained by removing “| group” from the term(s) concerned.